UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
INSTITUTO DE COMPUTAÇÂO
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

VITOR CURIEL TRENTIN CORRAL

Gerador de dados sintéticos para testes de rotinas de record linkage para o contexto brasileiro.

RIO DE JANEIRO
2021

VITOR CURIEL TRENTIN CORRAL

Gerador de dados sintéticos para testes de rotinas de record linkage para o contexto brasileiro.

Trabalho de conclusão de curso de graduação apresentado ao Departamento de Computação da Universidade Federal do Rio de Janeiro como parte dos requisitos para obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. Valeria M. Bastos
Co-orientador: Prof. Claudia Medina Coeli

RIO DE JANEIRO
2021

CIP - Catalogação na Publicação

C823g

Corral, Vitor Curiel Trentin
Gerador de dados sintéticos para testes de rotinas de record linkage para o contexto brasileiro / Vitor Curiel Trentin Corral. -- Rio de Janeiro, 2021.
46 f.

Orientadora: Valéria Menezes Bastos.
Coorientadora: Claudia Medina Coeli.
Trabalho de conclusão de curso (graduação) - Universidade Federal do Rio de Janeiro, Instituto de Matemática, Bacharel em Ciência da Computação, 2021.

1. Record Linkage. 2. Dados sintéticos. 3. Gerador de dados. 4. Saúde pública. I. Bastos, Valéria Menezes, orient. II. Coeli, Claudia Medina, coorient. III. Título.

Elaborado pelo Sistema de Geração Automática da UFRJ com os dados fornecidos pelo(a) autor(a), sob a responsabilidade de Miguel Romeu Amorim Neto - CRB-7/6283.

VITOR CURIEL TRENTIN CORRAL

Gerador de dados sintéticos para testes de rotinas de record linkage para o contexto brasileiro.

Trabalho de conclusão de curso de graduação apresentado ao Departamento de Computação da Universidade Federal do Rio de Janeiro como parte dos requisitos para obtenção do grau de Bacharel em Ciência da Computação.

Aprovado em 06 de Agosto de 2021

BANCA EXAMINADORA:

Valeria Menezes Bastos
D.Sc. (IC/UFRJ)

Claudia Medina Coeli
D. Sc. (IESC/UFRJ)

Myrian Christina de Aragão Costa
D. Sc. (COPPE UFRJ)

Adriana Santarosa Vivacqua, D. Sc.
(IC/UFRJ)

Dedico este trabalho aos meus pais Lucio e Luciana, sem seu apoio e seu carinho este trabalho e todas as demais conquistas de minha vida não seriam possíveis e não seriam tão alegres.

AGRADECIMENTOS

Agradecimento ao CNPq pelo incentivo ao desenvolvimento desta pesquisa. A professora Cláudia Medina pelas orientações e mentorias e um agradecimento especial à professora e orientadora Valeria Bastos pelos conhecimentos passados, suporte e parceria que teve início desde o primeiro dia do curso. E também a todos os nossos colegas da UFRJ que estiveram comigo nessa jornada, em especial aos amigos que ingressaram em 2015.1 e os da GEDAE (Daniel Artine, Silvio Mattos, Leonardo Schripsema , Willam Lacerda e Mateus Villas Boas).

RESUMO

Record linkage tem sido cada vez mais usado no Brasil, no entanto, apenas alguns estudos relatam a qualidade do processo de ligação, principalmente na área de saúde coletiva, onde é necessário efetuar a ligação dos pacientes em diversas bases de dados identificadas do SUS, para investigar causas e consequências das doenças e pacientes, e permitir estabelecer formas de controle e administração pública na área de saúde.
Dados de testes gerados sinteticamente podem ser usados para avaliar a qualidade do vínculo de dados, para desenvolver um gerador de dados sintéticos que crie conjuntos de dados de teste com atributos, características e erros semelhantes ao contexto brasileiros. Para isso, foi analisado o banco de dados de mortalidade (SIM) do Estado do Rio de Janeiro de 2013, para se conhecer as características e a distribuição de frequência dos atributos que o identificam (nome do paciente, nome da mãe, sexo, data de nascimento e endereço).
A metodologia de avaliação e geração de dados apresentado em (TRAN; VATSALAN; PETER, 2020) foi utilizado neste trabalho, com adaptações aos padrões de nomes brasileiros, sendo que suas principais rotinas foram reescritas em C++ e posteriormente em Python, compondo uma ferramenta de geração de dados pessoais para o padrão brasileiro. Os nomes brasileiros têm características específicas que os distinguem dos padrões de outros países: vários nomes de família são comuns, como nomes próprios compostos, com parte do nome do pai, ou da mãe, ou ambos, além da ocorrência frequente de homônimos. Devido às características nacionais específicas dos nomes no Brasil, a modelagem de dados sintéticos é uma atividade particularmente desafiadora e precisa ter regras mais flexíveis para gerar bancos de dados que permitam avaliar a qualidade dos processos de vinculação de dados com identificação.

Palavras-chave: Record Linkage. Dados Sintéticos. Gerador de Dados. Saúde Pública.

ABSTRACT

Record linkage has been increasingly used in Brazil. However, only a few studies report the quality of the linkage process. Synthetic test data can be used to evaluate the quality of data linkage.

To develop a synthetic data generator that creates test datasets with similar attributes and error characteristics found in the Brazilian databases. We analyzed the 2013 mortality database from Rio de Janeiro State to know the characteristics and frequency distribution of the database attributes (name, mother's name, sex, date of birth and address). We used initially C++ then Python to customize and add routines developed in (TRAN; VATSALAN; PETER, 2013), a personal data generation tool.

Brazilian names have specific characteristics that distinguish them from other countries' patterns: multiple family names are usual, as are composite first names, and, despite that, homonyms are frequent. Family names may include the full extension or only parts of either the father and mother's respective family names, or both, so there is a wide variation in progeny family names and not necessarily a common family name for all family members.

Due to the specific national characteristics of name building in Brazil, modeling synthetic data is particularly challenging and needs to have more flexible rules in order to generate databases that will actually allow assessing the quality of data linkage processes.

Keywords: Record Linkage. Synthetic Data. Data Generator. Public Helth.

LISTA DE ILUSTRAÇÕES

Figura 1 – O processo de *record linkage* 17
Figura 2 – Modelagem de variações e erros 24
Figura 3 – Linguagens mais populares no github em projetos de dados sintéticos . 26
Figura 4 – Utilização de frequências para construção de nomes 28
Figura 5 – Construção nome filho 31
Figura 6 – Frequência de nomes próprios extraídos da base SUS 40
Figura 7 – Frequência de nomes próprios obtida após a execução do gerador em padrões em brasileiro 40

LISTA DE CÓDIGOS

3.1 Importação de arquivos 33
3.2 extraiNome 34

LISTA DE TABELAS

Tabela 1 – Exemplo de dados gerados pelo GeCo 25
Tabela 2 – Dados gerados com sufixos nominais 29
Tabela 3 – Nomes desconexos 30
Tabela 4 – Construção de nomes (N NpM NP*) 32
Tabela 5 – Exemplo de duplicatas geradas com base no registro original 39

LISTA DE QUADROS

Quadro 1 – Comparativo colunas geradas 28

SUMÁRIO

1 INTRODUÇÃO 13
1.1 Contextualização 13
1.2 Problemática 13
1.3 Organização do documento 14

2 ESTADO DA ARTE 15
2.1 História 15
2.2 Processo de Record Linkage 16
2.2.1 Desafio dos dados históricos 17
2.2.2 Desafio dos dados atuais 18
2.2.3 Limpeza de dados 18
2.2.4 Padronização 20
2.2.5 Privacidade e segurança dos dados 21
2.2.6 Dados sintéticos 22
2.3 GeCo 22
2.3.1 Geração de atributos 23
2.3.2 Replicação 23
2.3.3 Corrompimento 23

3 METODOLOGIA 26
3.1 Preparação dos dados 26
3.2 Padrão australiano vs. padrão brasileiro 27
3.3 Padrão brasileiro de nomes 30
3.4 Ambiente de desenvolvimento em C 32
3.5 Ambiente de desenvolvimento Python 35
3.6 Replicando dados 35
3.7 Corrompendo a base 35

4 RESULTADOS 38

5 CONCLUSÃO 41

REFERÊNCIAS 43

APÊNDICE A – RESUMO 8ª SIAC **44**

APÊNDICE B – RESUMO 9ª SIAC **45**

APÊNDICE C – ABSTRACT IPDLN 2018 . . . . . **46**

1 INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO

Record linkage é uma técnica que serve para vincular um ou mais registros que se referem à mesma entidade, pertencente a um ou mais conjuntos de dados. O *record linkage* é necessário quando temos conjuntos de dados diferentes e cujas entidades podem ou não compartilhar um identificador único (por exemplo: CPF, CNPJ, chave do banco de dados, id), que pode ocorrer devido a diferenças na forma do registro, local de armazenamento, estilo ou preferência do curador.

O laboratório Link Data Pop, do Instituto de Estudos de Saúde Coletiva (IESC/UFRJ) trabalha com rotinas de *record linkage* em bancos de dados de saúde do SUS, com informações do estado do Rio de Janeiro, que são utilizados em processos de pareamento de dados. As bases de dados populacionais brasileiras não trazem um identificador único para cada habitante, fazendo com que a vinculação de bases seja realizada empregando-se a comparação de identificadores pessoais como nomes, datas de nascimento, nome da mãe, dentre outros.

Estes identificadores são empregados conjuntamente em algoritmos de *record linkage*, por exemplo, probabilístico, para o cálculo de um escore que indica o quão verossímil são dois ou mais registros, ou seja, pertencentes ao mesmo indivíduo, criando assim, a vinculação no histórico de saúde de um paciente, possibilitando estudos de comportamento de saúde populacional, identificação de padrões, predições de doenças com base em dados históricos e outros tipos de pesquisa.

1.2 PROBLEMÁTICA

Um dos objetivos deste trabalho é possibilitar acesso à informação restrita, permitindo o desenvolvimento de estudos como os citados anteriormente, através da geração de identificadores pessoais não verídicos, porém com a mesma constituição dos identificadores de uma base real.

Um dos problemas metodológicos a ser enfrentado é a avaliação da acurácia dos algoritmos de *record linkage*. Para essa avaliação é necessário identificar um "padrão ouro"que indique a classificação correta dos pares em falsos e verdadeiros. O padrão ouro pode ser criado por meio da revisão manual dos pares, entretanto, esse procedimento, além de ser demorado, é custoso e suscetível a erros.

Adicionalmente, por questões de privacidade dos indivíduos, o acesso a bases com identificadores pessoais é permitida somente em casos especiais, devendo o mesmo ser realizado em ambientes seguros, o que restringe o uso de dados reais para a avaliação da

qualidade dos algoritmos de *linkage*.

Nesse caso se faz necessário o uso de dados sintéticos, que espelham os ambientes reais de dados, já que não existem bases sintéticas que reproduzam os padrões de nomes brasileiros. Este trabalho propõe o desenvolvimento de um algoritmo que gera dados sintéticos, onde, por exemplo, os nomes dos indivíduos são criados a partir da distribuição de frequência dos nomes existentes em bases de dados reais, tomando-se como base (TRAN; VATSALAN; PETER, 2013). Todavia, foi necessário personalizar o gerador original, adaptando-o para os padrões brasileiros de nomes, com várias regras de criação e limitações.

1.3 ORGANIZAÇÃO DO DOCUMENTO

Este documento está organizado da seguinte forma: inicialmente, o capítulo 2 apresenta o estado da arte, principais boas práticas e estudos desenvolvidos na área de *record linkage*, o ambiente computacional e ferramentas utilizadas durante o desenvolvimento. No capítulo 3 é descrito o trabalho desenvolvido, as rotinas e os problemas que o gerador busca emular. No capítulo 4, foram avaliados os problemas e resultados obtidos durante o decorrer do trabalho. Por fim, o capítulo 5 apresenta as conclusões do trabalho e desenvolvimentos futuros.

2 ESTADO DA ARTE

Em muitos projetos de mineração de dados, as informações de várias fontes de dados precisam ser integradas, combinadas ou vinculadas para permitir uma análise mais detalhada. O objetivo destas ligações é mesclar todos os registros relacionados à mesma entidade, como, por exemplo, um paciente ou um cliente. Na maioria das vezes, o processo de vinculação é um desafio pela falta do identificador de entidade único comum, portanto, torna-se não trivial.

Vincular as grandes coleções de dados de hoje torna-se cada vez mais difícil através de técnicas tradicionais de vinculação. Sendo assim, são necessárias novas abordagens, como: probabilística para limpeza e padronização de dados aprimorados, métodos de indexação inovadores, paralelização e o desenvolvimento de um gerador de conjunto de dados indexados, foco deste trabalho, e que permite a criação de registros contendo nomes, endereços e outros dados pessoais (CHRISTEN; CHURCHES; HEGLAND, 2004).

O processo de *record linkage* possui muitas aplicações, tais como remoção de duplicatas, junção de novos registros no *dataset* unificado, encontro de perfis de pessoas/pacientes/clientes que aparecem em bases que foram unificadas, limpeza e enriquecimento dos dados para análises e mineração de dados, identificações geográficas, dentre outros. E essas aplicações podem ser realizadas em diversas áreas, como, por exemplo: imigração, impostos, censos, identificação de fraudes, crimes, análise de informações de compra, saúde e social.

Este trabalho tem como foco a área de saúde e a geração de dados indexados, e busca realizar engenharia reversa das técnicas citadas acima, assim como o trabalho de (TRAN; VATSALAN; PETER, 2013), para construção de uma base de dados sintética, com o objetivo de auxiliar na correção de algoritmos de *record linkage* e estimular estudos na área, já que dados sintéticos são de utilização aberta e não sensível.

2.1 HISTÓRIA

Processos de *record linkage* com ajuda computacional tiveram início na década de 50, onde foram desenvolvidos trabalhos baseados em métodos de heurísticas ad-hoc (NEWCOMBE HOWARD B.AND KENNEDY; AXFORD; JAMES, 1959). A ideia inicial de *record linkage* probabilístico foi introduzida em 1962 em (NEWCOMBE HOWARD B.AND KENNEDY, 1962) .Mais a frente, as teorias fundamentais da área foram desenvolvidas por (FELLEGI; SUNTER, 1969), onde, através de comparações de atributos e conciliação de pesos baseados nas frequências dos dados, estimadores de erros eram utilizados para identificar a probabilidade de conciliação ou não de dois registros. Como descrito ao longo do trabalho, probabilidade e frequência são as principais técnicas empregadas

para o entendimento do comportamento e distribuição dos registros.

Em um passado mais recente, o interesse em computação e a utilização de dados em diversos campos, como mineração de dados, inteligência artificial, engenharia do conhecimento, sistemas de informação e bibliotecas digitais, aceleraram o desenvolvimento neste campo, onde diversas técnicas e algoritmos de estado da arte estão sendo utilizados para *record linkage*. Estes métodos podem ser divididos em três categorias:

a) Determinísticos:
   - Onde todo processo de conciliação é feito através de regras, executadas sequencialmente e não possuem nenhuma retroalimentação. Esta metodologia pode ser aplicada manualmente, visto que é possível desenvolver um roteiro de validações e regras que são avaliadas e executadas por um grupo de pessoas sobre os dados ou computacionalmente, onde as regras são programadas em um script. No laboratório LinkDataPop técnicas de *record linkage* foram empregadas em (COELI CLáUDIA MEDINA, 2012);

b) Probabilísticos:
   - Utilizam informações e atributos dos registros e geralmente são informações pessoais, como nomes, endereços, data de nascimento e outros, calculando-se os pesos e as probabilidades de dois registros pertencerem a mesma pessoa. Um desafio desta metodologia é a conversão de dados de texto em números ou aproximações, e para isso são utilizadas técnicas de distanciamento, as quais serão evidenciadas neste trabalho;

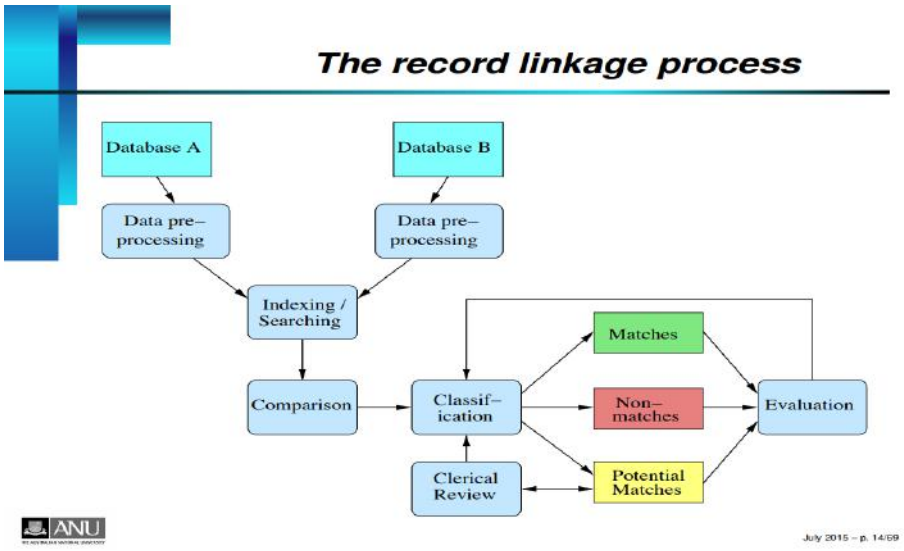
c) Computacionais:
   - Utilizam algoritmos de aprendizados de máquina, de classificação, regressão, além de outros algoritmos, podendo ser supervisionados, como árvores de decisão, ou não-supervisionados como clusterização ou redes neurais, por exemplo.

2.2 PROCESSO DE RECORD LINKAGE

Dadas duas bases de dados A e B, o processo de *record linkage* se inicia pelo preprocessamento dos registros de cada base, onde é feita a padronização das mesmas, tipagem de dados, remoções de erros, tratamento de valores faltantes, além da criação de novos campos que auxiliem na identificação de perfis e de outros processos, como veremos a frente subseção 2.3.1.

Em seguida, é realizada a indexação e buscas na base, com o objetivo de definir chaves pseudo-únicas, pois o *record linkage* se faz necessário quando não se tem uma chave única na base para identificar perfis. Logo, é preciso criar chaves utilizando a combinação de

Figura 1 – O processo de *record linkage*



Fonte: Cristen, Peter (2015, p. 15)

um ou mais campos para aproximar o máximo possível de uma chave única para aquele perfil.

Após a indexação e buscas na base, é efetuado o processo de comparação, onde, dependendo do método escolhido, são calculadas distâncias, probabilidades e scores de similaridade entre os registros. Estas informações são então utilizadas no passo de classificação, e, através de heurísticas ou probabilidades, é verificado se dois registros pertencem ao mesmo perfil ("Match"), não pertencem ("Non-match") ou estão em uma zona de incerteza ("*Potencial Matches*"), e neste último caso, a heurística utilizada não consegue definir com segurança se os dois registros pertencem ou não ao mesmo perfil.

Os casos potenciais passam então para uma etapa de revisão manual ou crítica sobre sua classificação, retroalimentando o passo de classificação. Uma vez classificados, os registros passam por uma avaliação que pode ser manual ou automática, dependendo do processo, e conforme o resultado retornam ao processo de classificação mais uma vez.

Durante o processo de *record linkage*, um par de registros pode passar diversas vezes pela etapa de classificação, com o objetivo de otimizar a quantidade de registros classificados corretamente, validando a acurácia do processo.

2.2.1 Desafio dos dados históricos

Como um dos objetivos do *record linkage* é a construção de perfis em bases de dados não diretamente relacionadas (que não possuem chave única), torna-se comum a utilização de técnicas para construção de perfis em bases de dados históricas, muitas vezes advindas de registros manuais, transcritos durante processos de digitalização. Anteriormente, a preocupação com a qualidade e rastreabilidade dos dados de bases históricas era baixa,

como, por exemplo, a base SINASC do SUS, onde a coleta dos dados é feita manualmente, e muitas das vezes por profissionais não orientados e sem um sistema padrão. Com isso, a baixa qualidade trouxe consigo os seguintes desafios durante o processo de *record linkage*:

a) Baixo nível de alfabetização:
   - Erros de escrita;
   - Incerteza em relação aos valores;

b) Alta incidência de Homônimos:
   - Muitas pessoas com mesmos nomes, principalmente mesmo primeiro nome;

c) Erros de transcrição, escaneamento ou de OCR ("Reconhecimento ótico de caracteres"):
   - Dados ilegíveis, incompletos e/ou errados;

2.2.2 Desafio dos dados atuais

Os dados das bases mais recentes oferecem benefícios quando comparados com dados históricos, principalmente relacionados a quantidade de informações e qualidade das mesmas, muitas vezes longe de um cenário que poderia ser considerado ideal, mas mesmo assim melhor, comparativamente. Trazem consigo tipos de dados mais complexos, como campos de texto aberto, multimídia e outros.

Entretanto, esses dados acarretam outros desafios, pois pertencem a pessoas vivas, e a segurança destas informações e a confidencialidade das mesmas é um tema critico, ainda mais quando se trata de dados de saúde, como o considerado neste trabalho.

Atualmente, existe maior disponibilidade de dados de diferentes formas, por exemplo, redes sociais, internet (*crawlers*), dados governamentais públicos, dentre outros, e que levantam as perguntas: "quais dados são mais adequados para a resolução deste problema?"e "quais dados teremos acesso?". A execução deste trabalho foi focada na utilização das bases de dados identificadas do SUS, que são de acesso restrito, porém, foram acessadas no laboratório LinkDataPop do IESC/UFRJ, curador de tais informações.

Outro objetivo do trabalho é justamente oferecer acesso à dados identificados e estimular pesquisas com os mesmos, através da geração de bases sintéticas geradas a partir das frequências de dados das bases de insumo do Sistema Único de Saúde.

2.2.3 Limpeza de dados

Dados reais não possuem padrões bem definidos e contém uma série de erros que torna difícil seu processamento. Os tipos mais comuns de erros encontrados, foram:

a) Tipográficos:

  - Durante digitação dos dados, o responsável pode clicar uma ou mais teclas erroneamente;
  - Fonéticos - muitas vezes os dados são coletados verbalmente e então transcritos ou digitados, essa etapa verbal pode ocasionar erros quando duas ou mais silabas com fonética são muito parecidas. No português, é possível citar como exemplos: "ss", "s", "c"e "ç", "p"e "t", "b"e "d";

b) Codificação dos dados:

  - Dados digitais possuem diferentes padrões de *encondig*, com isso, dependendo da codificação usada na leitura dos dados, os caracteres são interpretados de formas diferentes. Este problema pode ser resolvido através de identificadores desses padrões, mas muitas vezes dependemos do processo de aquisição para fazê-lo da forma correta, pois caso exista uma troca de *enconding* durante a aquisição, a informação original do caractere pode ser perdida;

c) Valores faltando:

  - Estes dados podem ter sido perdidos durante o processo de aquisição, podem não ter sido coletados, ou, devido à falta de padrões de coleta, podem ter sido excluídos, além de outros problemas. Isso acarreta problemas para o *record linkage*, pois muitas vezes aqueles campos eram essenciais para a criação do perfil do registro e a falta do mesmo causa incerteza no pareamento entre o mesmo registro na outra base.

d) Mudança nos dados ao longo do tempo:

  - Bases de dados históricos são afetadas pela passagem do tempo, visto que os dados podem mudar temporalmente. Tomando como exemplo dados pessoais, o nome pode mudar, dado que a pessoa se casa, ou até mesmo opta por mudar de nome. Mais comum que isso são as mudanças de endereço.

Outro fato é que nomes e endereços possuem grande incidência de erros durante a coleta de dados, principalmente por fatores fonéticos, como citado anteriormente. Além disso, a mesma pessoa pode fornecer seus dados de várias formas e em momentos distintos, criando diversas possibilidades para o nome da mesma pessoa. Por exemplo, meu nome completo é Vitor Curiel Trentin Corral, mas posso me apresentar como:

- Vitor Curiel
- Vitor Curiel Trentin
- Vitor Curiel Trentin Corral
- Vitor Trentin Corral
- Vitor Corral

– Vitor Curiel Corral

– Vitor C. T. Corral

Sem considerar os casos nos quais a ordem os nomes fornecidos são trocadas. O meu nome pode servir como exemplo também para erros fonéticos, visto que hoje no Brasil temos as variações "Victor"e "Vitor"para a mesma sonoridade, sem contar variações mais atípicas como "Viktor", "Victtor", dentre outras possibilidades.

Quando se trata de endereços, no Brasil eles são compostos por:

– Logradouro (avenida, rua, praça, Lote, etc.)

– Endereço (nome do logradouro)

– Complemento (apartamento, A/B, fundos, etc.)

– CEP

– Bairro

– Cidade

– Estado

A maioria destes dados, se não todos, podem sofrer erros fonéticos, tipográficos, de abreviação (Ex: Praça, Pç) e valores faltando. Geralmente, são campos livres para escrita e nem todos os sistemas possuem validações de CEP e/ou integrações com sistemas que possam obter as demais informações do endereço com base no CEP. Uma hipótese inicial seria desconsiderar endereços para realização do *record linkage*, entretanto os campos de endereço são chave para a identificação de homônimos, ou pessoas que possuem um nome idêntico a outra.

Para o desenvolvimento deste trabalho foram utilizadas macro categorias de código do município e do bairro, as quais não são exaustivas, visto que é possível ter dois homônimos no mesmo bairro da mesma cidade. Mas, como apresentado mais a frente, a utilização destes dados em conjunto com outras métricas calculadas melhorou a abrangência da identificação destes casos.

2.2.4 Padronização

Outra parte importante para o processo de *record linkage* é a padronização dos dados, onde dados semelhantes podem ter uma série de erros, formatações e escritas diferentes. A padronização dos dados é utilizada para auxiliar o passo de classificação, diminuindo as diferenças irrelevantes para o processo de *linkage*, aumentando as semelhanças entre os registros e a chance de encontrar pares de registros para a criação dos perfis.

Um bom exemplo de padronização é a utilização dos códigos de municípios e bairros em vez do endereço completo de um indivíduo. Desta forma, estamos diminuindo erros

de digitação e possíveis diferenças entre os registros e os aproximando mais. Porém, deve ser feito com cuidado por se tratar de um processo delicado, e se a padronização for muito genérica, é possível perder informações que poderiam ser relevantes para a classificação de dois registros.

Tradicionalmente a padronização de dados era feita através de regras e transformações, aplicadas manualmente ou durante a importação dos dados, gerando novos campos ou alterando campos já existentes. O problema deste método é que consome muito tempo, tanto para o reconhecimento quanto para a geração das transformações, complexas para se desenvolver e de difícil manutenção.

Atualmente as técnicas mais utilizadas são os métodos probabilísticos, que tem muita interface com este trabalho, visto que são flexíveis e robustos, pois utilizam dos próprios dados como fonte para estabelecer regras e para calcular as probabilidades. O contraponto deste método é que geralmente necessitam de dados de treinamento, onde o método precisa "aprender"a identificar estes padrões e muitas das vezes essas bases de treinamento são desenvolvidas manualmente ou revisadas após algum processo de padronização.

A padronização é um dos focos deste trabalho, visto que os dados são gerados sinteticamente, possível manter o controle da padronização também aplicável nas bases de treinamento, auxiliando o desenvolvimento de novos métodos e algoritmos.

O processo de padronização é dividido em dois passos: limpeza e segmentação. O primeiro é responsável por corrigir erros de escrita e remover caracteres e/ou campos desnecessários. A segmentação é responsável por repartir dados em informações menores, definindo um padrão. Um exemplo de segmentação utilizado neste trabalho é a segmentação de nomes.

2.2.5 Privacidade e segurança dos dados

Como dito anteriormente, o processo de *record linkage* é muito utilizado em bases de dados de saúde, visto que não possuem chave única para os registros (CPF, Id de cadastro, etc.) e são consideradas de grande interesse na construção de perfis para identificação de tendências e/ou padrões dentro de uma população, ou grupo de pessoas com alguma doença. O laboratório LinkDataPop, onde este trabalho foi desenvolvido, possui diversas aplicações para bases linkadas, tais como: Saúde da mulher, criança e juventude, Tuberculose, Diabetes, Avaliação de Serviços de Saúde, Intervenções em Saúde Pública, dentre outras.

Entretanto, essas aplicações e dados acarretam uma grande responsabilidade com a segurança dos dados e a privacidade das informações das pessoas presentes nas bases, visto que informações de nome, nome da mãe, endereço, histórico médico, relatórios médicos e outras informações sensíveis estão disponíveis.

Nesses casos, são necessários contratos de confidencialidade, onde são definidos padrões de segurança e manipulação, aos quais os dados devem seguir e quais pessoas podem ter

acesso e somente em equipamentos livres de acesso à internet e a usuários não autorizados.

Este cuidado com as informações é extremamente necessário, mas acarreta uma grande restrição no acesso a estas informações e, com isso, nem todos os pesquisadores podem trabalhar com essas bases, afetando também o potencial de descobertas que poderiam ser realizadas na área.

Uma possível abordagem para este problema de privacidade seria a desidentificação dos dados, removendo qualquer traço de informação relacionada diretamente a uma ou mais pessoas. Este tipo de abordagem é muito utilizada em censos, onde as informações são desidentificadas e acessadas de forma agrupada. Entretanto, essa forma de trabalho com as informações vai contra o objetivo do *record linkage*, sendo justamente a formação de perfis e históricos, e onde existe a necessidade de utilizar os nomes e endereços reais, pois são chaves para ligação entre dois registros.

Por este mesmo motivo, não é possível a utilização de criptografia dos dados, que substituiria os valores reais por valores criptografados, mas sem valor para o processo de *linkage*.

2.2.6 Dados sintéticos

Uma abordagem plausível para este problema, proposta em (TRAN; VATSALAN; PETER, 2013), é de gerar os dados sinteticamente, para não perder os dados (e erros) reais, mas sim "embaralhá-los", mantendo a confidencialidade.

> Uma alternativa ao uso de dados reais é gerar dados sintéticos (ou artificiais) baseados em dados reais. Esses dados gerados devem exibir características estatísticas semelhantes em comparação com os dados reais em que se baseiam. Por exemplo, os valores gerados, suas distribuições de frequência, as ocorrências, frequências de tipografia, outros erros e variações devem seguir dados reais. Dependências entre elementos de dados reais também devem ser modeladas. As vantagens metodológicas dos dados sintéticos são que eles podem ser gerados com frequência bem definida, distribuições, características e variações de erros; o status de quais registros foram gerados com base em cada outro é conhecido, permitindo, por exemplo, medir a precisão dos registros correspondentes quando dados sintéticos são usados para ligação de registros (geralmente não é possível em dados reais porque as correspondências verdadeiras são desconhecidas); e os dados gerados, como bem como o próprio programa gerador, podem ser publicados.
> (Christen and Vatsalan, 2012, p. 1 - Traduzido pelo autor)

2.3 GECO

O gerador foi desenvolvido como parte da suite de ferramentas e técnicas do projeto (CHRISTEN; CHURCHES; HEGLAND, 2004), com o objetivo de gerar dados sintéticos com alta flexibilidade e customização. É possível dividir as atuações do gerador em algumas sessões, que serão trabalhadas também ao longo do desenvolvimento do projeto.

2.3.1 Geração de atributos

O processo de geração de dados com base em tabelas de frequência é flexível e permite a criação de atributos de diferentes tipos de dados, tornando possível explorar os dados reais, criando tabelas de frequência que distanciam a informação dos dados reais. A geração de dados com base em funções programadas pelo usuário garante alta flexibilidade durante o processo. Atributos compostos, onde a geração de um valor depende de outro atributo ou função especificada pelo usuário, tornando fácil relacionar atributos e criarmos dependências.

Sendo assim, é possível criar quantos atributos diferentes forem necessários, e, conforme as formas especificadas acima, são definidos os nomes dos atributos e a ordem que os dados serão gerados.

2.3.2 Replicação

Na etapa de replicação, o objetivo é gerar diversas linhas com variações dos atributos listados, simulando a realidade de diferentes registros pertencentes ao mesmo perfil, encontrados no processo de *record linkage*.

O grande ganho desta etapa para o processo de *linkage* é que quando as cópias modificadas do registro original são geradas, também é mantida a relação entre eles através de um id, denominado "rec-number", tornando possível a utilização dos demais atributos gerados durante o processo de *linkage* e id dos dados em uma etapa de averiguação da acurácia do algoritmo. Com isso, pode-se identificar e qualificar em quais categorias de relações o algoritmo de *linkage* está falhando.

2.3.3 Corrompimento

Adjunto ao processo de replicação, o corrompimento dos dados é responsável por criar variações diferentes para os atributos de um mesmo registro original, emulando as diferentes possibilidades encontradas em atributos de um mesmo perfil durante o processo de *record linkage*.

Esta etapa é chamada corrompimento, pois ela engloba a simulação dos erros encontrados na base real (Figura 2), corrompendo então os dados que seriam gerados sem erros. No GeCo esta etapa pode englobar:

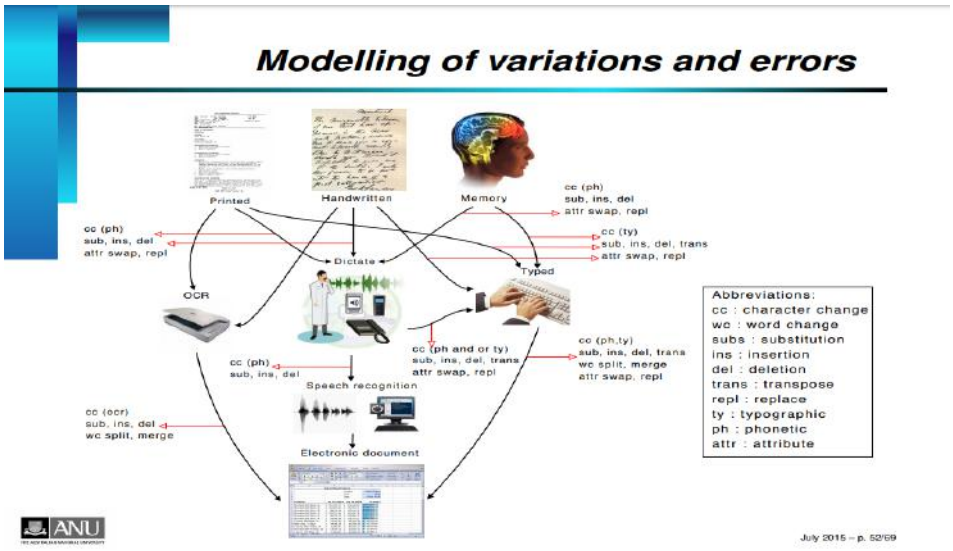
a) Edição aleatória de caracteres;

b) Mudança aleatória de valor da tabela de frequência, por outro valor da mesma tabela;

c) Mudança aleatória de caractere por tabela de reconhecimento ótico;

d) Aplicação aleatória de erro fonético;

Figura 2 – Modelagem de variações e erros



Fonte: Cristen, Peter (2015, p. 53)

A aplicação fornece ampla configuração em relação à frequência e aleatoriedade que estes erros e alterações podem acontecer durante o processo de geração, possibilitando gerações de bases "simples"com poucos erros, variações e bases complexas, onde a identificação de perfis seria quase impossível.

O objetivo do gerador desenvolvido por (TRAN; VATSALAN; PETER, 2013) é ser flexível, buscando se adaptar a novas regras, funções e bases de dados para a geração de dados sintéticos de diferentes realidades, oferecendo inicialmente configurações aplicadas nos trabalhos desenvolvidos na Austrália (Tabela 1) e no Japão. No presente trabalho, foram utilizadas as documentações e o código do GeCo (TRAN; VATSALAN; PETER, 2020), buscando a customização do mesmo para a realidade brasileira, a partir das características das bases de dados reais do SUS dos dados existentes no laboratório LinkDataPop.

Tabela 1 – Exemplo de dados gerados pelo GeCo

rec-number	gender	given-name	surname	postcode	city	telephone-number
rec-00-org	female	katelyn	krzyszton	2604	sydney	07 4614 9969
rec-00-dup-0	female	katelyn	krzyszdon	missing	sydney	
rec-00-dup-1	female	kaelyn	krzywzton	26p4	sydney	
rec-01-org	male	ruby	alderson	2914	sydney	02 5091 1848
rec-01-dup-0		rupy	alderson	291e	sysney	
rec-01-dup-1		rupy	alderson	2914	sydneyg	
rec-02-org	female	katherine	sukwatthananan	2906	perth	08 7387 6792
rec-02-dup-0		kadherine	jukwatthananan	29o6		08 7387 6792
rec-03-org	female	alexandra	white	2602	canberra	02 3729 9751
rec-03-dup-0	female	alexantra	whide	2602	kanberra	
rec-03-dup-1	female	alezxandra	hite	2602		02 3729 9751
rec-03-dup-2		alexantra	whit	missing	canberra	02 3729 9751

Fonte: Christen and Vatsalan (2012, p. 2)

3 METODOLOGIA

Para o desenvolvimento deste trabalho foi utilizado como base o gerador sintético desenvolvido em (TRAN; VATSALAN; PETER, 2013), realizadas as alterações necessárias para contextualização da ferramenta para o cenário brasileiro. Para isto, foi necessário preparar os dados da entrada, uma vez que o gerador utiliza bases de frequências para gerar os dados sintéticos.

Foi feito um levantamento dos padrões australianos, adaptados para o cenário do Brasil, com ênfase na construção de nomes, um dos principais atributos utilizados nos algoritmos de linkage.

No início do desenvolvimento foram realizadas buscas de diretórios de dados sintéticos no GitHub, encontrados diversos trabalhos sobre a aplicabilidade do processo, mas todos com escopos bem distintos do problema tratado aqui. Foi tomada de decisão com base na popularidade das linguagens nesta área de desenvolvimento, cujo resultado é mostrado na Figura 3 3, onde as linguagens utilizadas no desenvolvimento deste trabalho se encontram dentre às cinco linguagens mais utilizadas na plataforma GitHub.

Figura 3 – Linguagens mais populares no github em projetos de dados sintéticos

Fonte: Github (2017)

3.1 PREPARAÇÃO DOS DADOS

O primeiro passo para geração da base sintética utilizando o GeCo é a preparação dos dados, onde é feito um levantamento dos dados que serão gerados, com base nos

algoritmos de record linkage desenvolvidos no laboratório LinkDataPop (COELI, 2015). Através dos estudos desenvolvidos e técnicas aplicadas, foi possível identificar que as principais informações para a construção dos perfis, nos cenários desejáveis:

a) Nome completo do paciente;

b) Nome da Mãe do paciente;

c) Data de nascimento do paciente;

d) Endereço do paciente.

Esse estudo foi focado na geração de dados sintéticos semelhantes aos existentes nas bases SINASC (Sistema de Informações sobre Nascidos vivos) e SIM (Sistema de Informações sobre Mortalidade) do estado do Rio de Janeiro. Estas bases foram escolhidas, pois tem grande histórico de tempo, possuindo registros de nascimento/mortalidade, a partir de 1910.

O acesso às bases foi realizada no laboratório LinkDataPop. Sendo essa atividade parte de projeto aprovado pelo CEP do IESC/UFRJ (CAAE: 60094116.4.0000.5286).

Além disso, foram utilizadas em diversos estudos no laboratório, cujas pesquisas realizaram diversos tratamentos nos dados (COELI C.M., 2021), resultando em uma base sem muitos valores faltantes e sem erros de importação. As bases estão disponíveis no banco de dados PostgreSQL, em rede restrita, segura e interna ao laboratório, tornando fácil a extração de informações e a utilização de consultas SQL (*Structured Query Language*), tais como as frequências de nomes necessárias para a criação dos atributos pelo gerador.

Para gerar as informações selecionadas para a construção de perfis, foi necessário pré-processar os nomes, tanto do paciente quanto da mãe, em campos de nome e sobrenome, devido à estrutura de construção do gerador.

3.2 PADRÃO AUSTRALIANO VS. PADRÃO BRASILEIRO

Diferentemente do padrão australiano, onde os nomes nas bases de dados são compostos de "*given name*"e "*surname*", o padrão brasileiro não possui limitação na quantidade de sobrenomes de uma pessoa. A geração de mais de um sobrenome foi facilmente resolvido com a própria lógica do GeCo, bastando utilizar a mesma tabela de frequência de sobrenomes, criando três campos de sobrenomes diferentes (Figura 4), chamados "*surname*", "*ssurname*"(*second surname*) e "*tsurname*"(*third surname*).

Foi mantido o padrão de desenvolvimento do código em inglês, para aumentar o alcance da documentação e utilização do trabalho, visto que o inglês é uma das línguas mais faladas no mundo, e um dos objetivos do trabalho é disponibilizar o gerador de bases sintéticas, incentivando a utilização do mesmo em pesquisas. Esta formação de nomes

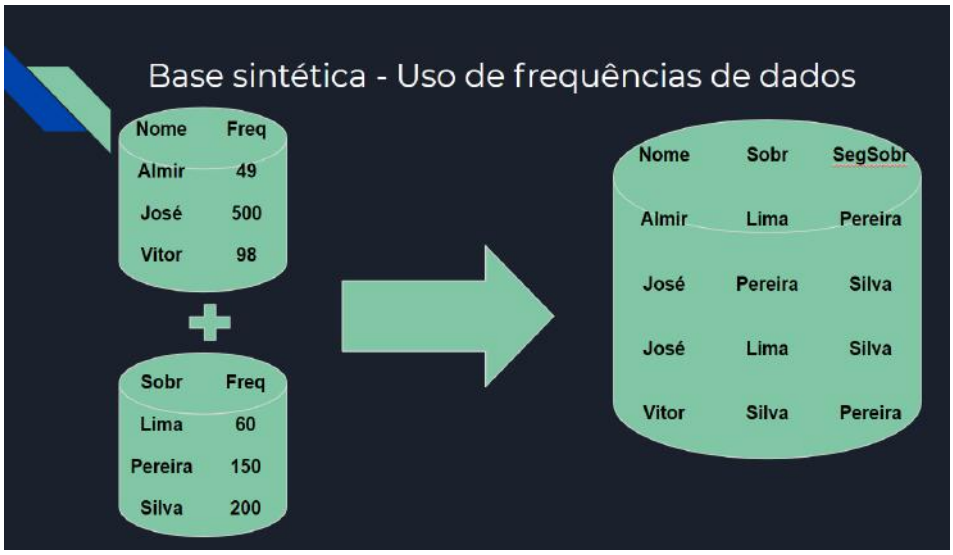
Quadro 1 – Comparativo colunas geradas

AUSTRALIA	BRASIL
given name	nome
surname	primeiro sobrenome
age	segundo sobrenome
salary	terceiro sobrenome
birth date	data de nascimento
gender	nome da mãe
city	primeiro sobrenome mãe
income	segundo sobrenome mãe
credit card	terceiro sobrenome mãe
blood pressure	código de município
	código de bairro

com múltiplos sobrenomes, por exemplo, pode ser encontrado também em outros países da América Latina.

Outra decisão foi a de gerar até no máximo três sobrenomes, pois ao contabilizar a quantidade de sobrenomes nas bases de natalidade e mortalidade, notou-se que a maioria expressiva estava contida entre um e três sobrenomes. Durante o desenvolvimento do gerador para o contexto brasileiro, toda a parametrização e programação foi feita de tal forma que adicionar mais sobrenomes é uma tarefa que pode ser desempenhada sem muito esforço, buscando a mesma customização estabelecida pelo GeCo.

Figura 4 – Utilização de frequências para construção de nomes



Fonte: Trentin (2018, p. 8)

É importante destacar que, diferentemente do GeCo, não foram geradas informações de gênero nos atributos levantados para o gerador de dados sintéticos, pois no cenário

Tabela 2 – Dados gerados com sufixos nominais

rec-id	given-name	surname	ssurname	tsurname
rec-00-org	antoniera	santos	albernaz	rodrigues
rec-01-org	josef	silverio	andrade	cunha
rec-02-org	angelica	rabello	conceicao	pinto
rec-03-org	merandulina	cardoso	telles	silva
rec-04-org	teresinha	**junior**	campos	gonzaga
rec-05-org	margareti	**filho**	silva	motta

brasileiro esta informação não possuía grande importância para a construção de perfis nos algoritmos de *linkage*. E para a geração de nomes femininos e masculinos, as bases de frequência de nomes masculinos e femininos foram unificadas, já que a própria distribuição de frequências encontradas nas bases são responsáveis por reproduzir a distribuição de gêneros da população.

Comparando as informações utilizadas pelo GeCo com o cenário brasileiro, foi possível ver que as bases de saúde pública australianas possuem informações muito mais específicas sobre os pacientes do que no Brasil, tais como: salário, numero do cartão de crédito e pressão arterial, dentre outras, necessário por parte do GeCo criar funções específicas para a geração destas informações em suas bases sintéticas. Estas funções não foram utilizadas durante o desenvolvimento deste trabalho, por serem específicas e não utilizadas nas tabelas do SUS.

Primeiramente, todas as funções de corrompimento do gerador GeCo foram desabilitadas, focando apenas na geração dos dados "originais"sem variações, para uma primeira avaliação do comportamento do gerador com as bases no cenário brasileiro, utilizando somente dos campos de: *given-name, surname, ssurname, tsurname*. Analisando os resultados obtidos com este cenário (Tabela 2), nos deparamos com uma grande incidência de sufixos nominais:

Isto ocorreu devido a grande frequência com que estes nomes aparecem na base de dados, pois eles não estão atrelados a uma árvore genealógica, podem sempre ser adicionados ao nome de qualquer família. Porém, neste caso, isto não pode ocorrer, pois, estes sobrenomes só podem aparecer como último sobrenome da pessoa, visto que o filho ou o neto tem o mesmo nome que seu predecessor, somente é adicionado o sufixo para diferenciação. Optou-se então pela remoção destes casos especiais, pois eles apresentavam baixa representatividade na construção de perfis nos trabalhos desenvolvidos no laboratório, já que as bases utilizadas no trabalho não apresentam informação de nome do pai, somente o nome da mãe.

Outra base de frequências foi gerada removendo sufixos nominais (Tabela 3), e após análise empírica dos dados gerados o resultado pareceu promissor. Em seguida, foram geradas as outras colunas, desta vez formando o nome completo do indivíduo e o nome

Tabela 3 – Nomes desconexos

rec-id	given-name	surname	ssurname	tsurname	mae-given-name	mae-surname	mae-ssurname	mae-tsurname
rec-00-org	WILSON	oliveira	pereira		LUCINDA	silveira	castro	abreu
rec-99-org	ALMIR	**silva**	**silva**		MARIA	santos	soares	fernandes

completo da mãe. Entretanto, foi possível notar que os nomes eram completamente desconexos em todos os casos, de forma que o nome do filho(a) não era composto pelo nome da mãe, como geralmente é no Brasil, onde o nome do filho(a) é uma junção dos sobrenomes da mãe e do pai. Existem casos em que o nome do filho(a), herda somente os sobrenomes do pai, e estes casos estariam sendo englobados nesta situação. Além disso, identificou-se a necessidade de modelar os demais casos de composição dos nomes e construção de um padrão brasileiro de nomes para geração dos dados.

3.3 PADRÃO BRASILEIRO DE NOMES

Devido a não limitação na quantidade de sobrenomes no Brasil, foram identificados alguns padrões culturalmente implementados na construção do nome de um indivíduo recém-nascido, buscando modelar as relações entre os nomes do filho(a), mãe e pai, de modo a aproximar da realidade brasileira. Para isso, os dados foram gerados para criar as relações que possibilitem identificar as composições de nomes de mãe e filho(a).

Para facilitar a identificação da herança dos nomes e construção dos padrões, foram utilizadas siglas para representar a herança de um sobrenome de cada lado parental, da seguinte forma:

a) N - Primeiro Nome;

b) NM - Sobrenome herdado da mãe;

c) NP - Sobrenome herdado do pai;

d) NpM - Sobrenome do pai da mãe;

e) NpP - Sobrenome do pai do pai;

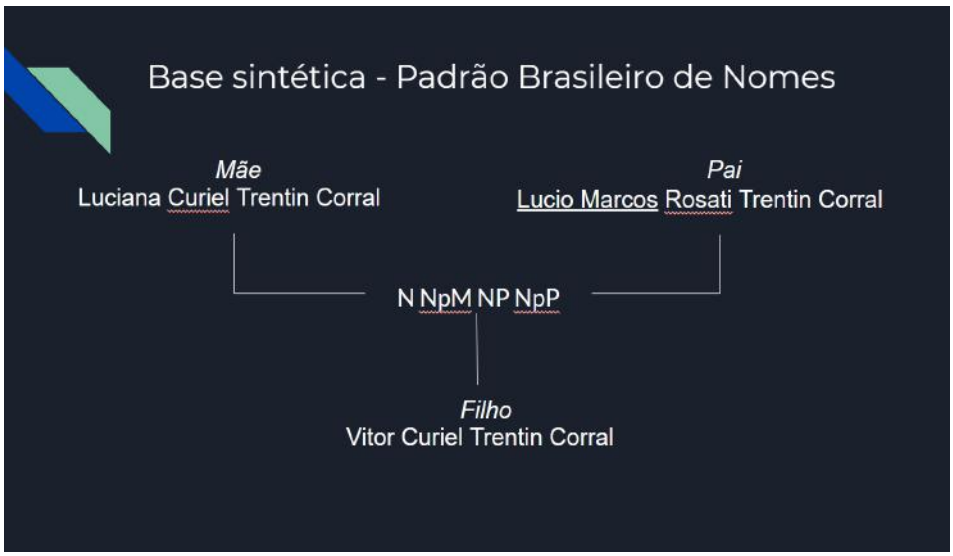
Os sobrenomes herdados do pai da mãe/pai foram separados, pois, culturalmente no Brasil, pelo histórico de patriarquia, é comum que seja priorizada a passagem do sobrenome do avô para seus netos, necessitando de uma modelagem para estes casos.

Tomando a construção do meu nome como exemplo, o padrão de construção seria N NpM NP NpP, então o gerador tem que ser capaz de gerar o primeiro nome do filho e o nome completo da mãe e do pai. Através do padrão de construção de nome recebido, formar o nome completo do filho, utilizando o primeiro sobrenome da mãe e os dois sobrenomes do pai, gerando assim um nome completo relacionado com a mãe e com o pai.

Interessante notar que devido ao casamento, minha mãe possui dois nomes vindo da família de meu pai, então poderíamos obter a mesma construção de nome através de outros padrões, como, por exemplo: N NpM NM NM e N NM NM NM, ou seja, o gerador tem que criar sobreposições nas construções dos nomes utilizando as siglas propostas.

Durante o desenvolvimento foi verificado que não existia necessidade de gerar o nome do pai separadamente, já que após a geração dos sobrenomes o mesmo não aparece na base final. Por isso, optou-se por gerar dois nomes completos e relacioná-los com um dos padrões de construção de nomes brasileiros, para substituir os sobrenomes do filho por alguns da mãe e os demais que não fossem substituídos representariam os sobrenomes do pai (Tabela 4). Desta forma, a geração de dados pelo gerador seria mais simples, reduzindo o tamanho da base e deixando mais fácil a implementação de novos padrões de construção de nomes.

Figura 5 – Construção nome filho



Fonte: Trentin (2018, p. 15)

O exemplo da construção também serve para exemplificar outro caso especial, o de nomes compostos. Inicialmente, estes casos foram tratados como sobrenomes, aos quais foram adicionadas as frequências que eles aparecem nas bases reais, substituindo o primeiro sobrenome do registro. Entretanto, isto acarretaria uma complexidade na implementação dos padrões de nome, podendo o nome do(a) filho(a) não herdar o primeiro sobrenome, caso fosse um nome composto.

A solução encontrada foi realizar um pré-processamento nas bases de natalidade e mortalidade, de forma que o nome (N), caso fosse composto, viesse na base de frequências, não sendo necessário nenhum tratamento ou regra extra, já que o nome composto não é utilizado no padrão de formação de nomes.

Tabela 4 – Construção de nomes (N NpM NP*)

Pai				Mãe			
1 José	2 Alberto	3 Carlos	4	5 Maria	6 Martins	7 Souza	8
Filho							
1 José	7 Souza	3 Carlos					

Com esta nova abordagem, sem a necessidade de gerar o nome do pai separadamente, NpP e NP podem ser agrupados em uma nova categoria, NP* onde, "*"representa aleatoriedade, dado que os sobrenomes herdados por parte de pai não são exibidos na base de dados final, não existe diferença entre selecionarmos o sobrenome por parte paterna ou materna do pai, pois a relação não será exibida no resultado final, logo, indiferente. Utilizamos o nome do gerador e simplesmente ajustamos e conectamos o nome do filho(a) ao nome da mãe para que representem a forma de construção tradicional brasileira. Após estas descobertas, os padrões de nomes foram simplificados da seguinte forma:

a) N NpM NP*;

b) N NP* NP*;

c) N NP*;

d) N NpM NP* NP*;

e) N NM NM;

f) N NM (ultimo sobrenome da mãe)

Os resultados encontrados até aqui, na utilização do GeCo, com a implementação do gerador em padrões brasileiros foi apresentada em (CORRAL, 2017), recebendo menção honrosa e nota máxima pela apresentação e pela originalidade do trabalho.

3.4 AMBIENTE DE DESENVOLVIMENTO EM C

Para o desenvolvimento do gerador de dados em padrões brasileiros, optou-se pelo desenvolvimento do mesmo em linguagem C, por ser mais familiar. Devido à simplicidade da linguagem, durante a importação de grandes arquivos de dados, ocorreram diversos erros de "*Stack Smashing*"e erros no gerenciamento dos separadores dos arquivos, necessário desenvolver o gerador para desconsiderar os formatadores de textos.

Para resolver este problema, o programa foi refeito, de modo a manter os nomes com vírgulas, e, dessa forma, mantém sempre os nomes formatados, porém, cada item do nome continua separado pelas vírgulas. Com isso, a linguagem de programação foi mudada para o C++, melhorando a forma de tratamento das *strings* pelo gerador.

Com a utilização de *strings*, foi necessário utilizar as bibliotecas "*iostream*"e "*fstream*"de manipulação de arquivos, para extrair os dados do arquivo "CSV", tomando como base a rotina descrita no código 3.1:

Código 3.1 – Importação de arquivos

```
// basic file operations
include <iostream>
include <fstream>
using namespace std;

int main () {
  ofstream myfile;
  myfile.open ("example.txt");
  myfile << "Writing this to a file.\n";
  myfile.close();
  return 0;
}
```

A partir da decisão de um nome sempre vir seguido da vírgula e com a utilização das bibliotecas da linguagem C++, manipular os nomes e exibi-los corretamente se tornou muito mais fácil.

Dado a não utilização de bibliotecas para leitura dos arquivos em formato *CSV*, a função "extraiNome"no gerador facilitou muito o processo de leitura dos nomes para o gerador brasileiro, responsável pela leitura e separação dos campos de nomes extraídos do *CSV*. A rotina recebe como argumento a linha que foi lida do arquivo, e retorna um trecho dele segmentado pela quantidade de vírgulas que separam os campos da linha. Função esta que foi futuramente trocada por bibliotecas de leitura CSV em python, como veremos no próximo item, que realizam o mesmo trabalho de forma otimizada e com implementação mais simples.

Código 3.2 – extraiNome

```
string extraiNome(string linha, int casaInicio, int casaFinal){
  int virgulas = 0;
  int countNomeDesejado =0;
  string nomeDesejado;
  for(int i =0; i < linha.size() ; i++){
    if(virgulas >= casaInicio && virgulas <= casaFinal){
      nomeDesejado += linha[i];
      countNomeDesejado++;
    }else if(virgulas > casaFinal){
      break;
    }
    if(linha[i] == ','){
      virgulas++;

    }
  }
  return nomeDesejado;
}
```

A troca de bibliotecas e adaptação do programa para C++ teve ótimas consequências, o código do programa se tornou mais legível e entendível, o código é composto de menos linhas, e a utilização da função "extraiNome"facilitou a execução do gerador para entradas que possuem muitos dados junto ao nome, visto que esses dados são extraídos e copiados em um bloco, tornando o programa mais customizável e prático.

A linguagem C++ é conhecida pela sua velocidade e desempenho de execução, no entanto, há um custo por ser uma linguagem de baixo nível. Durante o desenvolvimento deste trabalho, ocorreram várias dificuldades e erros envolvendo "questões simples"de implementação, tais como importação de arquivos ".CSV"e tratamento dos mesmos, necessário desenvolver funções para tratamento de separadores e para alocação dinâmica de memória.

Entretanto, esta limitação de arquitetura tornou-se um empecilho para a continuidade do trabalho, visto que o desenvolvimento e manutenção de novas funcionalidades no gerador tomava muito tempo.

Este trabalho foi apresentado em (CORRAL, 2017), e um dos avaliadores da banca perguntou: "Porque não desenvolver todas as novas funcionalidades diretamente em Python?". Diante deste questionamento, o processo de desenvolvimento do gerador foi reavaliado e sendo feita a mudança completa de arquitetura, reescrevendo todo código em Python.

3.5 AMBIENTE DE DESENVOLVIMENTO PYTHON

Com a mudança de linguagem, poderíamos utilizar de base algumas boas práticas e funcionalidades desenvolvidas no GeCo, centralizando todas as bibliotecas em, um projeto só, não sendo necessária escrita de arquivos intermediários para a transferência de informação entre o GeCO e o gerador brasileiro, como era necessário anteriormente. O maior ganho de todos, importação direta dos módulos do GeCO para utilização completamente integrada dentro do Python entre às duas ferramentas, tornando totalmente transparente a integração entre as ferramentas e suas funções desenvolvidas.

Após estudos e alguns testes com bons resultados, foi realizada a transcrição código de C++ para Python, o que resultou em uma diminuição na quantidade de linhas do código, de mais de trezentas linhas para menos de cem linhas. Em relação à redução de desempenho, esperada durante a mudança, não houve diferença aparente.

Isto se deve ao fato de que a arquitetura anterior requeria escrita de dados parciais em memória, e também devido ao tempo de escrita e posterior leitura para continuação do processamento, tarefa que não foi mais realizada na nova versão. A utilização da linguagem Python trouxe uma série de pequenas facilidades que aumentaram muito a produtividade do desenvolvimento gerador e posterior manutenção.

3.6 REPLICANDO DADOS

A replicação de dados corresponde a geração de uma linha com um registro "original"e, ainda que sintético, representa o registro completo, e, a partir dele são geradas duplicatas com variações de formatação, erros e valores faltantes, de modo a simular o cenário real.

Neste trabalho foi utilizada como base a estrutura de replicação do GeCo para gerar os dados duplicados, de forma que após a geração das duplicatas e erros, foi aplicado um pós-processamento para a padronização dos nomes adaptados para o contexto brasileiro.

3.7 CORROMPENDO A BASE

Agora a geração das informações e nomes em padrões brasileiros e a geração das duplicatas dos registros, a base foi corrompida com erros normalmente encontrados nas bases reais, de modo a criar diferenças entre as linhas geradas, dificultando a identificação de padrões entre os registros e corroborando para os tratamentos dos dados que o processo de *linkage* efetua durante a classificação dos pares.

Para isso, alguns tipos de erros que o GeCo aplica sobre os dados foram analisados e selecionados conforme a realidade do contexto brasileiro:

a) CorruptMissingValue:

- Esta função troca qualquer atributo por um valor em branco, podemos utilizar esta informação em quase todos os atributos, visto que as bases reais realmente podem sofrer perdas de dados durante as coletas e/ou processamentos. Uma ocorrência de dado faltante é quando uma pessoa não fornece seu nome completo, podendo omitir os nomes do meio, por exemplo;

b) CorruptValueEdit:

- Esta função substitui um caractere ou um grupo de caracteres por outros do mesmo agrupamento (letras, dígitos ou ambos) ou troca a ordem dos mesmos, utilizada para simular casos onde ocorre corrompimento na base e erros de tratamento;

c) CorruptValueKeyboard:

- Esta funcionalidade se utiliza do *layout* do teclado para simular erros de digitações das informações. Algumas bases de dados do SUS sobre natalidade e/ou mortalidade possuem coleta manual de informações, e, para esses casos foi necessária uma adaptação do *layout* teclado utilizado, tendo em vista que o padrão australiano não possui o caractere "ç"e nem acentuações. Para o contexto brasileiro de nomes, a implementação desta função foi alterada para contemplar esses casos;

d) CorruptValueOCR:

- Simula erros de leitura de arquivo e reconhecimento de imagem (exemplo: I e l, j e i). Este erro, teoricamente, não se aplica ao cenário brasileiro, pois as bases não passam por este processo programático. Entretanto, ele é muito interessante para replicar cenários os quais os dados são inseridos manualmente por um usuário que está sujeito a erros visuais (lendo informações escritas e/ou em uma tela), confundindo os caracteres durante a leitura/inserção dos dados. Este mesmo problema relacionado a similaridade visual dos caracteres pode afetar algoritmos de OCR, os quais também buscam coletar dados via imagem;

e) CorruptValuePhonetic:

- Alterações fonéticas replicam cenários onde a informação é transmitida verbalmente. Este caso é de extrema importância, principalmente, pois a língua portuguesa possui diversas similaridades fonéticas amplamente reconhecidas (ex: s, ss, ç, c, dentre outros). A utilização desta função não é simples, pois utiliza um arquivo que contém essas relações e indica onde podem ocorrer. Um exemplo é a sequência: "START,vic,vi,"que, neste caso, palavras que começam ("START"do inglês, começo) com "vic"podem ter as três letras alteradas por "vi". Nota-se que, para aproximar o máximo possível da realidade, é necessário um estudo

fonético detalhado, inserindo os casos mais comuns e identificados no arquivo utilizado pela função;

f) CorruptCategoricalValue:

- Alteração de valores com base em uma tabela de relações podem replicar trocas de valores comuns, por exemplo, dois botões próximos em uma interface para inserir os dados. Neste caso, é necessário criar uma relação entre esses dois valores, já que o usuário pode errar ao clicar na tela. Este erro não é aplicável ao cenário brasileiro, pois não foi mapeado nenhum caso onde isso ocorre ou onde seria relevante tal utilização;

g) SurnameMisspellCorruptor:

- O GeCo possui uma base com relações entre sobrenomes que poderiam ser escritos erradamente. Uma coluna identificando a escrita correta e outra com a variação de erros na escrita. Entretanto, os nomes contidos não condizem com o contexto brasileiro e a criação de uma base similar requer uma padronização de sobrenomes e a construção do que seria "o correto". O mesmo padrão é utilizado no GeCo para gerar variações de nomes.

Importante dizer que esta avaliação de contexto não impede a utilização dos erros para testes de estresse nos algoritmos de *linkage* e para avaliar o comportamento dos mesmos a diferentes categorias de erro. Entretanto, neste trabalho foram utilizados somente os erros que se aproximaram mais da realidade brasileira.

O trabalho realizado após a adaptação e a implementação das rotinas geradoras de erros na base foi apresentado em (CORRAL, 2018a), recebendo menção honrosa e nota máxima pela apresentação. Além disso, esse trabalho foi apresentado em Banff Canadá (CORRAL, 2018b).

4 RESULTADOS

Um conjunto de registros que formam um perfil (Tabela 5), obtidos pela execução do gerador é composto pelas seguintes colunas:

a) rec-gBR, identificador gerado pelo gerador de dados em padrão brasileiro (Ex: RT:5-MS:2-FS:4):
   - RT:#,id da rotina de construção de nomes foi utilizada, a fim de facilitar o reconhecimento e a manutenção em caso de erros;
   - MS:#,sobrenome da mãe foi utilizado;
   - FS:#,sobrenome da pai foi utilizado;

b) rec-id - identificador gerado pelo GeCo (Ex: rec-01-dup-0):
   - rec-##,id do registro, um registro possui de 0 a N duplicatas, sendo N um parâmetro configurável;
   - dup-##,id da duplicata;
   - rec-01-org, registro original e completo, sem corrompimento;

c) given-name - primeiro nome do filho;

d) surname - primeiro sobrenome do filho;

e) ssurname - segundo sobrenome do filho;

f) turname - terceiro sobrenome do filho;

g) dataNasc - data nascimento do filho;

h) codMunicipio - código do município do endereço;

i) codBairro - código do bairro do endereço;

j) mother-given-name - primeiro nome da mãe;

k) mother-surname - primeiro sobrenome da mãe;

l) mother-ssurname - segundo sobrenome da mãe;

m) mother-tsurname - terceiro sobrenome da mãe;

As linhas dos registros "originais", servem de base para avaliação e comparação com as duplicatas. Antes da execução das rotinas de linkage, os registros originais são filtrados, uma vez que os mesmos representam cenários irreais, pouco prováveis, onde todas as informações aparecem sem qualquer erro, o que facilitaria muito o processo de construção dos perfis. É plausível considerar a utilização das linhas "org"para algoritmos de linkage

Tabela 5 – Exemplo de duplicatas geradas com base no registro original

rec-gBR	rec-id	given-name	surname	ssurname	tsurname	dataNasc	codMunicipio	codBairro	mother-given-name	mother-surname	mother-ssurname	mother-tsurname
RT:5-MS:2-FS:4	rec-01-dup-0	ALMIR	borkes			10041937	33045r	16	MARIA HELENA	borkes	santana	jantos
RT:2-MS:4-FS:2	rec-01-dup-1	ALMR	silva			100419573		16	MARIA HELENA	borkes	sandana	zantos
RT:1-MS:2-FS:4	rec-01-dup-2	ALMIR	silga	jilva		10041973	3,30E+57		MARIA HELENA	borkes	santana	jantos
RT:2-MS:4-FS:4	rec-01-dup-3	ALMIR	alcandara			17041973		1t	MARIA	beremger	fieira	barroj
RT:1-MS:3-FS:2	rec-01-org	ALMIR	silva	silva		10041973	330455	16	MARIA HELENA	borges	santana	santos

com baixa maturidade, uma vez que o mesmo deveria ser capaz de, pelo menos, identificar perfis em casos sem erro.

Na (Tabela 5) é possível identificar diferenças entre as duplicatas, resultado da composição dos nomes e do corrompimento dos dados. A primeira linha contém os seguintes erros: o surname com um caractere errado, a dataNasc possui um dígito faltando, no codMunicipio o último dígito "5"foi substituído por "r"devido a proximidade no teclado, o mother-surname apresenta o mesmo erro de caracteres, assim como o mother-tsurname.

A segunda linha contém dois campos com missing values (ssurname e codMunicipio) e dois erros de escrita em mother-surname e mother-tsurname. A terceira linha apresenta um erro de proximidade do teclado nos campos surname e codBairro, missing values nos campos codMunicipio, mother-surname, mother-tsurname, e troca de caracteres no codMunicipio.

Na quarta linha, o filho recebe um sobrenome do pai que não aparece na base, contém erro de valores faltantes para codMunicipio, proximidade do teclado para codBairro, remoção de texto do nome composto da mãe, troca de valor do atributo de mother-ssurname e erro de escrita em mother-surname e mother-tsurname.

Durante a execução do processo de criação da base sintética é possível acompanhar detalhadamente a geração e o corrompimento dos erros pelos logs dos geradores, de forma a identificar mais corretamente qual erro foi aplicado em cada atributo, uma vez que a grande quantidade de erros modelados pode causar incerteza em relação a qual função foi aplicada.

As frequências extraídas das bases do SUS e a frequência dos dados gerados são apresentadas nas figuras (Fig. 6) (Figura 7) respectivamente. Neste exemplo, é possível ver que 7 dos 10 nomes que aparecem com maior frequência na base SINASC-SUS estão presentes nos dados gerados pelo gerador em padrões brasileiros, em sua grande maioria seguindo a mesma ordem. O resultado mostra que mesmo a volume de dado sendo menor, a proporção de nomes gerados é a mesma, aproximando os dados gerados da realidade. Importante ressaltar que, para esta categoria de análise foram filtradas somente as linhas de dados gerados para registros do tipo "org", correspondentes aos registros originais, não contendo incidências de erros, afetando a distribuição das frequências, visto que o mesmo nome poderia aparecer escrito de diversas formas diferentes.

Figura 6 – Frequência de nomes próprios extraídos da base SUS

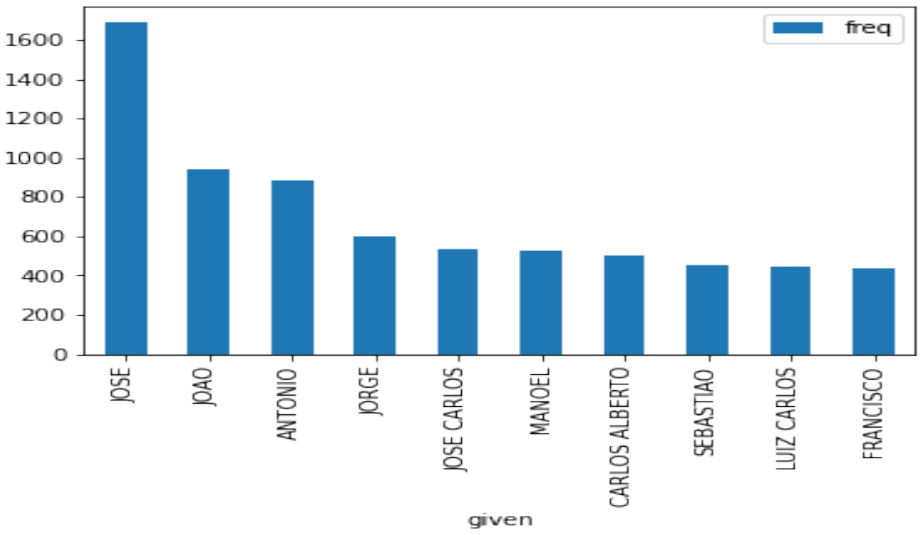
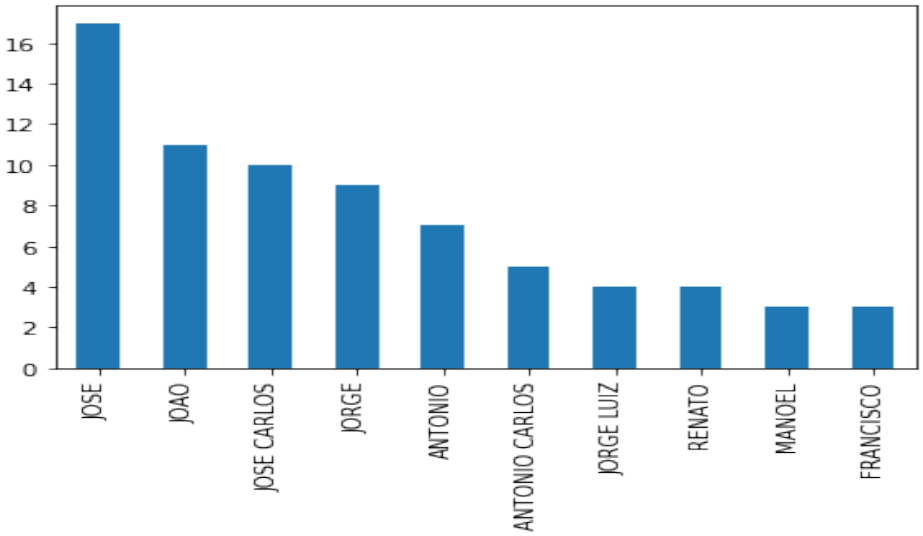


Figura 7 – Frequência de nomes próprios obtida após a execução do gerador em padrões em brasileiro



5 CONCLUSÃO

A partir do problema apresentado no início deste trabalho, é possível dizer que o objetivo de construir uma base sintética para padrões brasileiros, mantendo a confidencialidade e a segurança dos dados originais foi cumprido de forma satisfatória. A base sintética gerada mantém os atributos de identificação dos dados, possibilitando a utilização dos mesmos para correção e melhorias nos algoritmos de record linkage.

Uma vez que os dados gerados não possuem relações diretas com os dados originais, somente os dados agrupados e suas frequências foram utilizados, mantendo as colunas de rec-gBR e rec-id, preservando as relações de um perfil. Empiricamente, é possível afirmar também que os dados gerados, principalmente os nomes, se assemelham aos nomes encontrados no Brasil. E por fim, a base sintética mantém a relação direta com os dados reais em função das distribuições de frequência dos nomes e atributos que acompanham a distribuição das frequências dos dados reais. O código utilizando neste projeto é aberto e está disponível em (CORRAL, 2021).

Como próximos passos deste trabalho é interessante o estudo da utilização do gerador em diversos algoritmos de *linkage*, analisando o aumento do desempenho dos mesmos e as novas modelagens caso sejam necessárias. O gerador pode ser utilizado também para construção de um *rank* de comparação entre algoritmos e técnicas de *record linkage*, visto que todos os algoritmos utilizam a mesma base para comparação e o desempenho pode ser medido através do uso dos identificadores únicos.

Uma possível melhoria pode ser a expansão dos padrões de nomes construídos no gerador, englobando nomes em formatos estrangeiros de imigrantes ou descendentes. Isto pode ser feito através da implementação de novos padrões da mesma forma que visto neste trabalho, ou adicionando fatores e composições aleatórias na formação dos nomes, possibilitando uma maior quantidade de formações diferentes.

Outra sugestão é fazer a transformação da ferramenta para utilização de uma biblioteca Python chamada Pandas, a qual é completamente focada na manipulação e utilização de dados, sendo completamente integrada com bibliotecas matemáticas como *Numpy* e bibliotecas gráficas como Matplotlib, as quais facilitam o desenvolvimento de novas funcionalidades.

Uma opção interessante, buscando maior desempenho do gerador, seria a adaptação dos geradores para ferramentas com estruturas mais paralelizáveis, como Hadoop ou *Spark*, visto que a arquitetura destas ferramentas foram construídas para lidar com enormes quantidades de dados, e diferentemente do python, possuem uma sintaxe que se aproxima das consultas SQL, facilitando a manipulação e o trabalho com os dados. Esse é um passo além da utilização da biblioteca Pandas do Python, que busca esta aproximação e facilidade na manipulação de dados, porém não possui um bom desempenho.

Durante o desenvolvimento do trabalho, foi cogitada também a construção de uma aplicação web, onde os usuários poderiam utilizar o gerador de dados através de um auto-serviço, possibilitando um alcance maior da ferramenta e dos dados gerados, sendo necessário criar uma API para realizar a comunicação entre o gerador e o *front-end web*.

REFERÊNCIAS

CHRISTEN, P.; CHURCHES, T.; HEGLAND, M. **Febrl – A Parallel Open Source Data Linkage System**. [S.l.]: Springer, Berlin, Heidelberg, 2004. v. 01.

COELI, C. M. A qualidade do linkage de dados precisa de mais atenção. **Cadernos de Saúde Pública [online]**, v. 31, n. 7, 2015.

COELI CLáUDIA MEDINA, P. R. S. e. C. K. R. Conquistas e desafios para o emprego das técnicas de record linkage na pesquisa e avaliação em saúde no brasil. **Epidemiologia e Serviços de Saúde [online]**, v. 24, n. 4, 2012.

COELI C.M., S. V. M. P. Record linkage under suboptimal conditions for data-intensive evaluation of primary care in rio de janeiro, brazil. **BMC Med Inform Decis Mak**, v. 21, n. 190, 2021.

CORRAL, V. C. T. **DESENVOLVIMENTO DE GERADOR DE DADOS SINTÉTICOS PARA TESTES DE ROTINAS DE RECORD LINKAGE PARA O CONTEXTO BRASILEIRO**. 2017. <http://sistemas.macae.ufrj.br/8siac/arquivo/Caderno_de_resumos_CCMN_2017.pdf>. [Online; acessado 28-Julho-2021].

CORRAL, V. C. T. **DESENVOLVIMENTO DE GERADOR DE DADOS SINTÉTICOS PARA TESTES DE ROTINAS DE RECORD LINKAGE PARA O CONTEXTO BRASILEIRO**. 2018. <http://sistemas.macae.ufrj.br/9siac/cadernoController/gerarCadernoResumo/31000000>. [Online; acessado 28-Julho-2021].

CORRAL, V. C. T. **Synthetic data generator for testing record linkage routines in Brazil**. 2018. <https://ipdln.org/sites/default/files/2018ConcurrentSessions/AT-A-GLANCE-Final.pdf>. [Online; acessado 28-Julho-2021].

CORRAL, V. C. T. **Source code para gerador em padrões brasileiros**. 2021. <https://gitlab.com/vitorcuriel/gerador-dados-sinteticos-padrao-brasileiro.git>. [Online; acessado 06-Agosto-2021].

FELLEGI, I. P.; SUNTER, A. B. A theory for record linkage. **Journal of the American Statistical Association**, v. 64, n. 328, p. 1183–1210, 1969.

NEWCOMBE HOWARD B.AND KENNEDY, J. M. Record linkage: making maximum use of the discriminating power of identifying information. **Communications of the ACM**, v. 5, n. 11, p. 563–566, 1962.

NEWCOMBE HOWARD B.AND KENNEDY, J. M.; AXFORD, S. J.; JAMES, A. P. Automatic linkage of vital records. **Science**, v. 130, n. 3381, p. 954–959, 1959.

TRAN, K.-N.; VATSALAN, D.; PETER, C. Geco: an online personal data generator and corruptor. **CIKM: Conference on Information and Knowledge Management**, v. 22, n. 13, p. 2473–2476, 2013.

TRAN, K.-N.; VATSALAN, D.; PETER, C. **ANU Online Personal Data Generator and Corruptor (GeCo)**. 2020. <https://dmm.anu.edu.au/geco/>. [Online; acessado 28-Julho-2021].

APÊNDICES

APÊNDICE A – RESUMO 8ª SIAC

O laboratório Link Data Pop trabalha com rotinas de linkage em banco de dados de saúde do SUS Rio com objetivo de desenvolver algoritmos para traçar perfis e relações entre as pessoas existentes no banco.

Porém as bases de dados populacionais brasileiras não trazem um identificador unívoco, fazendo com que a vinculação de bases se utilize de informações pessoais, dados sensíveis, e visando proteger a privacidade, o acesso a bases com identificadores pessoais somente é permitida em casos especiais. Nesse caso se faz necessário o uso de dados sintéticos para análise, desenvolvimento e estudo de algoritmos computacionais para vinculação de dados (record Linkage).

Os dados sintéticos são importantes para o teste dos algoritmos de record linkage, pois com a manipulação correta dos dados é possível fazer melhorias na implementação do algoritmo de linkage. Dessa forma são encontrados mais vínculos em dados reais. Devido a inexistência de bases sintéticas para o contexto brasileiro, este trabalho propõe o desenvolvimento de um algoritmo que gere dados sintéticos se utilizando de dados reais e suas frequências para aproximar as bases geradas da realidade, mantendo a integridade e privacidade durante o uso das informações pessoais.

O laboratório Link Data Pop, do Instituto de Estudos de Saúde Coletiva (IESC/UFRJ) trabalha com rotinas de linkage em bancos de dados de saúde do SUS Rio, que são utilizados em processos de pareamento de dados. O objetivo desses algoritmos é identificar pares de registros, através da comparação de nomes, endereços e outros atributos que geralmente não serviriam como identificadores individuais, mas atendem aos critérios de classificação probabilística do algoritmo para avaliação do resultado. Porém, as bases de dados populacionais brasileiras não trazem um identificador único, fazendo com que a vinculação de bases se utilize de informações pessoais, dados sensíveis e protegidos eticamente contra divulgação a privacidade. Dessa forma, o acesso a bases com identificadores pessoais é permitida somente em casos especiais ou em ambientes seguros. Nesse caso se faz necessário o uso de dados sintéticos para análise, desenvolvimento e estudo de algoritmos computacionais para vinculação de dados (record Linkage). Os dados sintéticos são importantes para o teste dos algoritmos de record linkage, pois com a manipulação correta dos dados é possível fazer melhorias na implementação e desempenho do algoritmo de linkage. Devido a inexistência de bases sintéticas para o contexto brasileiro, este trabalho propõe o desenvolvimento de um algoritmo que gera dados sintéticos se utilizando de dados reais e suas frequências, aproximando as bases geradas da realidade, e preservando a integridade e a privacidade durante o uso das informações pessoais.

APÊNDICE B – RESUMO 9ª SIAC

O laboratório Link Data Pop, do Instituto de Estudos de Saúde Coletiva (IESC/UFRJ) trabalha com rotinas de linkage em bancos de dados de saúde do SUS Rio, que são utilizados em processos de pareamento de dados. As bases de dados populacionais brasileiras não trazem um identificador único, fazendo com que a vinculação de bases seja realizada empregando a comparação de identificadores pessoais como nomes, datas de nascimento e outros atributos que são empregados conjuntamente por algoritmos para record linkage para o cálculo de um escore que indica o quão verossímil é que dois registros pertençam à mesma entidade (em geral, o mesmo indivíduo), Um dos problemas metodológicos a ser enfrentado é a avaliação da acurácia dos algoritmos de record linkage Para essa avaliação é necessário um um padrão ouro que indique a classificação correta dos pares em falsos e verdadeiros. O padrão ouro pode ser criado por meio da revisão manual dos pares. Entretanto, esse procedimento é vulnerável a erros. Adicionalmente , visando garantir à privacidade dos indivíduos, o acesso a bases com identificadores pessoais é permitida somente em casos especiais, devendo a mesma ser realizada em ambientes seguros, o que restringe o uso de dados reais para a avaliação da qualidade dos algoritmos de linkage.

Nesse caso se faz necessário o uso de dados sintéticos. Até onde possamos saber, não existem bases sintéticas que reproduzem os padrões de nomes brasileiros. Após um ano de desenvolvimento, criamos um algoritmo que gera dados sintéticos a partir da distribuição de frequência de nomes de bases de dados reais. Inicialmente, utilizamos o GeCo (http://dlrep.org/dataset/GeCo), um gerador de dados pessoais personalizável desenvolvido por Tran et al.(DOI:10.1145/2505515.2508207). Todavia as personalizações do gerador tem seu limite perante os padrões brasileiros de dados, criando a necessidade de adaptar o gerador para o Brasil, criando rotinas que atuem em consonância com o GeCo. Inicialmente, desenvolvemos essas rotinas utilizando C, passamos à C++ até chegar em Python. Criamos um algoritmo capaz de gerar nomes de filhos e mães, emulando a construção nominal brasileira.

A próxima fase consistiu em emular os possíveis erros durante a coleta de dados e a diferença entre registros distintos de uma mesma pessoa, por meio de algoritmos de corrupção dados.

A corrupção é essencial para avaliar a eficácias das rotinas de record linkage no tratamento de erros usualmente observados nas bases reais.

APÊNDICE C – ABSTRACT IPDLN 2018

Synthetic data generator for testing record linkage routines in Brazil.

Record linkage has been increasingly used in Brazil. However, only a few studies report the quality of the linkage process. Synthetic test data can be used to evaluate the quality of data linkage.

To develop a synthetic data generator that creates test datasets with similar attributes and error characteristics found in the Brazilian databases. We analyzed the 2013 mortality database from Rio de Janeiro State to know the characteristics and frequency distribution of the database attributes (name, mother's name, sex, date of birth and address). We used Python and C++ to customize and add routines to GeCo (http://dlrep.org/dataset/GeCo), a personal data generation tool developed by Tran et al. (DOI:10.1145/2505515.2508207).

Brazilian names have specific characteristics that distinguish them from other countries' patterns: multiple family names are usual, as are composite first names, and, despite that, homonyms are frequent. Family names may include the full extension or only parts of either the father and mother's respective family names, or both, so there is a wide variation in progeny family names and not necessarily a common family name for all family members.

Due to the specific national characteristics of name building in Brazil, modeling synthetic data is particularly challenging and needs to have more flexible rules in order to generate databases that will actually allow assessing the quality of data linkage processes.