FÁBIO MARRA

MODELAGEM DE PAREAMENTO DE DADOS E ANÁLISE DESCRITIVA ENTRE DADOS DO SISTEMA PÚBLICO DE SAÚDE E DO SISTEMA PRIVADO DE SAÚDE DE BELO HORIZONTE, BRASIL

Monografia 2 focada em pesquisa tecnológica na área de Mineração de Dados, elaborada para a conclusão do curso de Sistemas de Informação da Universidade Federal de Minas Gerais.

Orientador: Wagner Meira Júnior, Ramon Gonçalves Pereira

Belo Horizonte

Abril de 2025

Resumo

Este trabalho apresenta uma abordagem para vinculação de registros com o foco na áera de saúde em relação a bases de dado do setor público e privado de saúde brasileiro, em especial da cida dede Belo Horizonte. Um dos focos principais da abordagem é na estimação automatizada dos parâmetros probabilísticos m e u do modelo de Fellegi-Sunter. Para isso, foi desenvolvido o **cPareia MU Estimator**, um algoritmo iterativo que ajusta esses parâmetros com base na proporção esperada de correspondências entre registros, utilizando um limiar dinâmico derivado dos pesos logarítmicos. Foram conduzidos experimentos com bases sintéticas representando características reais de dados de saúde e suas possíveis variações e erros, testando o desempenho do estimador proposto em comparação com ferramentas consolidadas como Splink e RecordLinkage Toolkit. Os resultados mostram que o estimador desenvolvido apresenta desempenho competitivo, especialmente em bases maiores, e representa uma alternativa viável em contextos sem dados rotulados. O estudo também destaca a importância do record linkage para a integração de dados em saúde no Brasil e propõe aprimoramentos futuros no uso de blocagem híbrida e na aplicação em bases reais.

Palavras-chave: **record linkage**, **m e u**, **Fellegi-Sunter**, **dados de saúde**, **cPareia**.

Abstract

This work presents a record linkage approach designed for future application to real datasets from the Brazilian public and private healthcare sectors, particularly in the city of Belo Horizonte. One of the main focuses of the approach is the automated estimation of the probabilistic parameters m and u from the Fellegi-Sunter model. To achieve this, the **cPareia MU Estimator** was developed—an iterative algorithm that adjusts these parameters based on the expected proportion of matches between records, using a dynamic threshold derived from log-likelihood weights. Experiments were conducted using synthetic datasets that simulate realistic characteristics, variations, and errors found in healthcare data, in order to evaluate the estimator's performance in comparison with established tools such as Splink and the RecordLinkage Toolkit. The results show that the proposed estimator performs competitively, especially on larger datasets, and represents a viable alternative in scenarios where labeled data is unavailable. The study also highlights the importance of record linkage for integrating health data in Brazil and proposes future improvements involving hybrid blocking and application to real-world datasets.

Keywords: **record linkage**, **m and u**, **Fellegi-Sunter**, **healthcare data**, **cPareia**.

Sumário

Capítulo 1

Introdução

A integração de bases de dados provenientes dos sistemas público e privado de saúde constitui um dos maiores desafios para a construção de uma política de saúde verdadeiramente integrada, eficiente e baseada em evidências no Brasil. Essa necessidade se torna ainda mais urgente diante da complexidade do sistema de saúde brasileiro, que é historicamente marcado pela fragmentação entre o Sistema Único de Saúde (SUS), responsável pelo atendimento universal e gratuito, e o setor suplementar, representado por operadoras como Unimed, Hapvida, entre outras, que atendem uma parcela significativa da população com acesso a planos privados. Além disso, iniciativas como o InfoSAS — sistema de monitoramento baseado em mineração de dados para detecção de anomalias na produção do SUS — ilustram o potencial do uso inteligente de dados para a melhoria dos serviços de saúde, mesmo diante de cenários com qualidade de dados limitada. [7]

Essa separação estrutural entre os dois sistemas não apenas dificulta o acompanhamento longitudinal de pacientes que transitam entre o setor público e o privado, mas também compromete a construção de indicadores epidemiológicos confiáveis e abrangentes. A ausência de integração de dados limita o desenvolvimento de políticas públicas eficazes, a avaliação de intervenções em saúde e a gestão racional de recursos, além de dificultar a detecção de padrões de morbidade e de iniquidades no acesso aos serviços.

Nos últimos anos, diversas iniciativas têm buscado avançar na interoperabilidade dos sistemas de informação em saúde no Brasil, sobretudo no contexto do SUS. Estudos recentes apontam avanços importantes, como a institucionalização do Conecte SUS e o uso crescente de tecnologias de Big Data e interoperabilidade semântica, mas também destacam barreiras persistentes, como a heterogeneidade nos formatos de dados, a baixa padronização de campos críticos (como nome, data de nascimento e CPF), e a

dificuldade de implantação de identificadores únicos nacionais.

Neste contexto, técnicas de vinculação de registros (*record linkage*) desempenham um papel fundamental. O *record linkage* é o processo de identificação e combinação de registros que se referem ao mesmo indivíduo, mas que estão presentes em diferentes bases de dados. As abordagens para vinculação de registros podem ser divididas, de forma geral, em duas categorias principais: o pareamento determinístico e o pareamento probabilístico. O pareamento determinístico se baseia em regras fixas de comparação, exigindo a igualdade exata entre campos específicos (como CPF, nome completo e data de nascimento) para considerar dois registros como pertencentes ao mesmo indivíduo. Embora eficiente e simples de implementar, esse método apresenta baixa tolerância a erros de digitação, campos faltantes ou variações ortográficas, o que o torna pouco eficaz em bases de dados com qualidade heterogênea. Já o pareamento probabilístico, por sua vez, incorpora medidas de incerteza e similaridade parcial entre os atributos dos registros. A principal referência nesse campo é o modelo proposto [18] que estabelece uma estrutura teórica baseada na estimativa de pesos para cada campo comparado, levando em conta a probabilidade de coincidência entre pares verdadeiros (m) e falsos (u). Esses pesos são somados para produzir uma pontuação final, permitindo classificar os pares como *matches*, *non-matches* ou *possíveis matches*, de acordo com limiares definidos. Essa abordagem é particularmente adequada para cenários com dados incompletos ou inconsistentes, como é comum em sistemas administrativos de saúde.

Em ambientes onde identificadores únicos não estão disponíveis de forma consistente, como é comum em bases administrativas de saúde no Brasil, métodos probabilísticos e baseados em aprendizado de máquina oferecem alternativas robustas para a realização dessa tarefa, ainda que envolvam desafios computacionais e estatísticos relevantes. Além disso, apesar do potencial das abordagens probabilísticas, sua eficácia depende fortemente da estimativa adequada dos parâmetros m e u, que representam, respectivamente, a probabilidade de um campo coincidir entre registros que pertencem ao mesmo indivíduo e entre registros que pertencem a indivíduos diferentes. A definição imprecisa desses parâmetros pode comprometer a acurácia do modelo, resultando em altos índices de falsos positivos ou negativos. No entanto, a estimativa desses valores nem sempre é trivial: em muitos cenários reais, especialmente em bases com dados ruidosos ou sem rótulos disponíveis, não é possível contar com um conjunto de pares previamente identificados como correspondentes ou não correspondentes. Essa limitação tem motivado o desenvolvimento de métodos automáticos de estimação, que buscam inferir valores apropriados de m e u a partir de propriedades estatísticas dos próprios dados e de suposições sobre a taxa de correspondência esperada.

O presente trabalho insere-se nesse contexto ao propor uma forma automatizada de calculo de m e u a ser acoplada em softwares existentes de pareamento bem como a modelagem de vinculação de registros entre uma base sintética representando o setor privado e os registros públicos do SUS, com foco na população da cidade de Belo Horizonte, Minas Gerais. A construção da base sintética permitiu simular cenários realistas de erro, omissão e duplicidade, frequentemente encontrados em bases reais de saúde.

Os objetivos principais deste estudo são:

- Desenvolver um algoritmo de estimação automática dos parâmetros probabilísticos m e u, fundamentais para a modelagem de record linkage segundo a metodologia de Fellegi e Sunter, com base em uma proporção de matches esperada entre os registros;

- Avaliar comparativamente o desempenho de diferentes ferramentas de record linkage — incluindo **Splink**, **RecordLinkage** e **cPareia** — com base em métricas como acurácia, número de pares gerados, número de blocos e tempo de execução;

- Identificar os desafios operacionais e metodológicos envolvidos na integração de dados entre os setores público e privado, e discutir o papel do record linkage na construção de sistemas de informação em saúde mais integrados e inteligentes.

Capítulo 2

Referencial Teórico

Esta seção descreve os conceitos principais abordados neste trabalho sobre pareamento de registros bem como informações relacionadas a registros da área da saúde, que foram utilizados neste experimento.

2.1 Pareamento de Registros

As técnicas e algoritmos de pareamento de registros seguem uma lógica específica. Primeiro se faz um **Pré Processamento de dados**, para preparação e estruturação dos mesmos. Depois é realizado a **Blocagem** que separa os dados em blocos com eliminar comparações desnecessárias. Posteriormente temos a **Estimativa de Parâmetros**, onde definimos o modelo probabilístico para o cálculo de semelhanças dos dados. Por fim, realizamos a comparação efetivamente, que determina os pesos para cada comparação nos blocos e definir o pareamento. A figura 2.1 ilustra o processo de pareamento de registros.

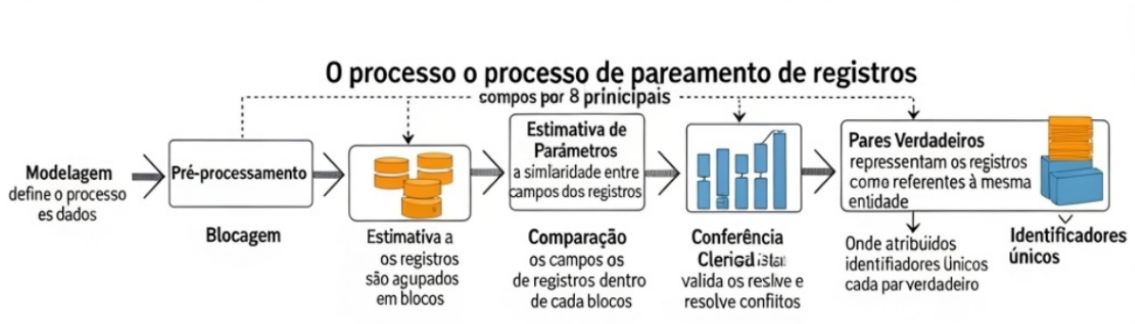


Figura 2.1. Processo de Pareamento de Registros

2.1.1 Pré Processamento

O pré-processamento é a etapa inicial responsável pela preparação dos dados que serão utilizados no processo de pareamento de registros. [6] Esta etapa é fundamental para garantir que a similaridade entre registros seja avaliada de forma justa e eficaz. Dados de diferentes bases de dados necessitam ter o mesmo formato de estrutura e de preenchimento de variável. Nesta etapa, são aplicadas técnicas de redução de variações de escrita dos dados, remoção de ruídos e inconsistências. [18]

2.1.2 Blocagem

Para parear dados na base de dados o ideal seria comparar todos os registros da base de dados e verificar a semelhança entre cada um deles. Entretanto, para base de dados muito extensas esse método, que escala de forma quadrática em relação quantidade de dados, se torna impraticavel computacionalmente.

A técnica de blocagem tem como objetivo reduzir o número total de comparações, garantindo que apenas registros com alta probabilidade de corresponder entre si sejam avaliados em profundidade. Os registros são divididos em blocos com características semelhantes, que serão verificadas *à priori*, para que as comparações ocorram internamente dentro desses blocos menores. Assim, um registro é incluído em um determinado bloco caso possua um conjunto pré-definido de *atributos* com valores idênticos — por exemplo, o *nome* ou a combinação de *nome* e *sobrenome*. Caso esses atributos iniciais não satisfaçam os critérios de igualdade, os demais campos do registro são automaticamente descartados da análise, evitando comparações desnecessárias e otimizando o desempenho do processo de vinculação. [18]

2.1.3 Estimativa de Parâmetros

Para avaliar o quão semelhante os registros comparados são, é necessário atribuir um peso a cada comparação de registros. Um modelo reconhecido é o de [18] que utiliza as probabilidades m e u para atribuir pesos a cada um dos atributos utilizados no pareamento de registros. **M**: Probabilidade de o campo coincidir dado que os registros são um match verdadeiro.

U: Probabilidade de o campo coincidir por acaso, em um par que não é match. Para cada atributo, esses parâmetros são utilizados para que um peso (nota) seja atribuída à comparação desse atributo. Esse peso é calculado como:

$$w = log2(m/u)$$

quando os atributos comparados coincidem ou:

$$w = log2(1 - m/1 - u)$$

quando eles são distintos. Dessa maneira o somatório desses pesos mensura a semelhança entre os registros comparados. Para estimar os pesos ideais para cada atributo existem algumas abordagens, a mais utilizada quando não temos os rótulos - não sabemos se as linhas realmente são iguais ou não - é a *Expectation-Maximization (EM)* que funciona utilizando os valores atuais dos parâmetros m e u - inicialmente atribui-se um valor - depois é calculada a semelhança entre os registros e por fim estipulado quais são matchs e quais não são (unmatchs), geralmente baseado num threshold. Uma vez que temos estipulados os matchs e unmatchs calculamos para cada atributo os novos parâmetros m e u. Esses passos são repetidos até a convergência. [5]

2.1.4 Comparação

Na etapa de comparação os registros são selecionados, par a par, dentro dos blocos criados, e comparados, atributo por atributo, utilizando os pesos estimados na etapa de estimativa de parâmetro. São utilizadas diferentes técnicas de comparação de atributos, como comparação exata, distância de edição com funções como Levenshtein [25] e Jaro-Winkler [31], proximidade numérica e de datas. Nos algoritmos Pareia e Record Linkage a comparação é feita de forma binária, sendo um match ou não. Porém algoritmos como o Splink abrangem a possibilidade de resultados, considerando a igualdade exata mas também possíveis variações mínimas para que casos aceitavelmente parecidos não recebam uma nota tão ruim quanto um atributo completamente diferente. Para cada variação de comparação desse algoritmo parâmetros MUs são estipulados pelo mesmo, sendo então calculada a nota de cada caso. Após a comparação atributo por atributo os pesos de cada comparação é somada e então a nota do par é atribuída e, a partir de um treshold determinado, rotulado como match ou unmatch

2.2 Ferramentas de Record Linkage

O avanço das técnicas de vinculação de registros tem sido acompanhado pelo desenvolvimento de ferramentas especializadas, que implementam abordagens probabilísticas e heurísticas para facilitar a integração de grandes volumes de dados. Entre essas ferramentas, destacam-se o **Splink** e o **RecordLinkage Toolkit**, amplamente utilizadas em projetos de ciência de dados e estatística aplicada. Por último, o **cPareia**

uma ferramenta desenvolvida a fim de processar grandes volumes de dados utilizando como core a linguagem C.

2.2.1 Splink

O *Splink* é uma biblioteca de código aberto desenvolvida pelo Ministério da Justiça do Reino Unido, com o objetivo de viabilizar a vinculação probabilística de registros em ambientes de larga escala. Baseado no modelo clássico de Fellegi-Sunter, o Splink utiliza algoritmos de Expectation-Maximization (EM) para estimar automaticamente os parâmetros probabilísticos m e u, bem como os pesos logarítmicos associados à comparação entre campos. Sua arquitetura é orientada a SQL, o que permite a execução eficiente de consultas em motores como Apache Spark e DuckDB, tornando-o adequado para processar conjuntos de dados com milhões de registros.

Além de estimar parâmetros, o Splink também realiza a blocagem dos dados com múltiplas estratégias, gera visualizações dos pesos calculados e dos histogramas de similaridade entre pares, e permite a formação de agrupamentos de registros vinculados. A ferramenta tem sido aplicada em diversos contextos, como integração de registros educacionais, judiciais e de saúde no setor público britânico, demonstrando alto desempenho mesmo em bases despadronizadas [22].

2.2.2 RecordLinkage Toolkit

O *Python RecordLinkage Toolkit*, também conhecido como `recordlinkage`, é uma biblioteca acadêmica mantida por pesquisadores da Universidade de Maastricht. Desenvolvido em Python, o pacote oferece uma implementação modular do modelo de Fellegi-Sunter, com suporte a diferentes etapas do processo de vinculação, incluindo blocagem, comparação de atributos, classificação e avaliação dos resultados. [14]

A blocagem pode ser realizada de forma indexada, canônica ou combinada, e os campos podem ser comparados por meio de funções clássicas como Levenshtein, Jaro-Winkler, Soundex e distância euclidiana. A biblioteca permite ainda a aplicação de classificadores supervisionados, como regressão logística e florestas aleatórias, além de técnicas não supervisionadas, como agrupamento baseado em limiares.

Por sua flexibilidade e facilidade de integração com estruturas como `pandas`, o `recordlinkage` é amplamente utilizado em estudos de prototipagem, especialmente em cenários com volume moderado de dados e necessidade de controle mais detalhado sobre cada etapa do processo. Sua aplicação é particularmente vantajosa em contextos

exploratórios, onde se deseja avaliar diferentes combinações de funções de similaridade e estratégias de blocagem de forma iterativa.

2.2.3 Pareia e cPareia

A dissertação de mestrado de Walter dos Santos Filho, intitulada *Algoritmo Paralelo e Eficiente para o Problema de Pareamento de Dados* [16], representa uma contribuição pioneira no contexto brasileiro ao abordar a escalabilidade e o desempenho do processo de vinculação de registros em grandes volumes de dados. O autor propõe um algoritmo paralelo com foco na eficiência computacional, utilizando técnicas de blocagem como estratégia fundamental para reduzir o número de comparações necessárias entre os registros.

A abordagem apresentada combina blocagem fonética e heurísticas de pré-processamento para agrupar registros potencialmente correspondentes em blocos menores, sobre os quais são aplicadas funções de comparação. A arquitetura paralela do algoritmo permite distribuir a carga de processamento entre múltiplos núcleos de CPU, tornando-o adequado para aplicações com milhões de registros. Além disso, o trabalho discute critérios de avaliação de desempenho com base em medidas como precisão, revocação e tempo de execução, enfatizando a importância da otimização tanto da qualidade do pareamento quanto da viabilidade computacional em cenários reais. Embora o foco principal do estudo esteja na eficiência operacional, o modelo de decisão adotado é compatível com frameworks probabilísticos inspirados na teoria de Fellegi-Sunter, o que permite uma aproximação conceitual com modelos mais modernos.

O *cPareia* é uma evolução direta do sistema PAREIA, proposto originalmente por Santos Filho [16], com o objetivo de tornar o processo de vinculação de registros ainda mais eficiente e acessível para execução em ambientes computacionais locais. Desenvolvido em linguagem C, o *cPareia* busca otimizar o desempenho de tarefas de blocagem, comparação e decisão de pareamento, mantendo a robustez metodológica da versão anterior, porém com ganhos significativos em velocidade e uso de memória.

Enquanto o PAREIA original foi concebido para execução paralela em ambientes distribuídos ou multiprocessados, o *cPareia* foi projetado para rodar de forma eficiente em máquinas standalone, o que amplia sua aplicabilidade prática em instituições com infraestrutura computacional limitada. A reescrita do código em C permitiu a eliminação de sobrecargas associadas a linguagens interpretadas e a maior controle sobre estruturas de dados e gerenciamento de memória.

2.3 Dados em Saúde no Brasil

O Sistema Único de Saúde do Brasil, o SUS, é formado pelo conjunto de todas as ações e serviços de saúde prestados por órgãos e instituições públicas federais, estaduais e municipais, da administração direta e indireta e das funções mantidas pelo Poder Público. À iniciativa privada é permitido participar desse Sistema de maneira complementar. [11] A assistência médica privada sempre participou da organização da política de saúde brasileira, com influência progressiva na oferta e na estrutura dos serviços nacionais. Esta participação teve diferentes impactos no modelo de saúde vigente no país, dentre eles, a complementariedade de serviços privados no Sistema Único de Saúde. Na atenção de média complexidade, a oferta depende fortemente do setor privado, mesmo após a implantação do SUS, devido à insuficiência da rede pública para o atendimento dos usuários, mantendo-se nesta área a lógica de compra de ações e serviços de maneira complementar. Os serviços e ações de média complexidade visam atender às demandas de saúde e agravos da população, cuja prática clínica dependa de profissionais especializados e uso de recursos tecnológicos de apoio diagnóstico e terapêutico. [15]

Historicamente, a experiência de tratamento de dados em saúde no Brasil tem sido acompanhada da implementação de múltiplos sistemas de informação, voltados para diferentes fins: epidemiológico, demográfico e de produção de serviços. No setor público, podemos citar como exemplos aqueles vinculados ao DATASUS, o Departamento de Informática do SUS: Sistema de Informação sobre Nascidos Vivos (Sinasc), Sistema de Informações sobre Agravos de Notificação (Sinan), Sistema de Informações Hospitalares (SIH), Sistema de Informação de Mortalidade (SIM), Sistema de Informação de Atenção Básica (SIAB), Sistema de Cadastramento de Usuários (CADSUS), Cadastro Nacional de Estabelecimentos de Saúde (CNES), Sistema de Informações do Programa Nacional de Imunizações (SI-PNI), Sistema de Informação do Câncer do Colo do Útero e Sistema de Informação do Câncer e Mama (SISCOLO e SISMAMA), Sistema de Cadastramento e Acompanhamento de Hipertensos e Diabéticos (HIPERDIA), Sistema de Acompanhamento da Gestante (SISPRENATAL), Sistema de Informações Ambulatoriais do SUS (SIA) entre outros [13].

Em meio a esse cenário, ao longo das últimas três décadas, atores do sistema público e privado de saúde no Brasil produziram sistemas de informação em saúde. Assim, diversos sistemas foram desenvolvidos para atender às demandas de planejamento e gestão local, da mesma maneira que foram produzidas bases de dados em saúde1. Porém, esses sistemas de informação emsaúde continuam fragmentados e o Ministério da Saúde, por inúmeras vezes,tentou contratar sistemas de registro eletrônico de saúde

(RES), sem sucesso. [20]

Algumas estratégias foram criadas com foco em interoperabilidade e integração de sistemas. A Estratégia e-SUS Atenção Básica (e-SUS AB) foi uma iniciativa do Ministério da Saúde para promover a integração dos SNIS, visando reduzir a duplicidade de registros e melhorar a qualidade das informações . No entanto, a integração completa ainda é um desafio, especialmente devido à coexistência de múltiplos sistemas que não se comunicam adequadamente [9].

A interoperabilidade de dados na saúde é regulamentada pela Portaria 2.073 do Ministério da Saúde, que estabelece padrões e normas para a troca de informações entre sistemas públicos e privados . Apesar dos benefícios potenciais, como a melhoria na continuidade do cuidado e na eficiência dos serviços, a adoção dessas práticas ainda enfrenta barreiras culturais e de investimento [10].

Assim, um trabalho que visa vincular registros de saúde de instituições com perfis diferentes (Público e Privado) se posiciona como um desafio relevante e tendo como resultado dados longitudinais com cobertura quase que completa da assistência em saúde de determinadas regiões. Permitindo a realização de estudos, implementação de novas políticas públicas de saúde e entender o comportamento da sociedade no uso de diferentes tipos de sistemas de saúde.

2.4 Trabalhos Relacionados a Vinculação de Dados

A base dos pareamentos de registro residem em métodos que permitem a união de dados de diferentes fontes, mesmo na ausência de identificadores únicos. O objetivo é identificar registros que pertencem à mesma entidade. A base teórica principal da grande maioria dos métodos é a de [18]. Entretanto, trabalhos focados em utilização de pareamento de registros, evolução de técnicas, modelos e frameworks para execução.

[19] cria um guia que oferece um framework conceitual e orientações práticas para a realização de record linkage em pesquisas de serviços de saúde. Abrange desde o planejamento e preparação de dados até a execução do pareamento e as considerações éticas e de privacidade, incluindo conformidade com regulamentações como HIPAA.

No trabalho de [29], uma abordagem bayesiana é proposta para o record linkage, modelando-o como um problema de pareamento bipartido. A metodologia bayesiana permite a quantificação de incertezas e a incorporação de conhecimento prévio, levando a estimativas de pares verdadeiros, particularmente útil com dados ruidosos ou incompletos. Isso contrasta com métodos probabilísticos mais tradicionais.

No guia da Oxford, [24], a ideia é descrever métodos para avaliar a qualidade do

record linkage. Ele detalha a determinação da precisão e do recall dos links, utilizando abordagens como a comparação com um "padrão-ouro"ou a análise de características estatísticas dos dados ligados e não ligados, o que é crucial para validar a confiabilidade dos conjuntos de dados resultantes.

Aprofundando na avaliação de qualidade, [8] foca em métodos para estimar a precisão (precision) e o recall (recall), métricas para avaliar o desempenho de algoritmos de record linkage, tanto determinísticos quanto probabilísticos. Os autores discutem o cálculo dessas métricas, mesmo na ausência de um "padrão-ouro"completo.

A preocupação com a qualidade é recorrente. [sci] faz um apelo por maior atenção à qualidade e transparência nos estudos de pareamento de registros. Ele argumenta que a validação rigorosa dos resultados do pareamento é necessária para a confiabilidade das análises, e destaca a importância de reportar a qualidade da ligação para que pesquisadores possam interpretar os dados com precisão.

Um trabalho relevante na saúde pública brasileira que utilizou a teoria de [18] é o de [21] que descreve um esforço de pareametno de registros para construir um banco de dados nacional de saúde centrado no indivíduo. Detalha a metodologia e os desafios da ligação de registros administrativos e epidemiológicos de 2000 a 2015, destacando a importância da união de dados para pesquisa e gestão em saúde pública no país.

Com o aumento do volume e da complexidade dos dados, novos desafios surgem, exigindo métodos mais sofisticados, incluindo o uso de técnicas de aprendizado de máquina e a consideração de aspectos como a privacidade dos dados. [12] revisita e aprofunda o método probabilístico de pareamento de registros. Ele discute os fundamentos teóricos, como o modelo Fellegi-Sunter, e explora maneiras de otimizar a aplicação desse método, contribuindo para uma compreensão das nuances do pareamento de registros probabilístico.

Um estudo comparativo, [17], avalia o desempenho de diferentes algoritmos de record linkage (determinísticos e probabilísticos) em um contexto de saúde pública. A pesquisa mede a eficácia e a precisão desses algoritmos na identificação de pares correspondentes, oferecendo insights sobre a adequação de métodos para diferentes cenários e qualidades de dados.

No Brasil, [27] apresenta um estudo de caso de aplicação de vinculação de registros em larga escala, utilizando dados de registro de rotina. Demonstra a aplicação do pareamento de registros para estudos de saúde pública e epidemiologia, focando na mortalidade de recém-nascidos, utilizando 17.6 milhões de registros ligados para gerar insights sobre a saúde populacional.

Vindo da ecologia, [32] embora do campo da ecologia, destaca desafios de integração de dados que são análogos aos erros de pareamento em record linkage. Os

principais pontos levantados incluem a dificuldade de ligar dados com heterogeneidade de estrutura, tipo e granularidade, a ocorrência de inconsistências e erros de registro nas fontes, e o desafio de conciliar diferenças de escala e resolução. Tais problemas podem levar a pareamentos incorretos ou perdidos, sublinhando que a integração de dados bem-sucedida exige a superação dessas barreiras de diversidade, qualidade e escala.

O trabalho de [2] destaca a crescente importância do record linkage como uma ferramenta epidemiológica e de pesquisa translacional na medicina. Tecnicamente, ele enfatiza como a integração de diversas fontes de dados de saúde – como registros de atenção primária, internações hospitalares, dados de laboratório e registros de óbito – aprimora significativamente o poder analítico e a validade externa dos estudos. Ao ligar esses dados, é possível construir coortes longitudinais mais abrangentes, investigar trajetórias de doenças em populações maiores, avaliar a efetividade de intervenções em ambientes de "vida real"e identificar padrões epidemiológicos que seriam inatingíveis com fontes de dados isoladas. O artigo sublinha que o record linkage permite novas investigações clínicas e epidemiológicas, transformando dados fragmentados em um recurso coeso para descobertas médicas.

O Guia de Melhores Práticas para Ligação de Dados em Saúde no NHS [30] focado no contexto do Serviço Nacional de Saúde (NHS) do Reino Unido, oferece uma introdução às melhores práticas e considerações operacionais e metodológicas para a ligação de dados em saúde. Tecnicamente, ele aborda: Princípios de Data Governance, Etapas do Processo de Ligação, Desafios Técnicos e Operacionais reconhecendo os obstáculos comuns, como a qualidade variável dos dados de entrada, a ausência de identificadores únicos consistentes entre bases, a necessidade de expertise em ciência de dados e a infraestrutura computacional para lidar com grandes volumes de dados de saúde. Benefícios e Casos de Uso: Ilustra como a ligação de dados pode ser aplicada para melhorar a qualidade do atendimento, a eficiência do sistema de saúde e a capacidade de pesquisa, conectando dados de atenção primária, secundária, saúde mental e serviços sociais.

Sobre o futuro, [3] aborda a promessa e Desafios da Ligação de Dados de Saúde explorando o potencial transformador e os desafios inerentes à ligação de dados em saúde. Do ponto de vista técnico, ele destaca que a integração de dados de saúde de diversas fontes (clínicas, genômicas, de estilo de vida, ambientais) permite a criação de conjuntos de dados mais ricos e multidimensionais. Isso, por sua vez, pode levar a: Descobertas em Medicina de Precisão, Farmacovigilância Aprimorada, Pesquisa Populacional Mais Abrangente. No entanto, o artigo também discute os desafios técnicos e éticos. A complexidade técnica envolve a harmonização de dados heterogêneos, a necessidade de algoritmos de pareamento robustos que lidem com erros de dados e a

escalabilidade computacional. As questões de privacidade e segurança são enfatizadas, exigindo soluções como o Privacy-Preserving Record Linkage (PPRL) e o uso de ambientes de dados seguros para proteger informações sensíveis do paciente. Ele ressalta a necessidade de infraestrutura adequada e de expertise interdisciplinar para maximizar o valor da ligação de dados de saúde.

Mais recente, abordando a escalabilidade, [23] propõe o uso de frameworks de processamento distribuído, como Apache Spark, e bibliotecas de machine learning para realizar record linkage em grandes volumes de dados de saúde. Discute a eficiência e a capacidade de lidar com a massa de dados, bem como desafios de privacidade e segurança.

Finalmente, [4] foca no papel da vinculação de registros no campo da saúde pública. Ele detalha como a integração de dados permite uma compreensão das tendências de saúde, padrões de doenças e o impacto de intervenções, impulsionando a vigilância, a pesquisa e a formulação de políticas eficazes.

A literatura revisada destaca a importância crescente do pareamento de registros para a saúde pública, facilitando a integração de dados e gerando insights valiosos. Uma preocupação recorrente nesses trabalhos é a qualidade do pareamento, frequentemente avaliada por meio de precisão e recall. Projetos em larga escala, tanto no Brasil quanto no Reino Unido, ilustram o potencial dessa técnica para a pesquisa e gestão em saúde.

No entanto, há uma lacuna nos estudos existentes. Embora a validação da qualidade dos links seja bem explorada, poucos trabalhos se aprofundam na análise dos estimadores e pesos que os algoritmos de pareamento utilizam. A discussão costuma focar nos resultados finais, sem investigar como as diferentes abordagens para o cálculo desses estimadores e pesos impactam a performance e a robustez dos modelos. Da mesma forma, a comparação sistemática de algoritmos sob a perspectiva da estimação de parâmetros e seus efeitos na qualidade dos links ainda é uma área com potencial. Este trabalho busca preencher essa lacuna, investigando e comparando diferentes abordagens para a estimação de pesos e parâmetros em algoritmos de pareamento de registros, com foco em dados de saúde pública no Brasil, oriundos do sistema público e do sistema privado.

Capítulo 3

Metodologia

Este trabalho propõe um novo método de cálculo de m e u e avalia uma estratégia de vinculação de registros entre bases sintéticas representando os sistemas público (Instituição A) e privado (Instituição B) de saúde, com foco na cidade de Belo Horizonte, Minas Gerais.

3.1 cPareia MU Estimator - Cálculo de M e U

Este algoritmo customizado foi inspirado no modelo de Fellegi e Sunter [18] para o cálculo iterativo das probabilidades m (de concordância entre registros correspondentes) e u (de concordância entre registros não correspondentes).

O algoritmo parte de valores iniciais fixos para todos os campos. Sendo eles $m = 0{,}5$ e $u = 0{,}25$. Durante a execução do algoritmo, esses valores são ajustados de forma iterativa com base nos pares comparados e na proporção de correspondências esperada entre as bases, definida como um parâmetro de entrada pelo usuário (por exemplo, 80%). A nota mínima de corte para considerar um par como correspondente é derivada diretamente dos pesos logarítmicos dos campos com base nas estimativas atuais de m e u. Essa nota de corte é dada pela função:

$$threshold = (max\_score - min\_score) * match\_proportion$$

Onde *max_score* é a maior nota que uma comparação pode atingir, ou seja, todos os campos existem e são iguais, *min_score* é a menor nota que uma comparação pode atingir, ou seja, todos os campos existem e são diferentes e *match_proportion* é o parâmetro de proporção definido pelo usuário. O algoritmo 1 ilustra o pseudo código do algoritmo desenvolvido.

input : Pesos MUs, Dados, Campos, Executor do Pareia, *threshold*, Valor de atualização mínimo
output: Novos pesos MUs e *threshold*
Function `cPareiaMuEstimator`(MUs, $Data$, $fields$, $pareiaAlgorithm$, $threshold$, $minDiffToConverge$)**{**

```
maxDiff ← minDiffToConverge;
newMUs;
while maxDiff ≥ minDiffToConverge do
    maxDiff ← 0;
    results ← pareiaAlgorithm(data, MUs);
    matchs ← results["notas" ≥ threshold;
    unmatchs ← results["notas«threshold];
    for fieldinfields do
        fieldTrueMatchs ← matchs[field == MUs[field][mWeigth]];
        fieldFalseMatchs ← unmatchs[field ==
          MUs[field][mWeigth]];
        m ← fieldTrueMatchs/matchs;
        u ← fieldFalseMatchs/unmatchs;
        if abs(m − MUs[field][m]) > maxDiff then
            maxDiff ← abs((m − MUs[field][m]));
        end
        if abs(u − MUs[field][u]) > maxDiff then
            maxDiff ← abs((u − MUs[field][u]));
        end
        mWeigth ← log2(m/u);
        uWeigth ← log2(1 − m/1 − u);
        newMus[field] ← (m, u, mWeigth, uWeigth);
    end
end
MUs ← newMUs
return MUs;
```

Algorithm 1: cPareia MU estimator

O processo é executado até convergência dos parâmetros, sendo monitorada a estabilidade dos valores e o número de pares classificados como positivos em cada iteração. Esta abordagem deriva dos métodos tradicionais como *Expectation e Maximization* e cálculos pré existentes. O desafio reside em encontrar um *threshold* robusto, Uma vez que este que será o principal ponto de apoio para o cálculo de parâmetros além de determinar, ao final do algoritmo, quais registros são suficientemente semelhantes para serem considerados um match.

Os algoritmos SpLink e Record Lingage possuem seus próprios estimadores de M e U já implementados. Para fins de teste e comparação do nosso cálculo que foi utilizado no algoritmo cPareia, este algoritmo foi testado nas bases de dados com os parâmetros M e U calculados pelo cPareia MU Estimatoro e também com os parâmetros M e U estimados pelo Python Record Linkage Toolkit. Os valores de M e U do SpLink não foram utilizados pois não possuem a mesma estrutura proposta.

3.2 Base de Dados Sintética

Diante da indisponibilidade de acesso aos dados reais, este trabalho utilizou uma base de dados sintética que simula um cenário realista de vinculação de dados entre os setores público e privado de saúde. A tabela 3.1 demonstra como foram criadas as bases de dados.

Tabela 3.1. Bases de dados Sintéticas Geradas

Base de Dados	**Total de Registros**	**Registros da Instituição A**	**Registros da Instituição B**	**% de Correspondência**
1	30000	20000	10000	80%
2	50000	30000	20000	80%
3	100000	60000	40000	80%
4	1000000	600000	400000	80%

Os campos disponíveis incluem nome completo, data de nascimento, sexo, município, nome da mãe e CPF (com campos faltantes ou corrompidos em A), permitindo a aplicação de técnicas de pareamento de registros baseadas em similaridade textual, blocagem e classificação. Os campos utilizados foram extraídos das variáveis disponíveis nos Sistema de Informação Ambulatorial (SIA) e do Sistema de Informação Hospitalar (SIH) do SUS no Brasil. [28]

A base sintética foi construída de forma a representar características reais de registros administrativos de saúde. As variáveis presentes nas bases da **Instituição A** (SUS) e da **Instituição B** (setor privado) são descritas na Tabela 3.2.

Tabela 3.2. Descrição das variáveis das bases sintéticas

Variável	Tipo	Descrição
id	String	Identificador único do registro dentro da base de origem.
nome_completo	Texto	Nome completo do paciente.
data_nascimento	Data (YYYY-MM-DD)	Data de nascimento do paciente.
sexo	Categórico (M/F)	Sexo biológico do paciente.
nome_mae	Texto	Nome completo da mãe do paciente.
nome_do_responsável	Texto	Nome completo do responsável do paciente.
cpf	Texto (11 dígitos)	Cadastro de Pessoa Física.
nacionalidade_do_paciente	Texto	Nacionalidade do Paciente.
tipo_logradouro	Texto	Tipo do endereço do paciente, como Rua, Avenida, Bloco.
logradouro	Texto	Logradouro de residência do paciente. Pode aparecer com variações ortográficas.
numero_casa_paciente	Numérico	Número de residência do paciente. Pode aparecer com variações ortográficas.
bairro	Texto	Bairro de residência do paciente. Pode aparecer com variações ortográficas.
código_município_paciente	Texto	O código do município de residência do paciente.
uf_paciente	Texto	O código do estado de residência do paciente.
cep_paciente	Texto	O código postal de residência do paciente.
ground_truth	Inteiro	Identificador único do paciente correspondente da base A.

Dados faltantes e inconsistentes foram inseridos de forma controlada na base A,

simulando erros comuns: nomes abreviados, datas trocadas, campos vazios e variações fonéticas. A variável `ground_truth` foi incluída apenas na base B e não participa do processo de vinculação, sendo usada exclusivamente para avaliação dos pares corretos.

3.3 Estratégia de Pareamento

Em todos os experimentos realizados e para todos os algoritmos testados, na etapa de Blocagem, para reduzir a complexidade computacional, os registros foram agrupados por três regras para limitar a comparação entre pares plausíveis. Foram elas:

- Regra 1: NM_PACIENTE (nome do paciente), SEXO_PAC (sexo do paciente), NM_MAE_PAC (Nome da mãe do paciente)
- Regra 2: CPF (cpf do paciente)
- Regra 3: CEP_PAC (Código postal do paciente), SEXO_PAC (sexo do paciente)

Na etapa de comparação e estimativa de parâmetros houveram pequenas diferenças atendendo as singularidades de cada um dos algoritmos. Por padrão, a comparação pode ser definida como exata ou por similaridade. Os atributos comparados através da comparação exata foram CPF, Sexo, Bairro, Unidade Federativa, Raça/Cor, CEP, Nacionalidade. Os atributos comparados por similaridade foram: nome do paciente, nome da mãe, nome do responsável, tipo de logradouro, logradouro, complemento de endereço. O campo data de nascimento foi tratado de maneira diferente entre os algoritmos de pareamento de registros testados. No cPareia e o Python Record Linkage Toolkit, ela foi comparada por proximidade enquanto o SPLink possui uma função específica de comparação de data.

Neste caso, para a comparação de datas, o algoritmo Splink adota uma abordagem mais robusta, considerando seis casos de correspondência distintos. Cada um desses casos possui parâmetros M e U específicos, permitindo a atribuição de notas diferenciadas de acordo com o grau de similaridade da data. Correspondência exato das datas; Diferença de edição de Levenshtein; Diferença entre as datas de até 1 mês; Diferença entre as datas de até 1 ano; Diferença entre as datas de até 10 anos; Qualquer outro caso, onde então se recebe uma nota mínima baseada nos pesos MUs. Desse modo possíveis variações ou erros nas digitações das datas, que foram inseridas propositalmente na Base de Dados A para simular erros reais nas bases dos SUS, foram melhor avaliadas, dando um desempenho interessante para o Algoritmo Splink. Além disso,

no SPLink, quando um campo é avaliado pela métrica de similaridade Jaro-Winkler, a correspondência exata e a correspondência por similaridade possui peso diferente na comparação.

3.4 Ferramentas Utilizadas

O processamento e análise dos dados foram realizados em Python, utilizando as seguintes bibliotecas: pandas e numpy para manipulação de dados; Jaro Winkler para medidas de similaridade; cPareia, RecordLinkage e Splink como ferramentas especializadas em pareamento de registros.

3.5 Avaliação dos Resultados

A avaliação do desempenho dos métodos de pareamento foi realizada por meio de: Matriz de confusão; Acurácia, Precisão, Revocação (Recall) e F1-score. Á acurácia representa a porcentagem de amostras corretamente classificadas. Por exemplo, a porcentagem de indivíduos corretamente rotulados como doentes ou controles. Precisão representa a porcentagem de amostras previstas como "positivas" que realmente são "positivas". Por exemplo, a porcentagem de variantes identificadas que foram corretamente previstas. Revocação (ou sensibilidade) representa a porcentagem de amostras "positivas" que foram corretamente previstas. Por exemplo, a porcentagem de casos de câncer de mama corretamente identificados. F1-score representa a média harmônica entre precisão e revocação. Por exemplo, minimizando tanto os diagnósticos não detectados (falsos negativos) quanto os diagnósticos incorretos (falsos positivos) em um algoritmo de teste genético. [26]

Capítulo 4

Resultados

Para a avaliação do algoritmo proposto cPareia MU Estimator e da vinculação de dados nas bases sintéticas foram realizados experimentos visando medir o desempenho dos algoritmos de vinculação em termos de acurácia, eficiência computacional e escalabilidade. As 4 bases de dados descritas na seção **??** foram utilizadas nestas análises. Todas as bases seguiram o mesmo padrão estrutural, contendo campos de identificação pessoal como nome, data de nascimento, nome da mãe, sexo, município e CPF, com variações no grau de ruído e completude.

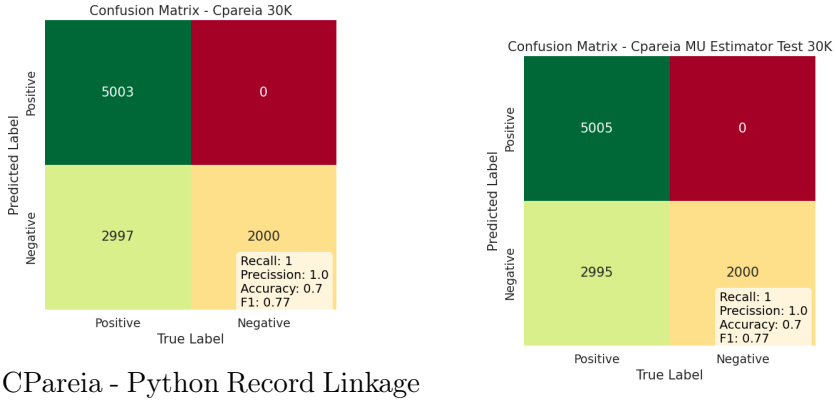
Foram testadas quatro configurações distintas de algoritmos, descritas a seguir:

- **cPareia - Record Linkage M e U**: Execução do sistema *cPareia* utilizando os parâmetros probabilísticos m e u estimados automaticamente pelo algoritmo clássico de Record Linkage, com base no modelo de Fellegi-Sunter.
- **cPareia - MU Estimator Test**: Execução do *cPareia* com os valores de m e u estimados pelo algoritmo proposto neste trabalho, que utiliza uma abordagem baseada em limiar de correspondência esperada e iteração até convergência.
- **Python Record Linkage**: Execução da biblioteca `recordlinkage` em Python, utilizando seus mecanismos internos para a estimativa dos parâmetros e pareamento probabilístico.
- **Splink**: Execução da ferramenta `Splink`, com blocagem e estimação dos parâmetros m e u via Expectation-Maximization (EM), e aplicação do modelo de decisão probabilístico nativo da biblioteca.

Cada configuração foi aplicada com foco na análise dos seguintes aspectos: Acurácia, precisão, revocação (sensibilidade) e F1-score, calculados a partir da comparação

com a variável `ground_truth` presente na base sintética; Número de pares candidatos gerados após a etapa de blocagem e estabilidade dos parâmetros estimados em cada técnica.

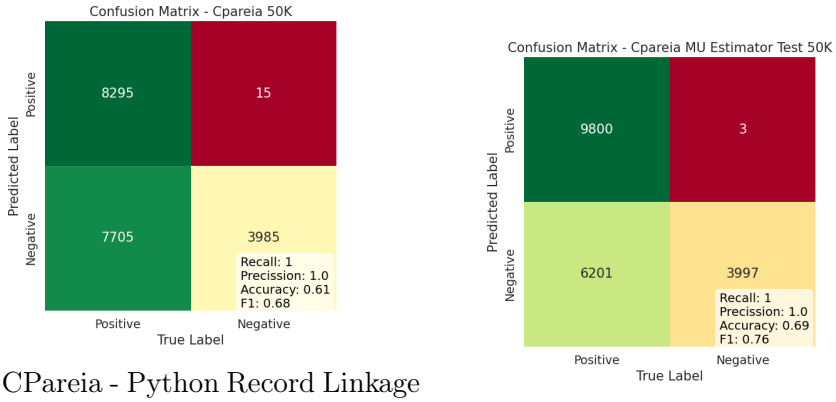
Para o algoritmo cPareia MU Estimator, foram testados dois diferentes *thresholds* de 0,5 e 0,7 como mínimo de igualdade de atributos para considerar um par verdadeiro, onde os pesos calculados foram similares. Entretanto, com o *threshold* mais baixo o algoritmo demorou mais para convergir. Assim, foi utilizado nos experimentos o threshold mais alto (70%). A figura **??** exibe a matriz de confusão para cada uma das bases testadas na comparação.



(a) CPareia - Python Record Linkage MU

(b) CPareia - MU Estimator Test

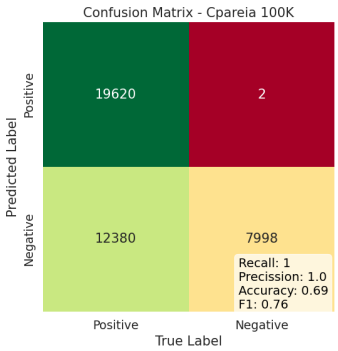
Figura 4.1. Matriz de Confusão dos Algoritmos Na base de Dados com 30.00 pacientes



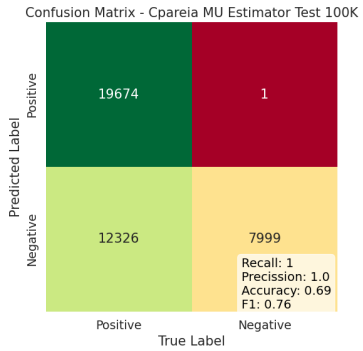
(a) CPareia - Python Record Linkage MU

(b) CPareia - MU Estimator Test

Figura 4.2. Matriz de Confusão dos Algoritmos Na base de Dados com 50.00 pacientes

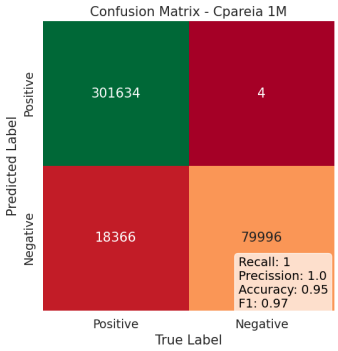


(a) CPareia - Python Record Linkage MU

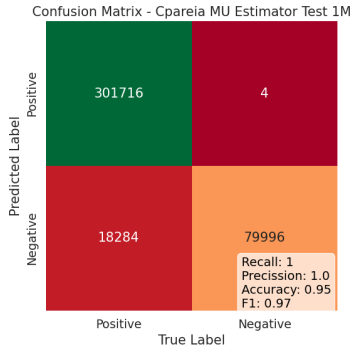


(b) CPareia - MU Estimator Test

Figura 4.3. Matriz de Confusão dos Algoritmos Na base de Dados com 100.00 pacientes



(a) CPareia - Python Record Linkage MU



(b) CPareia - MU Estimator Test

Figura 4.4. Matriz de Confusão dos Algoritmos Na base de Dados com 1.000.000 pacientes

Os resultados obtidos pelo estimador proposto mostraram-se bastante similares aos dos métodos de referência, o que, por si só, já indica a consistência do modelo, uma vez que ele se aproxima do desempenho de estimadores amplamente utilizados. No entanto, destaca-se que, no cenário específico da base com 50.000 registros, o algoritmo com estimativa própria de m e u superou os demais em todas as métricas avaliadas.

As tabelas 4.1 até 4.4 resumem a performance dos diferentes algoritmos testados neste trabalho. É importante lembrar que a quantidade de registros em cada base foi aumentada a fim de se testar velocidade de execução, entretanto, para as 4 bases o tempo de execução foi insignificante em todos os métodos.

Tabela 4.1. Base de Dados - 30 Mil de pacientes

Algoritmo	F1	Recall	Precission	Accuracy
CPareia - Python Record Linkage MU	0.77	0.63	1.00	0.70
CPareia - MU Estimator Test	0.77	0.63	1.00	0.70
Python Record Linkage	0.77	0.62	1.00	0.70
SpLink	0.98	0.96	1.00	0.97

Na tabela 4.1, é possível observar que que os três primeiros algoritmos avaliados — CPareia com parâmetros estimados por Python Record Linkage, CPareia com o estimador proposto e o Python Record Linkage — apresentaram desempenho idêntico em todas as métricas: F1-score de 0,77, revocação de 0,63, precisão perfeita de 1,00 e acurácia de 0,70. Isso indica que, embora todos os métodos tenham sido conservadores na identificação de pares positivos (alta precisão), a sensibilidade foi relativamente baixa, o que compromete a cobertura dos pares verdadeiros. Por outro lado, o Splink obteve desempenho significativamente superior, com F1-score de 0,98, revocação de 0,96, precisão de 1,00 e acurácia de 0,97, mostrando-se mais eficaz na recuperação de verdadeiros correspondentes sem perda de precisão.

Tabela 4.2. Base de Dados - 50 Mil de pacientes

Algoritmo	F1	Recall	Precission	Accuracy
CPareia - Python Record Linkage MU	0.68	0.52	1.00	0.61
CPareia - MU Estimator Test	0.76	0.61	1.00	0.69
Python Record Linkage	0.76	0.61	1.00	0.69
SpLink	0.98	0.96	1.00	0.97

A tabela 4.2, os resultados demonstram uma maior distinção entre os algoritmos. O CPareia com estimador proposto e o Python Record Linkage apresentaram desempenho idêntico, com F1-score de 0,76, revocação de 0,61, precisão de 1,00 e acurácia de 0,69. Já o CPareia com parâmetros estimados por Record Linkage obteve um desempenho inferior, com F1-score de 0,68 e revocação de apenas 0,52, embora mantendo precisão máxima. O destaque mais uma vez foi o Splink, que obteve F1-score de 0,98, revocação de 0,96 e acurácia de 0,97. Esses resultados reforçam a robustez do estimador proposto, que superou a versão baseada em parâmetros externos, apresentando uma sensibilidade maior sem perda de precisão.

Tabela 4.3. Base de Dados - 100 Mil de pacientes

Algoritmo	F1	Recall	Precission	Accuracy
CPareia - Python Record Linkage MU	0.76	0.61	1.00	0.69
CPareia - MU Estimator Test	0.76	0.61	1.00	0.69
Python Record Linkage	0.76	0.61	1.00	0.69
SpLink	0.98	0.96	1.00	0.97

A tabela 4.3, com o aumento para 100 mil registros, os três primeiros métodos — CPareia (ambas as versões) e Python Record Linkage — apresentaram desempenho idêntico em todas as métricas, com F1-score de 0,76, revocação de 0,61, precisão de 1,00 e acurácia de 0,69. Isso sugere estabilidade entre os métodos, mas ainda com limitações na sensibilidade. O Splink, por sua vez, manteve o melhor desempenho, com F1-score de 0,98, revocação de 0,96 e acurácia de 0,97, demonstrando que sua abordagem com estimativa via Expectation-Maximization e blocagem avançada continua eficaz mesmo com o crescimento da base de dados.

Tabela 4.4. Base de Dados - 1 Milhão de pacientes

Algoritmo	F1	Recall	Precission	Accuracy
CPareia - Python Record Linkage MU	0.97	0.94	1.00	0.95
CPareia MU Estimator Test	0.97	0.94	1.00	0.95
Python Record Linkage	0.98	0.96	1.00	0.96
SpLink	0.98	0.95	1.00	0.96

A tabela 4.4 mostra que os resultados se aproximaram significativamente entre os métodos. O CPareia, tanto com parâmetros do Python Record Linkage quanto com o estimador proposto, obteve F1-score de 0,97, revocação de 0,94, precisão de 1,00 e acurácia de 0,95. O Python Record Linkage e o Splink apresentaram desempenho levemente superior, com F1-score de 0,98 e acurácia de 0,96. Esses resultados indicam que, em escalas maiores, todos os métodos convergem para uma boa performance, com destaque para o comportamento consistente do estimador próprio, que se manteve competitivo frente a ferramentas amplamente utilizadas na literatura.

De maneira geral, o Splink apresentou o melhor desempenho entre os métodos avaliados, com os maiores indicadores em todas as bases de dados analisadas. Esse resultado pode ser atribuído, em grande parte, à abrangência de sua abordagem de comparação, que permite uma avaliação mais detalhada entre os pares de registros. Em relação aos demais algoritmos, observa-se que o aumento no volume de dados contribuiu positivamente para a melhoria do desempenho, especialmente na base com 1.000.000

de registros, na qual todos os métodos apresentaram ganhos significativos em F1-score, acurácia e revocação. Tal comportamento reforça a hipótese de que a disponibilidade de um maior número de exemplos durante a etapa de estimativa dos pesos probabilísticos influencia diretamente na qualidade do pareamento. Cabe destacar que, para garantir a comparabilidade entre os métodos, a estimação dos parâmetros m e u foi realizada utilizando a totalidade de cada base de dados em todos os experimentos conduzidos. A precisão de 1 para todos os restes reforça que foi utilizado uma nota alta para aceitar um par como verdadeiro, o que garante segurança e aumenta a especificidade entretanto diminui a sensibilidade.

Tabela 4.5. Comparações Geradas

Algoritmo	30K	50K	100K	1M	Total
CPareia - Python Record Linkage MU	13233	43166	53747	642912	753058
CPareia - MU Estimator Test	13233	43166	53747	642912	753058
Python Record Linkage	7746	15609	31754	420692	475801
SpLink	13234	26673	53746	639867	733520

A Tabela 4.5 apresenta o número de comparações geradas por cada algoritmo nas diferentes bases. Como a única diferença entre o *CPareia - Python Record Linkage MU* e o *CPareia - MU Estimator Test* está na forma de estimar os parâmetros probabilísticos m e u, é natural que ambos produzam exatamente o mesmo número de comparações, já que compartilham a mesma estratégia de blocagem. Ao compará-los com o *Python Record Linkage*, observa-se que este último realiza significativamente menos comparações, o que se reflete em um menor custo computacional. No entanto, essa economia vem acompanhada de uma taxa de *recall* consideravelmente mais baixa, indicando que muitos pares verdadeiros deixaram de ser identificados.

No caso do CPareia, apesar do maior número de comparações, o desempenho de *recall* também foi modesto, o que pode estar relacionado ao uso de um limiar (*threshold*) excessivamente restritivo. Essa hipótese foi testada ao considerar apenas os pares com notas positivas, mas não se verificou uma melhora expressiva. Por outro lado, o *Splink* mais uma vez se destacou ao apresentar a maior acurácia e uma taxa de *recall* significativamente superior, mesmo gerando um número de comparações apenas ligeiramente inferior ao do CPareia. Isso evidencia a eficácia de suas estratégias internas de blocagem e estimativa de parâmetros na identificação de correspondências verdadeiras.

Capítulo 5

Conclusão

Os resultados deste trabalho indicam que o método proposto para a estimação dos parâmetros probabilísticos m e u, especialmente quando integrado ao algoritmo *cPareia*, apresentou desempenho comparável ao de estimadores amplamente utilizados na literatura. Em cenários com limiares de decisão bem calibrados e volumes de dados suficientemente grandes para permitir uma boa inferência estatística, o estimador demonstrou potencial para alcançar resultados ainda mais robustos. Esses achados sugerem que o algoritmo desenvolvido é, no mínimo, tão eficaz quanto os métodos tradicionais, oferecendo uma alternativa viável e eficiente para contextos em que não se dispõe de pares rotulados para treinamento supervisionado.

No que diz respeito ao desempenho geral dos algoritmos de pareamento, o *Splink* destacou-se de forma consistente como a ferramenta mais precisa e sensível. Tal superioridade pode ser atribuída à sua modelagem probabilística refinada, que permite uma estimativa mais granular dos pesos e uma melhor mensuração das semelhanças entre os registros. Essa abordagem mais detalhada resulta em maior capacidade de discriminação entre pares verdadeiros e falsos, especialmente em bases de dados com alta variabilidade e ruído, reforçando sua posição como uma das soluções mais eficazes disponíveis para vinculação probabilística de registros.

5.1 Limitações e Trabalhos Futuros

Embora os resultados obtidos neste estudo tenham sido promissores, algumas limitações devem ser reconhecidas. A principal delas diz respeito ao uso de bases de dados sintéticas. Apesar de cuidadosamente construídas para simular cenários realistas, essas bases não reproduzem integralmente a complexidade, os padrões de erro e a heterogeneidade presentes em dados reais oriundos de sistemas públicos e privados de

saúde. Consequentemente, a generalização dos resultados para ambientes de produção requer validação adicional.

Outra limitação está relacionada à utilização de um único limiar de decisão (*threshold*) fixo para todos os experimentos. Ainda que tal abordagem tenha permitido a comparação direta entre métodos, diferentes bases ou domínios podem demandar estratégias mais flexíveis e adaptativas para definição de pontos de corte. Além disso, a avaliação dos algoritmos focou majoritariamente em métricas de desempenho clássicas (F1, revocação, precisão e acurácia), sem considerar diretamente aspectos como escalabilidade em tempo real, uso de memória ou interpretabilidade dos modelos.

Como trabalhos futuros, propõe-se a aplicação do estimador desenvolvido a bases reais oriundas de sistemas como o SIH/SUS e dados de operadoras privadas, de modo a avaliar sua robustez em contextos operacionais Por fim, é recomendada a incorporação de técnicas de blocagem híbrida e seleção automática de atributos, de forma a otimizar ainda mais o equilíbrio entre qualidade do pareamento e eficiência computacional.

Referências Bibliográficas

[sci] We must pay more attention to record linkage quality. *SciELO Public Health*. 11

[2] (2022). Data linkage in medical research. *BMJ Medicine*. 12

[3] (2023). Better together: the promise of health data linkage and its challenges. *Lifebit*. 12

[4] (2025). Advancing public health through data linkage. *Number Analytics*. 13

[5] Asher, J.; Resnick, D.; Brite, J.; Brackbill, R. & Cone, J. (2020). An introduction to probabilistic record linkage with a focus on linkage processing for wtc registries. *International journal of environmental research and public health*, 17(18):6937. 6

[6] Benhar, H.; Idri, A. & Fernández-Alemán, J. (2020). Data preprocessing for heart disease classification: A systematic literature review. *Computer Methods and Programs in Biomedicine*, 195:105635. 5

[7] Campos, F. E. & Haddad, A. E. (2019). Interoperabilidade entre sistemas de informação em saúde no brasil: avanços e desafios. *Cadernos de Saúde Pública*, 35(12). 1

[8] Chipperfield, J.; Hansen, N. & Rossiter, P. (2018). Estimating precision and recall for deterministic and probabilistic record linkage. *International Statistical Review*, 86(2):219--236. 11

[9] Coelho Neto, G. C. (2019). Integração entre sistemas de informação em saúde: o caso do e-sus atenção básica. Dissertação de mestrado, Escola Paulista de Medicina, Universidade Federal de São Paulo. 10

[10] Costa, M. V. da S., C. M. C. S. V. S. M. N. . M. U. V. d. S. (2025). Avanços e desafios da interoperabilidade no sistema Único de saúde. *Journal of Health Informatics*. 10

[11] da Saúde. Secretaria Executiva, B. M. (2001). *SUS–princípios e conquistas*. Ministério da Saúde. 9

[12] Dasylva, A.; Goussanou, A.; Ajavon, D. & Abousaleh, H. (2019). Revisiting the probabilistic method of record linkage. 11

[13] de Aragão, S. M. & Schiocchet, T. (2020). Lei geral de proteção de dados: desafio do sistema único de saúde. *Revista Eletrônica de Comunicação, Informação & Inovação em Saúde*, 14(3). 9

[14] de Bruin, J. (2019). Python record linkage toolkit: A toolkit for record linkage and duplicate detection in python. 7

[15] Domingos, C. M.; Ferraz, E. d. M. & Carvalho, B. G. (2019). Governança das ações e serviços de saúde de média complexidade em uma região de saúde. *Saúde em Debate*, 43:700--711. 9

[16] dos Santos Filho, W. (2009). Algoritmo paralelo e eficiente para o problema de pareamento de dados. Dissertação de mestrado em ciência da computação, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil. Acessado em 25 de junho de 2025. 8

[17] Duffin, J. A.; Ma, Y.; Zhai, F. & Wang, L. (2020). Comparing methods for record linkage for public health action: Matching algorithm validation study. *PubMed Central*, 8(10):e006627. 11

[18] Fellegi, I. P. & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183--1210. 2, 5, 10, 11, 14

[19] for Healthcare Research, A. & (AHRQ), Q. (2014). Linking data for health services research: A framework and instructional guide. Relatório técnico. 10

[20] Fornazin, M. (2015). *A informatização da saúde no Brasil: uma análise Multi-Paper inspirada na Teoria Ator-Rede*. Tese de doutorado. 10

[21] Junior, A. A. G.; Pereira, R. G.; Gurgel, E. I.; Cherchiglia, M.; Dias, L. V.; Ávila, J. D.; Santos, N.; Reis, A.; Acurcio, F. A. & Junior, W. M. (2018). Building the national database of health centred on the individual: administrative and epidemiological record linkage-brazil, 2000-2015. *International Journal of Population Data Science*, 3(1):446. 11

[22] Linacre, R. (2020). Splink: Probabilistic data linkage at scale. https://github.com/moj-analytical-services/splink. Accessed: 2025-06-25. 7

[23] Liu, X.; Wang, T.; Ma, Y.; Deng, L.; Cao, Y.; Gao, X.; Tang, J.; Yang, Z.; Wang, W. & Wang, X. (2024). Distributed record linkage in healthcare data with apache spark. *arXiv preprint arXiv:2404.07939*. 13

[24] Lowe, D.; Berecki-Gisella, J.; Boyle, F.; Forbes, A.; Haining, C.; Kelly, J.; Morris, E.; Palmer, A.; Petrie, D.; Roberts, A.; Schmidt, M.; Watson, S.; Welsh, J. & Wilkinson, M. (2017). A guide to evaluating linkage quality for the analysis of linked data. *International Journal of Population Data Science*, 2:25. 10

[25] Miller, F. P.; Vandome, A. F. & McBrewster, J. (2009). Levenshtein distance: Information theory, computer science, string (computer science), string metric, damerau? levenshtein distance, spell checker, hamming distance. 6

[26] Naidu, G.; Zuva, T. & Sibanda, E. M. (2023). A review of evaluation metrics in machine learning algorithms. In *Computer science on-line conference*, pp. 15--25. Springer. 19

[27] Paixao, E. S.; Blencowe, H.; Falcao, I. R.; Ohuma, E. O.; dos Santos Rocha, A.; Alves, F. J. O.; Costa, M. d. C. N.; Suárez-Idueta, L.; Ortelan, N.; Smeeth, L. et al. (2021). Risk of mortality for small newborns in brazil, 2011-2018: a national birth cohort study of 17.6 million records from routine register-based linked data. *The Lancet Regional Health–Americas*, 3. 11

[28] Sá, D. A. d. (2002). *Atenção à saúde no Brasil: um estudo do acesso à assistência a partir do Sistema de Informações Ambulatoriais do SUS (SIA/SUS)*. Tese de doutorado, Dissertação (Mestrado em Saúde Pública). Centro de Pesquisas Aggeu Magalhães 16

[29] Sadinle, M. (2016). Bayesian estimation of bipartite matchings for record linkage. 10

[30] Unit, N. H. E. (2022). A guide to data linkage. Relatório técnico. 12

[31] Wang, Y.; Qin, J. & Wang, W. (2017). Efficient approximate entity matching using jaro-winkler distance. In *International conference on web information systems engineering*, pp. 231--239. Springer. 6

[32] Zipkin, E. F.; Zylstra, E. R.; Wright, A. D.; Saunders, S. P.; Finley, A. O.; Dietze, M. C.; Itter, M. S. & Tingley, M. W. (2021). Addressing data integration challenges

to link ecological processes across scales. *Frontiers in Ecology and the Environment*, 19(1):30--38. 11