META 2: FORTALECIMENTO DO OBSERVATÓRIO NACIONAL DO MERCADO DE TRABALHO

META 2 – Produto 2: Relatório metodológico e aplicação de anonimização e integração de bases de dados selecionadas

TERMO DE FOMENTO Nº 01/2021 - PLATAFORMA+BRASIL Nº 919592

Dezembro de 2023

DIEESE
DEPARTAMENTO INTERSINDICAL DE ESTATÍSTICA E ESTUDOS SOCIOECONÔMICOS

EXPEDIENTE DO MINISTÉRIO DO TRABALHO E EMPREGO

Presidente da República
Luiz Inácio Lula da Silva

Ministro do Trabalho e Emprego
Luiz Marinho

Secretário Executivo
Francisco Macena da Silva

Subsecretária de Estatística e Estudos do Trabalho
Paula Montagner

Subsecretaria de Estatística e Estudos do Trabalho
Felipe Vella Pateo

Esplanada dos Ministérios - Bloco F – Ed. Sede
Brasília – DF
70059-900
(61) 2021- 5449

Obs.: Os textos não refletem necessariamente a posição do Ministério do Trabalho e Emprego

DIEESE DEPARTAMENTO INTERSINDICAL DE ESTATÍSTICA E ESTUDOS SOCIOECONÔMICOS

Escritório Nacional: Rua Aurora, 957 – 1º andar
CEP 05001-900 São Paulo, SP
Telefone (11) 3874-5366 / fax (11) 3874-5394
E-mail: en@dieese.org.br
www.dieese.org.br

Ficha Técnica

Coordenação do Projeto
Patrícia Toledo Pelatieri – Coordenadora geral

Equipe Executora
DIEESE

Apoio
Equipe administrativa do DIEESE

Entidade Executora
Departamento Intersindical de Estatística e Estudos Socioeconômicos - DIEESE

SUMÁRIO

Sumário

1. Apresentação

Com o advento da Lei nº 13.709, de 14 de agosto de 2018, Lei Geral de Proteção de Dados Pessoais (LGPD), torna-se necessário a anonimização de bases de dados que contenham informações pessoais e/ou informações sensíveis, tais como Relação Anual de Informações Sociais (RAIS), Cadastro Geral de Empregados e Desempregados (CAGED), Seguro-Desemprego (SD), Abono Salarial e benefício Emergencial de Manutenção do emprego e da Renda (BEm), dentre outras.

Isto posto, o presente produto faz parte do plano de trabalho do Termo de Fomento nº 001/2021 PLATAFORMA+BRASIL Nº 919592, celebrado com o então Ministério do Trabalho e Previdência (MTP) do governo federal, está inserido na Meta 2 – "Fortalecimento do Observatório Nacional do Mercado de Trabalho", e tem como objetivo verificar possibilidades de anonimizar determinadas bases de dados, garantindo que as informações pessoais contidas não sejam identificadas e identificáveis. Importante que o processo de anonimização não seja revertido utilizando exclusivamente meios próprios, ou quando, com esforços razoáveis, puder ser revertido.

Propõe-se que os dados de identificação dos trabalhadores sejam substituídos por códigos fictícios, capazes de manter um elo entre as bases de dados, e que, além disso, sejam utilizados outros meios necessários para a não identificação do titular da informação, como a técnica de generalização, em que os dados precisos são substituídos por categorias mais amplas e genéricas, por exemplo, idades exatas são convertidas em faixas etárias.

O desenvolvimento desta etapa passa pela realização de oficinas (online) metodológicas com especialistas e instituições produtoras e usuárias das bases, de modo que a proposta contemple os diversos usos.

2. Atividades realizadas

No período de abril de 2022 a dezembro de 2023 foram realizadas oficinas técnicas e reuniões com diversos especialistas com o objetivo de debater as possibilidades, dada a complexidade do desafio proposto – desenvolvimento de uma metodologia de anonimização de grandes bases de forma interligada.

O DIEESE produziu subsídios e organizou todas essas atividades, conforme relatado abaixo.

2.1. Oficina Técnica Interna – 13/05/2022

Pauta: Desenvolvimento do plano de trabalho: definição de variáveis a serem consideradas

Horário: 8h30 -12h30

Participantes: Equipe do Dieese

Pauta: Levantamento de profissionais com experiência no tema. Ficou estabelecido que seria feito contato individualmente com cada especialista sugerido, a fim de explicar em detalhes o objetivo do trabalho e discutir a possibilidade da metodologia de seu conhecimento ser a mais adequada.

2.2. Reuniões com especialista – 17/05 e 20/05/2022

Pauta: Metodologia de anonimização de grandes bases

Horário: 14h às 17h

Participantes: Equipe Dieese e Agusto Fadel – graduado em Estatística pela Escola Nacional de Ciências Estatísticas (ENCE/IBGE), com mestrado em Computação pela Universidade Federal Fluminense (UFF), na área de Algoritmos e Otimização, é atualmente doutorando na mesma área, onde desenvolve pesquisa sobre confidencialidade de dados.

2.3. Reunião com especialista – 23/05/2022

Pauta: Metodologia de anonimização de grandes bases

Horário: 14h às 17h

Participantes: Equipe Dieese e Sábado Girardi, médico, coordenador do Núcleo de Educação em Saúde Coletiva – NESCON - UFMG, e trabalha com a "reidentificação" do CNES do Datasus.

2.4. Reunião com especialista – 30/05/2022

Pauta: Metodologia de anonimização de grandes bases

Horário: 14h às 17h

Participantes: Equipe Dieese e Karollaynne Silva - Mestre em População, Território e Estatísticas Públicas na Escola Nacional de Ciências Estatísticas (ENCE/IBGE) e graduada em Ciências

Econômicas pela Universidade Federal de São João De Rei (UFSJ). Possui experiência em Métodos Quantitativos, com ênfase em Métodos Econométricos e Estatística Espacial.

2.5. Oficina com Especialistas no tema – 20/06/2022

Desafios da anonimização de grandes bases de dados

Abertura

14h00 - 14h15 - Patrícia Pelatieri - Diretora Técnica Adjunta do DIEESE

Demandas para disponibilização das bases de dados do Ministério do Trabalho e Previdência

14h15 - 14h45 - Felipe Vella Pateo (Coordenador-Geral de Cadastros, Identificação Profissional e Estudos do Ministério do Trabalho e Previdência)

Roda de Conversa: Desafios e Caminhos possíveis para anonimização de grandes bases

14h45 - 15h15 - Augusto Fadel (IBGE) - Uma visão geral de técnicas de preservação de privacidade aplicadas ao compartilhamento seguro de microdados

15h15 - 15h45 – Erivelton Pires Guedes – (IPEA) A experiência de anonimização da RAIS

16h00 - 16h30 – Thais Paiva (UFMG) - A experiência da UFMG com anominização de bases de dados

Debate

16h30 - 17h30

Encerramento

17h30 - 18h

Relatório da Oficina no Anexo I

2.6. Reunião com especialista – 16/09/2022

Pauta: Apresentação de proposta detalhada de trabalho (no Anexo II)

Horário: 10h às 13h

Participantes: Equipe Dieese e Karollaynne Silva

2.7. Reunião com especialista – 07/11/2022

Pauta: Metodologia de anonimização de grandes bases

Horário: 14h às 17h

Participantes: Equipe Dieese e Thais Paiva - Professora do Departamento de Estatística da Universidade Federal de Minas Gerais (UFMG) desde 2016. Possui graduação em Ciências Atuariais pela UFMG, mestrado em Estatística também pela UFMG, e doutorado em Estatística na Duke University, EUA.

Alguns dos seus projetos de pesquisa incluem métodos de imputação para dados sintéticos, simulação de coordenadas geográficas sintéticas para bases de dados confidenciais, e imputação de variáveis aleatórias contínuas multivariadas para dados ausentes não aleatórios.

2.8. Reunião com especialista – 07/11/2022

Pauta: Metodologia de anonimização de grandes bases

Horário: 14h às 17h

Participantes: Equipe Dieese e empresa Hacklab, empresa de tecnologia, com equipe multidisciplinar e experiência em projetos de anonimização

2.9. Reunião com Ministério do Trabalho e Emprego – 16/12/2022

Pauta: Anonimização e Integração das Bases

Horário: 15h às 17h

Participantes: Equipe Dieese e Equipe do Ministério do Trabalho (Raigner Rezende do Nascimento; Augusto Veras Soares M Albuquerque, Welton Resende de Oliveira, Robert Paula Gouveia, Felipe Vella Pateo, Viviani Renata Anze Greer , Eloa Nascimento dos Santos e Amilton Lobo Mendes Junior

No anexo III - planilha de controle do Ministério, com solicitações recebidas para acesso a dados identificados e descrição resumida da justificativa apresentada.

2.10. Reunião com Assessoria Jurídica – 15/02/2023

Pauta: Anonimização e a LGPD - Lei Geral de Proteção de Dados Pessoais

Apresentação do trabalho de análise, andamento, objetivos e questões que demandam suporte e acompanhamento jurídico. Escuta das elaborações, normas e demais aspectos jurídicos já estruturados por essas áreas relacionadas à natureza deste projeto.

Horário: 14h às 16h

Participantes: Equipe Dieese e Assessoria Jurídica

2.11. Reuniões com Ministério do Trabalho e Emprego – 06/10/2023; 11/10/2023 e 04/11/2023

Pauta: Dúvidas sobre as bases – planilha no Anexo IV

Horário: 15h às 17h

Participantes: Equipe Dieese, equipe Hacklab e Equipe do Ministério do Trabalho

3. Relatório Final: Prova de Conceito para anonimização de bases de dados do Ministério do Trabalho e Emprego

1. Introdução

1.1 Contexto

Diante do interesse público pelas bases de dados do Ministério do Trabalho e Emprego por diversos setores da sociedade, o Ministério e o Dieese em parceria com a Hacklab, ASK - Associated Researchers e Lago & Lago, desenvolveram uma metodologia de anonimização para cinco bases de dados:

- Benefício Emergencial de Preservação do Emprego e da Renda (BEM);
- Cadastro Geral de Empregados e Desempregados (CAGED);
- Relação Anual de Informações Sociais (RAIS) – estabelecimentos;
- Relação Anual de Informações Sociais (RAIS) – vínculos;
- Seguro Desemprego (SD).

Essa anonimização considerou tanto a Lei de Transparência e Acesso à Informação, como a Lei Geral de Proteção de Dados, como contornos jurídicos para as tomadas de decisões quanto à metodologia, diagnóstico, técnicas de anonimização e desenvolvimento da prova de conceito em si.

1.2 Justificativa

A Lei da Transparência (LC 131/2009) orienta a disponibilização de informações focadas no uso dos recursos públicos, com abertura e publicização das receitas e despesas da administração pública. Sua aprovação e regulamentação foi estruturante para a consequente regulamentação do Acesso à Informação (LEI 12.527), promulgada em 2011, a qual ampliou e complexificou a questão da transparência, garantindo acesso à informação - conforme previsto na Constituição Federal - a qualquer pessoa e organização que a solicite.

Com isso, desde 2009, a administração pública opera com a demanda de acesso e publicização de suas informações e, ao mesmo tempo, com a necessidade de estruturação de suas informações para que se tornem públicas, dada que a coleta, organização e acesso a tais informações identificadas são, com exceções controladas, até então, de uso interno - o que

refletiu em metodologias de coleta, gestão e organização destas bases de dados comprometidas apenas com as finalidades da própria administração.

Diante das questões éticas e políticas que envolvem o acesso a essas bases de dados produzidas pela administração pública e também pelo setor privado, em 2018 foi promulgada a Lei Geral de Proteção de Dados Pessoais (LGPD[1]) que apresentou novos parâmetros para a divulgação, comercialização e uso interno de dados - demandando um novo e mais complexo trabalho desde a coleta até a publicização deles.

Em seu artigo 5º, a LGPD define que dado pessoal é toda informação relacionada a pessoa natural "identificada" ou "identificável" e determina alguns princípios norteadores que o tratamento desses dados deve considerar[2]. Ou seja, ao seguir os marcos legais estabelecidos, as organizações públicas ou privadas indicam que os dados pessoais coletados, mediante consentimento do titular, são necessários, mínimos, corretos, de qualidade e que atendem a uma finalidade válida.

1.3 Objetivo geral

O principal objetivo deste projeto foi desenvolver uma metodologia de anonimização, como uma prova de conceito, para a promoção do acesso mais amplo às bases de dados, de acordo com os parâmetros das Leis de Transparência, Acesso à Informação e LGPD.

Ou seja, ao mesmo tempo deseja-se garantir o acesso à informação pública relevante e assegurar o respeito à privacidade de cidadãos e cidadãs, considerando que a privacidade é elemento indutor da autonomia e da cidadania, bem como pressuposto de uma sociedade democrática moderna (DONEDA, 2019, p. 128-129).

1.4 Objetivos Específicos

Vale ressaltar que a anonimização absoluta de dados sensíveis e pessoais é uma questão que a Ciência da Computação não responde com infalibilidade. A reidentificação é uma possibilidade, sobretudo quando se trata de bases complexas, com múltiplas chaves de identificação disponíveis não só na mesma base de dados como também em outras bases de dados que não são de domínio do Ministério do Trabalho e Emprego e não possuem necessariamente os mesmos parâmetros comuns de anonimização.

São, portanto, objetivos específicos:

1. Garantir uma documentação da metodologia, que revele seus parâmetros de análise das variáveis;
2. Garantir uma documentação que aponte as brechas de reidentificação dasvariáveis publicizadas;
3. Garantir uma documentação que aponte as técnicas de anonimização recomendadas com justificativas;
4. Garantir insumos para uma análise jurídica futura.

2. Planejamento e gestão

2.1 Acordos iniciais

Com a definição do escopo de trabalho, o primeiro passo para a execução do presente trabalho envolveu a formação de uma equipe especializada em anonimização e publicização de bases de dados públicas. Foram articuladas e formalizadas parcerias entre DIEESE, hacklab/, ASK - Associated Researchers e Lago& Lago.

Após a formação da equipe, foram realizadas reuniões entre DIEESE e Ministériodo Trabalho e Emprego para melhor compreensão do significado das variáveis de cada base.

2.2 Acesso à amostra das bases

Anteriormente à disponibilização das bases de dados, foram definidas as diretrizes para o acesso com a formalização de documentos como termos de confidencialidade entre as partes e a metodologia de extração de amostragem das bases a serem anonimizadas.

[1] Lei nº 13.709/2018 - Lei Geral de Proteção de Dados Pessoais (LGPD) está disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm

[2] Art. 6º da LGPD define estes 10 princípios de tratamento de dados pessoais: (i) finalidade, (ii) adequação, (iii) necessidade, (iv) livre acesso, (v) qualidade dos dados, (vi) transparência, (vii) segurança, (viii) prevenção, (ix) não discriminação, (x) responsabilização e prestação de contas.

[3] DONEDA, Danilo. Da privacidade à proteção de dados pessoais: fundamentos da Lei Geral de Proteção de Dados. 2. ed. São Paulo: Revista dos Tribunais, 2019.

A equipe do DIEESE promoveu uma extração das bases de 1 município, definido aleatoriamente, por região do Brasil, de cada uma das bases, garantindo a diversidade avaliada como necessária pelo órgão.

2.3 Metodologia de execução do trabalho

O método Kanban (de produção) foi escolhido por ser um método ágil, de abordagem simples, eficiente e direta para a administração das demandas de desenvolvimento de soluções tecnológicas. As etapas clássicas (cada uma com sua coluna no quadro kanban) são:

- to do: o que deve ser feito;
- doing: tarefas em execução;
- done: tarefas concluídas.

Além do acompanhamento e gestão das sprints por seu quadro Kanban, o Gitlab também foi adotado como ferramenta de versionamento de códigos / documentos, tribuição de tarefas da equipe e de controle dos tempos despendidos em cada tarefa.

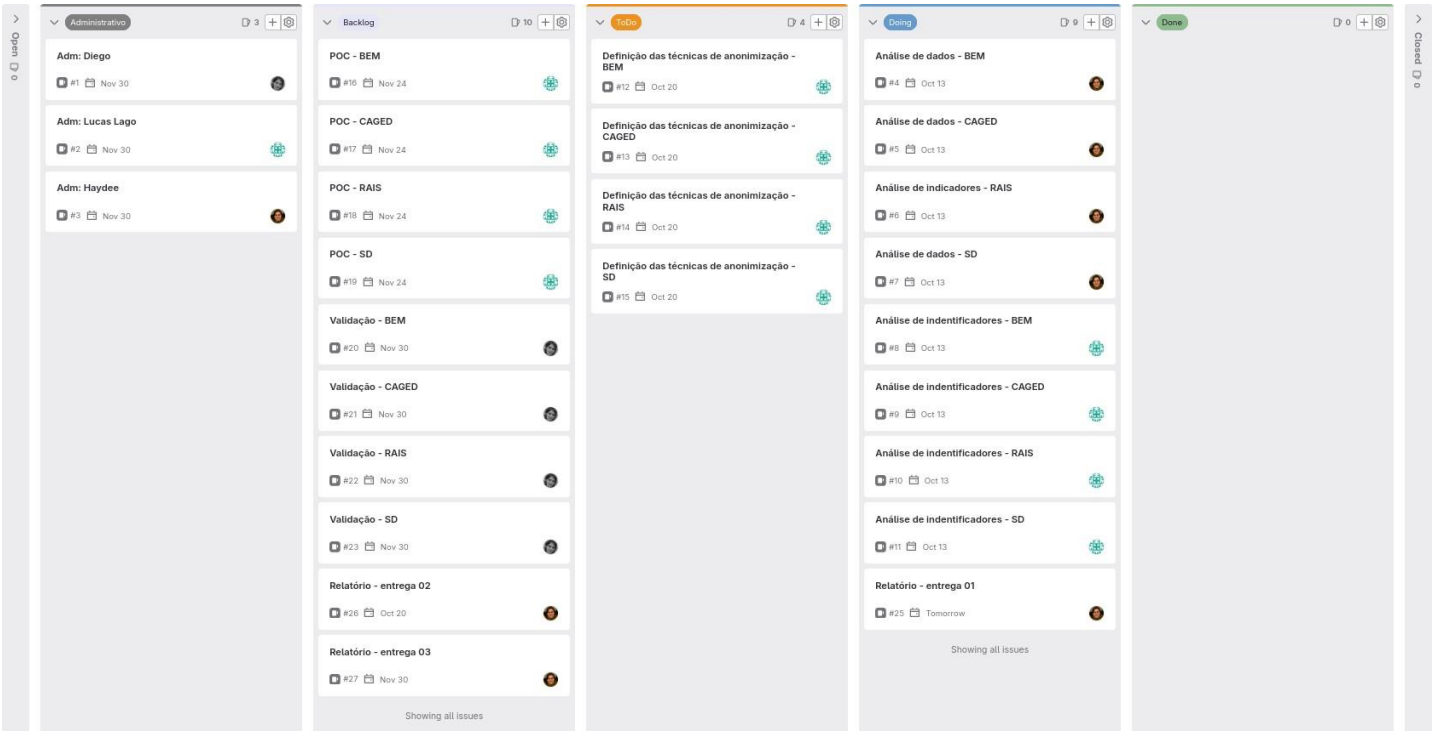


Figura 01 - Board do Gitlab das atividades previstas no planejamento - status em 21/09/2023

3 Definições de tecnologia

Para o desenvolvimento da Prova de Conceito (POC) foi escolhida a linguagem de programação python, tendo como principais motivos:

- quantidade de recursos próprios disponíveis para a finalidade do projeto;
- recursos existentes de forma integrada na própria linguagem;
- velocidade de desenvolvimento que permite;
- linguagem muito difundida entre desenvolvedores facilitando formação de equipes de desenvolvimento e manutenção futuras, tendo em vista a incorporação da POC.

Foi objetivo usar o mínimo de dependências/bibliotecas possível tornando a adoçãoda POC o mais fácil possível. Assim, foram utilizadas apenas:

- Pandas[4] para facilitar a leitura e a manipulação de dados tabulares;
- Linter[5] para implementar regras de estilo do Python, cobrindo as funcionalidades das seguintes ferramentas: flake8[6], pyflakes[7], pycodestyle[8], pylint[9] e black[10].

Como o objetivo é desenvolver uma POC e não um serviço efetivo, e considerando acesso a um volume de dados limitado (amostra reduzida) o foco não foi estabelecer um ambiente de produção.

[4] https://pandas.pydata.org/
[5] https://astral.sh/ruff
[6] https://flake8.pycqa.org/en/latest/
[7] https://pypi.org/project/pyflakes/
[8] https://pypi.org/project/pycodestyle/
[9] https://readthedocs.org/projects/pylint/
[10] https://black.readthedocs.io/en/stable/

4 Bancos de dados: análise preliminar

Foram analisadas amostras de 5 bases de dados, a saber:

- Benefício Emergencial de Preservação do Emprego e da Renda (BEM),
- Cadastro Geral de Empregados e Desempregados (CAGED),
- Relação Anual de Informações Sociais (RAIS) - estabelecimentos, e

- Relação Anual de Informações Sociais (RAIS) - vínculos, e
- Seguro Desemprego (SD).

Base de dados	Quantidade de variáveis	Quantidade de observações
BEM	112	73.178
CAGED	40	326.175
RAIS - estabelecimentos	31	36.710
RAIS - vínculos	78	293.617
SD	366	110.156

As 627 variáveis das 5 tabelas de dados foram avaliadas buscando:

1. **compreender seu significado**, dado que nem sempre os nomes são suficientemente auto-explicativos;
2. **classificar se a variável é um identificador direto**[11], ou seja, se contém uma informação pertinente ou possa ser atribuída diretamente a uma pessoa sem que seja preciso mais de um dado para identificar um indivíduo. Por exemplo, impressão digital ou número de identificação nacional como o CPF.
3. **classificar se a variável é um identificador indireto** (ou quase-identificador[12]),ou seja, se contém uma informação que possa estar vinculada a um indivíduo específico sendo necessárias informações adicionais para a identificação deum indivíduo. Por exemplo, o cruzamento de informações de sexo, estado civil e data de nascimento.

Um resumo do resultado da classificação das 5 bases de dados, com base nas amostras e informações é apresentado no quadro a seguir:

Base de dados	Quantidade de identificadores diretos individuais	Quantidade de identificadores indiretos[13]	Quantidade deatributos sensíveis[14]
BEM	11	27	0
CAGED	3	19	4
RAIS - estabelecimentos	3	10	0
RAIS - vínculos	6	41	5
SD	35	84	2

[11] Segundo SIMI, M. S.; NAYAKI, K. S.; ELAYIDOM, M. S. An extensive study on data anonymization algorithms based on k-anonymity. IOP Conference Series: Materials Science and Engineering, IOP Publishing, v. 225, p. 012279, aug 2017. Disponível em: https://iopscience.iop.org/article/10.1088/1757-899X/225/1/012279

[12] Segundo DOMINGO-FERRER, J.; SORIA-COMAS, J. From t-closeness to differential privacy and vice versa in data anonymization. Knowledge-Based Systems, v. 74, p. 151–158, 2015. ISSN 0950-7051. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0950705114004031.

5 Técnicas de anonimização e pseudo-anonimização

Há um equilíbrio a ser buscado, em especial por entidades de fins públicos, entre o respeito à privacidade e a promoção da transparência pública. À medida que a utilidade dos dados aumenta, a privacidade diminui e nem sempre de forma proporcional. Pequenos aumentos na utilidade podem gerar maiores reduções na privacidade, e pequenos aumentos na privacidade causam grandes reduções na utilidade[15,16]. Assim, para que haja alguma utilidade dos dados, estes serão imperfeitamente anônimos[17]. Tendo isso em vista, e compreendendo que nunca haverá uma anonimização 100% perfeita e infalível, recomenda-se seguir o ciclo avaliativo indicado na Figura 02 e considerar as possibilidades de técnicas de anonimização e pseudo-anonimização apresentadas na sequência.

Abaixo, segue descrição sobre as técnicas de anonimização conceituadas e aplicadas atualmente:

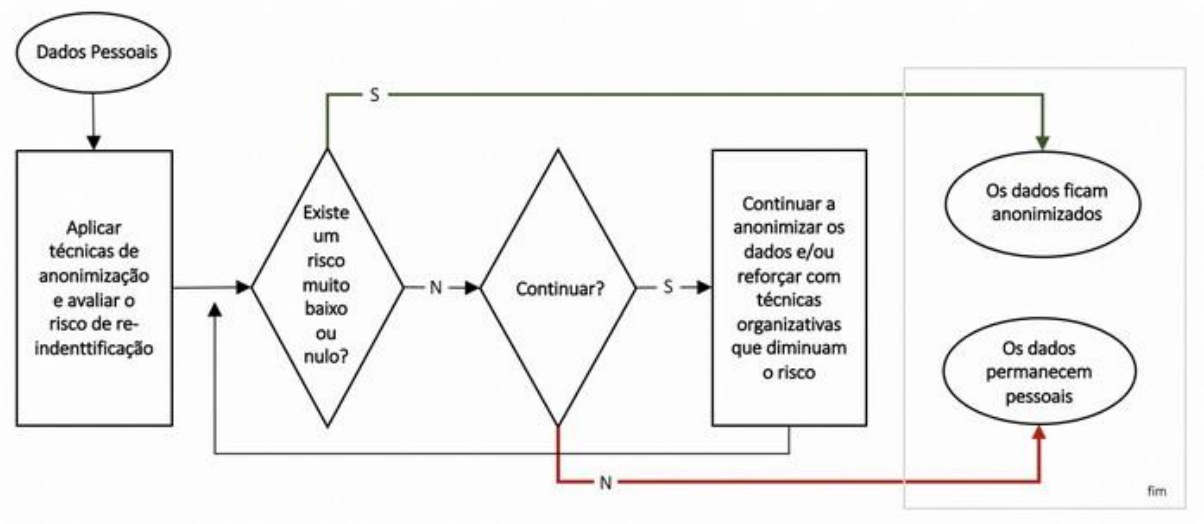


Figura 02 - Fluxograma de anonimização - Fonte: Universidade de Coimbra[18]

[13] Podem incluir, ou não, identificadores diretos.
[14] Segundo a Lei Geral de Proteção de Dados, Da LGPD é um dado pessoal sensível todo "dado pessoal sobre origem racial ou étnica, convicção religiosa, opinião política, filiação a sindicato ou a organização de caráter religioso, filosófico ou político,dado referente à saúde ou à vida sexual, dado genético ou biométrico, quando vinculado a uma pessoa natural"
[15] Brickell, J., & Shmatikov, V. (2008, August). The cost of privacy: destruction of data-mining utility in anonymized data publishing. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 70-78). Disponível em: https://www.cs.utexas.edu/~shmat/shmat_kdd08.pdf
[16] SHIN, S-Y.; KIM, H-S. (2021) Data pseudonymization in a range that does not affect data quality: Correlation with the degree of participation of clinicians. J Korean Med Sci, The Korean Academy of Medical Sciences, v. 36, n. 44. Disponível em: https://pubmed.ncbi.nlm.nih.gov/34783216/.
[17] NEGRI, S. M. C. D. Á.; GIOVANINI, C. F. R. (2020). Dados não pessoais: a retórica da anonimização no enfrentamento à covid-19 e o privacywashing. *Revista Internet e Sociedade*. Disponível em: https://revista.internetlab.org.br/dados-nao-pessoais-a-retorica-da-anonimizacao-no-enfrentamento-a-covid-19-e-o-privacywash ing/

[18] Universidade de Coimbra: Proteção de Dados Pessoais / Anonimização e Pseudonimização. Disponível em: https://www.uc.pt/protecao-de-dados-e-informacao-administrativa/protecao-de-dados-pessoais/anonimizacao-e-pseudonimizacao/

5.1 Encobrimento de caracteres

Descrição: Essa técnica é também chamada de mascaramento de dados e consiste em usar caracteres neutros, como o asterisco (*), para encobrir parcial ou totalmente as informações importantes e pessoais.

Exemplo: No caso de um CPF, os asteriscos podem preencher todos os espaços depois de um determinado número de caracteres.

Observação: É relevante avaliar se o comprimento do atributo (número de dígitos) traz informações relevantes sobre os dados originais.

Principais pontos fracos: Ainda se sabe o número de dígitos do dado original.

Pode permitir reidentificação: Sim

Utilizado na POC: Sim

5.2 Supressão

Descrição: Consiste em remover informação da base de dados.

Pode permitir reidentificação[19]: Não

5.2.1 De atributos (colunas)

Descrição: Consiste em remover os atributos da base de dados, suprimindo variáveis referentes a informações únicas.

Exemplo: Eliminação de variáveis como data de nascimento, nome completo, etc.

Observação: É indicada quando um atributo não é relevante, necessário ou quando não é possível anonimizá-lo de outra forma.

[19] Segundo MARQUES, J. F.; BERNARDINO, J. Analysis of data anonymization techniques. In: KEOD. [s.n.], 2020. p. 235–241. Disponível em: https://www.scitepress.org/Papers/2020/101423/101423.pdf

Principais pontos fracos: Possível perda de utilidade.

Utilizado na POC: Sim

5.2.2 De registros (linhas)

Descrição: Consiste em remover os registros da base de dados, suprimindo completamente a linha com informações de todos os atributos desse indivíduo.

Exemplo: Eliminação de registros que identificam unicamente indivíduos ao cruzar dados de gênero, raça/cor e grau de instrução de um determinado município.

Observação: É indicada para retirada de outliers e/ou quando é possível identificar unicamente indivíduos pelo cruzamento de algumas variáveis que, por sua relevância e/ou frequência, não podem ser eliminadas.

Principais pontos fracos: Possível perda de utilidade.

Utilizado na POC: sim

5.3 Generalização

Descrição: Consiste em transformar a escala ou magnitude dos dados. Quando são dados discretos, é possível generalizá-los agrupando em intervalos ou utilizando métricas de sumarização como média, mediana, entre outras.

Pode permitir reidentificação[20]: Não

5.3.1 Agregação

Descrição: Quando os registros individuais não são necessários e os dados agregados fornecem informação suficiente.

Exemplo:

ID	Cidade	Idade	Saldo
1	Natal	30	300,00
2	Belo Horizonte	43	450,00
3	Parnamirim	30	300,00
4	Sete Lagoas	40	500,00

Dados originais

ID	Cidade	Idade	Saldo
1	Natal	[30-39]	300,00
2	Belo Horizonte	[40-49]	450,00
3	Parnamirim	[30-39]	300,00
4	Sete Lagoas	[40-49]	500,00

(b) Dados agregados por faixa etária

Principais pontos fracos: perda de utilidade por conta da perda de granularidade da informação.

Utilizado na POC: sim, foi usada generalização estabelecendo faixas etárias, faixas salariais e transformando data no formato DD/MM/AAAA para MM/AAAA.

5.3.2 K-anonimato

Descrição: Garante que cada registro seja semelhante a, pelo menos, k−1 outros registros do conjunto de dados[21], com isso, cada registro tem probabilidade máximade 1/k de ser ligado a um indivíduo[22].

Exemplo:

ID	Cidade	Idade	Saldo
1	Natal	30	300,00
2	Belo Horizonte	43	450,00
3	Parnamirim	30	300,00
4	Sete Lagoas	40	500,00

5.3.2.1 Caso que não atende aos critériosdo 2-anonimato

ID	Cidade	Idade	Saldo
1	RN	[30-39]	300,00
2	MG	[40-49]	450,00
3	RN	[30-39]	300,00
4	MG	[40-49]	500,00

5.3.2.2 Caso não atende aos critérios do2-anonimato

Principais pontos fracos: ataque de ligação ao atributo. Esta técnica é sensível ao ataque em que se infira atributos sensíveis de um indivíduo mesmo sem o reidentificar.

Utilizado na POC: sim, foi estabelecido k = 2

5.3.3 L-diversidade

Descrição: deve ser aplicado a um conjunto de dados já k-anônimo e fornece proteção contra o ataque de ligação ao atributo ao qual o k-anonimato é sensível[23]**.**Cada classe de equivalência gerada pelo k-anonimato devem conter pelo menos L valores distintos para os atributos sensíveis, cuja probabilidade máxima de deduçãoé de 1/L[24]**.**

Exemplo:

ID	Cidade	Idade	Saldo
1	RN	[30-39]	300,00
2	MG	[40-49]	450,00
3	RN	[30-39]	300,00
4	MG	[40-49]	500,00

5.3.3.1 Caso que não atende aos critérios da2-diversidade

ID	Cidade	Idade	Saldo
1	RN	[30-39]	300,00
2	MG	[40-49]	450,00
3	RN	[30-39]	230,00
4	MG	[40-49]	500,00

5.3.3.2 Caso que atende aos critériosda 2-diversidade

[21] SWEENEY, L. Guaranteeing anonymity when sharing medical data, the datafly system. In: AMERICAN MEDICAL INFORMATICS ASSOCIATION. Proceedings of the AMIA Annual Fall Symposium. 1997. p. 51. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2233452/

[22] MEDKOVÁ, J. High-degree noise addition method for the k-degree anonymization algorithm. In: IEEE. 2020 Joint 11th International Conference on Soft Computing and Intelligent Systems and 21st International Symposium on Advanced Intelligent Systems (SCIS-ISIS). 2020. p. 1–6. Disponível em: https://ieeexplore.ieee.org/document/9322670.

[23] BRITO, F. T.; MACHADO, J. C. Preservação de privacidade de dados: Fundamentos, técnicas e aplicações. Jornadas de A tualização em Informática, 2017. Disponível em: https://www.researchgate.net/publication/318726149_Preservacao_de_Privacidade_de_Dados_Fundamentos_Tecnicas_e_Aplicacoes

[24] PINHO, F. A. Anonimização de bases de dados empresariais de acordo com a nova Regulamentação Europeia de Proteção de Dados. 2017. Tese (Doutorado) — Dissertação (Mestrado em Segurança Informática), Departamento de Ciência de

Principais pontos fracos: perda de utilidade, ataque de assimetria e ataque de similaridade.

Utilizado na POC: não

5.3.4 T-proximidade

Descrição: deve ser aplicado a um conjunto de dados já k-anônimo e L-diverso comobjetivo de superar a limitação em relação ao ataque de assimetria. Uma classe de equivalência tem t-proximidade quando a distância da distribuição do atributo sensível desta classe para a distribuição desse mesmo atributo em todo o conjuntode dados não for maior que um limite t definido.

Principais pontos fracos: perda de utilidade, falta de flexibilidade e o uso da métricaEMD[25] não é apropriada para ataques de ligação ao atributo quando os mesmos sãonuméricos[26].

Utilizado na POC: não

5.4 Aplicação de técnicas de criptografia

Descrição: Consiste em transformar os dados originais, legíveis, em versões impossíveis de compreender e utilizar. Para decodificar as informações é necessário acesso a uma chave que somente o detentor da base possui. Entre os diversos tiposde criptografia estão: chave simétrica, chave assimétrica, funções hash, *advanced encryption standard* (AES), *secure and faster encryption routine* (Safer), *data encryption standard* (DES) e *international data encryption algorithm* (Idea).

Computadores, Faculdade de Ciências, Universidade do Porto., 2020. Disponível em: https://cracs.fc.up.pt/sites/default/files/MSI_Dissertacao_FINAL.pdf.

[25] A métrica EMD (Earth Mover's Distance) mede a distância entre as distribuições da classe de equivalência e global.

[26] Rajendran, Jayabalan e Rana 2017 RAJENDRAN, K.; JAYABALAN, M.; RANA, M. E. A study on k-anonymity, l-diversity, and t-closeness techniques. IJCSNS, v. 17, n. 12, p. 172, 2017. Disponível em: https://www.ijirst.org/articles/IJIRSTV6I6015.pdf.

Observação: Criptografia pode ser forte ou fraca. A força de criptografia é medida a partir dos recursos (tempo, dinheiro, esforço computacional) exigidos para recuperaros dados. Com uma criptografia forte é muito difícil decifrar sem ter a ferramenta apropriada para decodificação. Contudo, a força da criptografia tende a decrescer com o tempo, dado que o poder computacional vem aumentando a cada dia.

Principais pontos fracos: É uma técnica reversível, classificada de pseudonimização, mas que pode ser combinada com outros métodos para gerar uma mudança completa e irreversível, a depender da necessidade.

Pode permitir reidentificação: sim

Utilizado na POC: sim, função hash com pepper

5.4.1 Hash

Descrição: As funções hash são algoritmos do tipo “one-way”, assim, em tese, não são possíveis de reverter o seu resultado. Também são determinísticas: dada uma certa entrada, ela produzirá sempre a mesma saída.

Observação: Os algoritmos de hash foram projetados para que algoritmos de reversão/quebra sejam lentos. Os hash de senha devem usar tantos recursos quanto possível para tornar os ataques de força bruta mais lentos e mais caros**.**

Principais pontos fracos: Apesar de ataques de força bruta serem custosos para cadeias de caracteres arbitrárias, no caso de cadeias de caracteres que possuem padrões o custo do ataque de força bruta é reduzido.

Utilizado na POC: sim, função hash com pepper

5.4.2 Hash com salt

Descrição: Consiste na utilização da função hash com um dado aleatório usado como uma entrada adicional. Em uma configuração típica, o *salt* e o valor original são concatenados,

processados pela função hash e a saída resultante é armazenada junto do salt em um banco de dados.

Observação: Protege bem contra ataques como como *rainbow table*, forçando o invasor a re-computar considerando o *salt* e bloqueando o paralelismo.

Principais pontos fracos: Não impedem ataques de força bruta ou dicionário.

Utilizado na POC: não

5.4.3 Hash com pepper

Descrição: Consiste na utilização da função hash com um dado aleatório usado como uma entrada adicional, similar ao *salt*. Mas, diferentemente do *salt*, esse valor aleatório é único para cada dado de entrada e não é armazenado junto com o *hash*. Além disso, o pepper deve ser mantido separado em algum outro meio, com controles rígidos de segurança. A maior vantagem de uma pepper é o fato de que ela não é mantida no banco de dados. Portanto, no caso de uma violação de dados, mesmo com acesso a todas as senhas com hash, o invasor ainda precisaria fazer força bruta no banco de dados.

Observação: O *pepper* desempenha uma função comparável à do *salt* ou de uma chave de criptografia, mas enquanto o *salt* não é secreto (apenas exclusivo) e pode ser armazenado junto com o resultado do *hash*, o *pepper* é secreto, único por dado de entrada e não deve ser armazenado com o resultado. Estas são as maiores vantagens do pepper: ser único por registro e não ser mantido no banco de dados. Logo, no caso de uma violação de dados, mesmo com acesso todos *hashs*, o invasor ainda precisaria fazer força bruta no banco de dados. E no caso de vazamento do pepper de um único dado, os demais hashes ainda assim estão protegidos.

Principais pontos fracos: É necessário gerar uma tabela de *peppers*, que não pode ser mudada e demanda gerenciamento adequado de acessos e política de segurança, um eventual vazamento dos dados dessa tabela torna esse processo equivalente ao processo de hash simples.

Utilizado na POC: sim, com um pepper exclusivo para cada indivíduo da base

5.5 Adição de ruídos aos dados

Descrição: Também chamada de perturbação de dados, consiste em mudar consistentemente as informações da base de acordo com alguns critérios, de modo a eliminar vestígios de identificação e garantir o uso para o objetivo principal. Geralmente aplicada a atributos numéricos ou datas, pode se dar por regra de arredondamento ou onde o valor original de um atributo v é substituído por v+r, sendo r o ruído adicionado[27].

Exemplo:

Pessoa	Altura (cm)	Peso (kg)
1	161	50
2	177	70
3	158	48
4	173	75
5	169	77
6	176	77

(a) Dados originais

Pessoa	Altura (cm)	Peso (kg)
1	160	50
2	180	70
3	160	50
4	175	75
5	170	80
6	175	75

(b) Dados arredondados

Observação: O grau de ruído deve ser proporcional à ordem de grandeza dos valores do atributo. Se for muito pequeno, o efeito de anonimização será mais fraco; por outro lado, se for muito grande, os valores finais serão muito diferentes dos originais, reduzindo a utilidade do conjunto de dados.

Principais pontos fracos: a utilidade pode ser reduzida pelo efeito do ruído.

Pode permitir reidentificação: Sim

Utilizado na POC: não

[27] VIRUPAKSHA, S.; DONDETI, V. Anonymized noise addition in subspaces for privacy preserved data mining in high dimensional continuous data. Peer-to-Peer Networking and Applications, Springer, v. 14, n. 3, p. 1608–1628, 2021. Disponível em: https://link.springer.com/article/10.1007/s12083-021-01080-v.

5.6 Privacidade Diferencial

Descrição: consiste em tornar os dados dos indivíduos anônimos adicionando um ruído às respostas das consultas realizadas na base de dados, fornecendo informações estatísticas sobre o conjunto de dados sem afetar a privacidade dos indivíduos.

Observação: Não fornece um conjunto de dados anonimizados, mas sim responde a uma consulta aos dados[28].

Principais pontos fracos: pode se tornar muito complexa de implementar e gerar muito ruído, gerando perda significativa da utilidade dos dados.

Pode permitir reidentificação: não

Utilizado na POC: não

5.7 Embaralhamento

Descrição: Também chamada de troca ou permutação, consiste em embaralhar valores do atributo perdendo a relação chave-valor, mas mantendo os dados originais e a distribuição individual por atributo. Funciona melhor em grandes conjuntos de dados[29].

Observação: Há a variação da técnica do embaralhamento em grupo. Embaralhamento em grupo é utilizado quando as informações agrupadas precisam ser anonimizadas em conjunto, onde um grupo de atributos é embaralhado.

Principais pontos fracos: O embaralhamento nem sempre fornece a anonimização dos dados, neste caso, deve ser utilizado em conjunto com outras técnicas. Esta técnica pode reduzir a utilidade do dado a depender da análise que se deseja fazer.

[28] SINGAPORE, P. D. P. C. Guide to basic data anonymisation techniques. 2018. Disponível em https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Guide-to-Anonymisation_v1-(250118).pdf

Pode permitir reidentificação: sim

Utilizado na POC: não

Exemplo:

Nome	Telefone	Idade
Mulher Maravilha	92435-7563	45
Batman	94712-5831	32
Homem de Ferro	98756-4286	52

(a) Dados originais

Nome	Telefone	Idade
Mulher Maravilha	94712-5831	52
Batman	98756-4286	45
Homem de Ferro	92435-7563	32

(b) Dados embaralhados

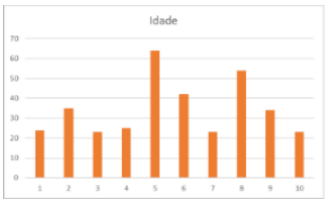
5.8 Uso de dados sintéticos

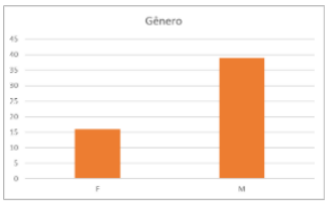
Descrição: Consiste na publicação de dados que respeitam a distribuição e algumas estatísticas, mas não informa valores reais. Os dados sintéticos são gerados por meio de modelos estatísticos construídos a partir do conjunto de dados original e buscam preservar suas propriedades estatísticas[30]. Por serem dados artificiais e não dizerem respeito a pessoas reais não estão sujeitos às mesmas restrições legais e de privacidade que os dados pessoais.

[29] Segundo SINGAPORE, P. D. P. C. Guide to basic data anonymisation techniques. 2018. Disponível em: https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Guide-to-Anonymisation_v1-(250118).pdf.
[30] Segundo JAMES, S. et al. Synthetic data use: exploring use cases to optimise data utility. Discover Artificial Intelligence, Springer, v. 1, n. 1, p. 1–13, 2021. Disponível em: https://link.springer.com/article/10.1007/s44163-021-00016-y.

Exemplo:

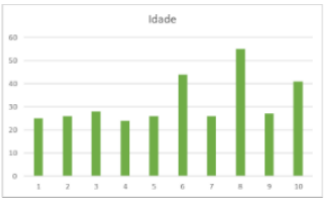
ID	Gênero	Idade
1	F	24
2	F	35
3	M	23
4	F	25
5	M	64
6	M	42
7	M	23
8	M	54
9	F	34
10	M	23

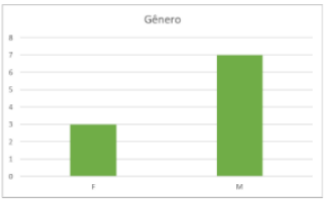




(a) Dados originais

ID	Gênero	Idade
1	M	25
2	M	26
3	F	28
4	F	24
5	F	26
6	M	44
7	M	26
8	M	55
9	M	27
10	M	41





(b) Dados sintéticos gerados através do pacote SDV[31]

Observação: Bastante útil para treino de algoritmos e testes de carga.

Principais pontos fracos: É necessário ter cuidado com os dados gerados para quenão contenham valores sem sentido para determinados atributos.

Pode permitir reidentificação: não

Utilizado na POC: não

5.9 Remoção do identificador direto

Descrição: É uma forma de supressão que consiste em apenas alterar a identificação direta de cada entrada. No lugar de cada identificador, é gerado um valor (código ou token) de forma aleatória ou determinista.

Observação: Essa supressão pode ser usada como um método de anonimização e como pseudonimização. Para ser considerada uma técnica de anonimização, é preciso usar tokens aleatórios e não referenciá-los em nenhum outro local. Assim, em tese, não há meios para identificar alguém. Para ser considerada uma técnica de pseudonimização, pode-se usar uma tabela associando os indivíduos aos valores aleatórios (pseudônimos) e usar esses valores em outras combinações de dados.

[31] SDV. The synthetic data vault.https://sdv.dev/

Exemplo:

Nome	Nota
Mulher Maravilha	7
Batman	5
Homem de Ferro	3

(a) Dados originais

Nome	Nota
15873248	7
45871325	5
97354850	3

(b) Dados anonimizados

Principais pontos fracos: havendo o desejo ou necessidade de reidentificação, a listade valores utilizados na geração dos pseudônimos necessita cuidados adicionais desegurança e privacidade como gestão adequada do controle de acesso.

Pode permitir reidentificação: sim, se os pseudônimos não forem gerados de forma aleatória.

Utilizado na POC: não

6. Prova de Conceito

6.1 Arquitetura

O projeto contém um arquivo/script principal (process.py) de execução do fluxo de anonimização, um arquivo de configuração para cada base de dados a ser anonimizada (arquivos *.json dentro da pasta settings) e um módulo python com osdiversos algoritmos de anonimização a serem utilizados (algoritmos.py). A Figura 03 detalha um pouco mais as componentes desta Prova de Conceito (POC).

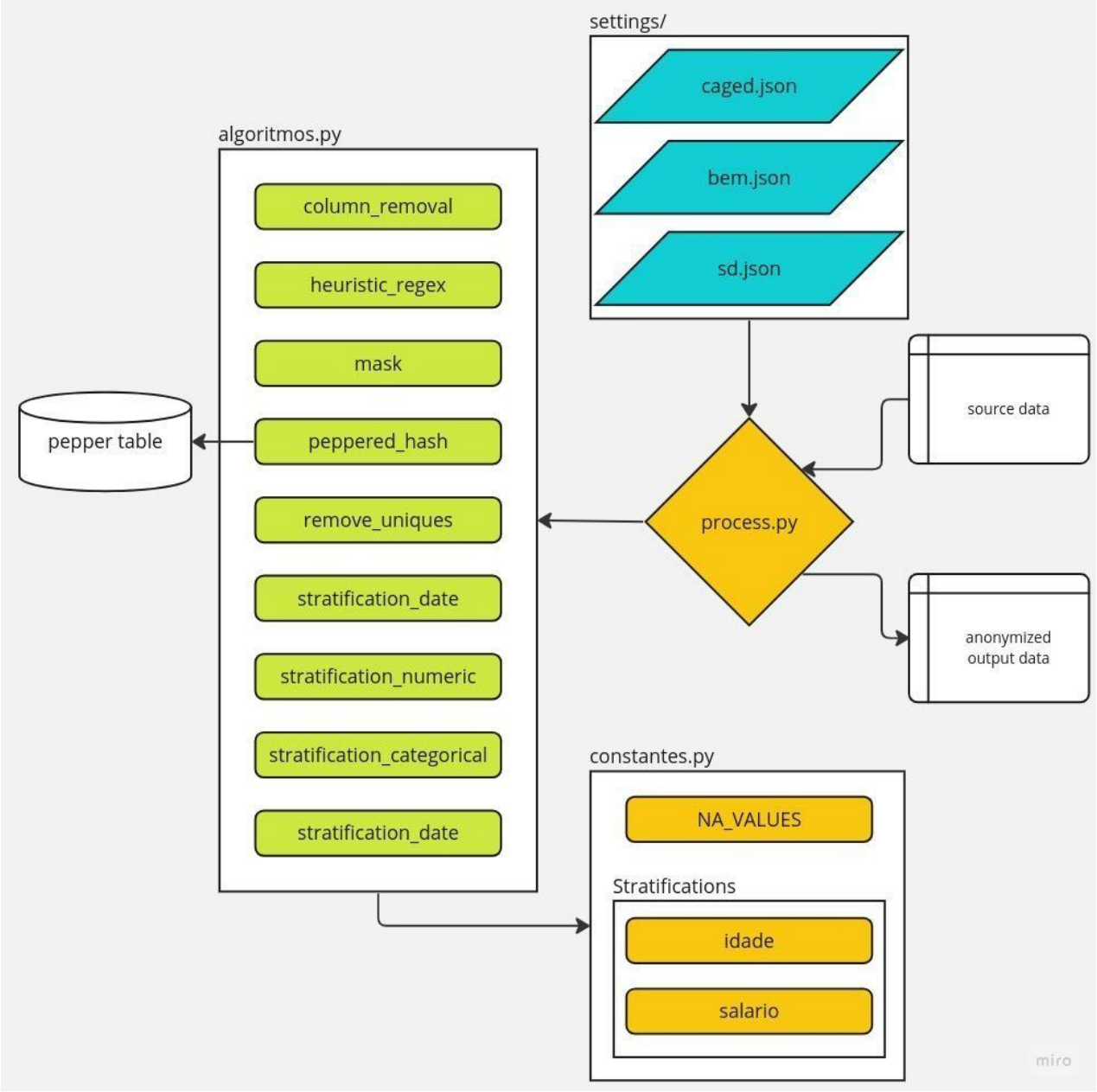


Figura 03 - Componentes da Prova de Conceito

6.2 Desenvolvimento e implementações

Todo o código desenvolvimento está disponível em repositório privado do GitLab cujos acessos são assegurados a responsáveis deste projeto a partir da URL: https://dieese.gitlab.io/-/piloto-anonimizacao-bases-2023/-/jobs/5698845772/artifa cts/html/algoritmos.html.

6.2.1 process.py

O script principal recebe como argumento o “apelido” da base de dados a ser anonimizada naquela execução (p.ex.: caged, sd, etc).

6.2.2 settings

Os arquivos de configuração de cada base são no formato json, se encontram nodiretório settings e contém as seguintes informações:

- file: (ex: "files/Dados_CAGED_Anonimizaca_2022_2023.csv")
- file_out: (ex: "files/Dados_CAGED_Anonimizados.csv")
- separator: Qual o separador utilizado no arquivo de dados de origem (em geral vírgula ou ponto e vírgula)
- decimal: Qual o separador de casas decimais utilizado no arquivo de dados de origem (em geral vírgula ou ponto)
- dtypes: mapeamento entre nome de variável e tipo da variável (float, int, string, etc) para leitura dos dados de entrada. Não é necessário listar todas asvariáveis, somente aquelas que podem ser interpretadas de forma errônea naleitura, como por exemplo o CPF (que deve ser lido como string e não como número)
- columns: lista contendo cada uma das colunas que precisam passar por algum procedimento de “anonimização”, com as respectivas configurações de anonimização.

  O mínimo de informações necessárias para cada item é:

  - nome: nome da coluna
  - algoritmo: qual algoritmo de anonimização será aplicado.
  - configuration: dicionário com as configurações daquele algoritmo.

6.2.3 algoritmos.py

A seguir é apresentada a documentação automatizada gerada a partir do código implementado para a POC, contendo as funções de anonimização, suas variáveis e retorno das mesmas.

6.2.3.1 Mask

Descrição

Mascara o conteúdo de uma coluna: Aplica uma máscara em um determinado percentual do texto de um determinada coluna/variável/atributo. A máscara pode ser aplicada da esquerda para a direita ou da direita para a esquerda. O caractere usado na máscara também é passado de via configuração.

Argumentos

- df: pandas dataframe com os dados de referência/origem;
- column_name: nome da coluna de referência no dataframe;
- configuration: dicionário contendo as respectivas configurações para a execução desta função. Aqui são esperadas as seguintes entradas:
    - mask_direction: Direção na qual a máscara será aplicada (left/right). Se "right", seleciona os x% primeiros caracteres da string. Se "left", seleciona os "x%" últimos caracteres da string.
    - mask_char_size: Quantos caracteres serão ofuscados. Número inteiro,substitui o mask_percentage.
    - mask_percentage: Qual o percentual que será utilizado. Número inteirode 0 a 100.
    - mask_character: Qual o carácter que será utilizado na máscara e substituirá os caracteres originais da string.
    - boolean_column: Coluna usada como controle para saber se a máscara deve ou não ser aplicada em cada célula. Ao iterar pelas linhas, olha esta coluna booleana para saber se a máscara é aplicada ou se o valor original é retornado. Com isso pode-se ter um controle "linha a linha" da aplicação da máscara.
    - in_place: variável booleana para saber se devemos sobrescrever ou não a coluna original no dataframe de saída. Se não sobrescrevermos,os dados com hash irão para uma nova coluna que será criada com o nome "anon_" (default: True).

RETORNO

Dataframe com a coluna/variável/atributo solicitada mascarada.

6.2.3.2 Column Removal

Descrição

Remove a coluna/variável/atributo especificada do dataframe final.

Argumentos

- **df**: pandas dataframe com os dados de referência/origem;
- **column_name**: nome da coluna de referência no dataframe;
- **configuration**: não é necessária nenhuma configuração.

Retorno

Dataframe com a coluna/variável/atributo removida.

6.2.3.3 Remove Uniques

Descrição

Remove as linhas/ registros identificados univocamente pelas colunas/variáveis/atributos passados.
Dada uma coluna/variável/atributo (column_name) e uma lista de colunas/variáveis/atributos pareadas (paired_columns), remove todas as linhas/registros que possuem estas tuplas como únicas no dataframe.

Argumentos

- **df**: pandas dataframe com os dados de referência/origem;
- **column_name**: nome da coluna de referência no dataframe;
- **configuration**: dicionário contendo as respectivas configurações para aexecução desta função. Aqui é esperada a seguinte entrada:
    - **paired_columns**: lista de colunas a serem pareadas com a colunaprincipal passada.

Retorno

Dataframe com as mesmas coluna/variável/atributo originais mas sem as linhas/registros removidas de acordo com o critério de unicidade.

6.2.3.4 Stratification Numeric

Descrição

Estratifica dados numéricos como forma de generalização.

Dada uma lista de faixa de valores, com os respectivos intervalos, cria uma nova coluna/variável/atributo que aloca os valores originais às faixas passadas.

Argumentos

- **df**: pandas dataframe com os dados de referência/origem;
- **column_name**: nome da coluna de referência no dataframe;
- **configuration**: dicionário contendo as respectivas configurações para a execução desta função. Aqui são esperadas as seguintes entradas:
    - **stratification_type**: Um dos tipos pré-definidos de estratificação que podem ser encontrados no arquivo constants.py. O uso desta variável éexcludente com a variável seguinte.
    - **stratification_list**: lista de estratos para construir a estratificação. Cada item desta lista (estrato) é um dicionário com três chaves:
        - **name**: Nome do estrato;
        - **min_value**: valor mínimo do intervalo do estrato;
        - **max_value**: valor máximo do intervalo do estrato.
    - **in_place**: variável booleana para saber se devemos sobrescrever ou não a coluna original no dataframe de saída. Se não sobrescrevermos,os dados com hash irão para uma nova coluna que será criada com o nome "anon_" (default: True).

Retorno

Dataframe original com a coluna/variável/atributo "column_name" modificada.

6.2.3.5 Stratification Date

Descrição

Estratifica datas como forma de generalização.

Converte uma data no formato "dia/mês/ano" para "mês/ano" ou "ano", a depender da configuração passada.

Argumentos

- **df**: pandas dataframe com os dados de referência/origem;
- **column_name**: nome da coluna de referência no dataframe;
- **configuration**: dicionário contendo as respectivas configurações para aexecução desta função. Aqui são esperadas as seguintes entradas:
    - **format**: Formato final da "data", com as seguintes opções:
        - "**mes-ano**": Mantém mês e ano;
        - "**ano**": Mantém apenas ano.
    - **in_place**: variável booleana para saber se devemos sobrescrever ou não a coluna original no dataframe de saída. Se não sobrescrevermos,os dados com hash irão para uma nova coluna que será criada com o nome "anon_" (default: True).

Retorno

Dataframe original com a coluna/variável/atributo "column_name" modificada.

6.2.3.6 Peppered Hash

Descrição

Aplica *peppered hash* numa determinada coluna/variável/atributo.

Esta função cria ou usa um arquivo de *pepper hash* para codificar uma determinada coluna/variável/atributo do dataframe de origem.

Ao final da execução é criada uma nova coluna com os valores codificados (*hashed*)seguindo o algoritmo *peppered hash*. A coluna/variável/atributo original pode ou não ser mantida no resultado final a depender da configuração passada.

Argumentos

- **df**: pandas dataframe com os dados de referência/origem;
- **column_name**: nome da coluna de referência no dataframe;
- **configuration**: dicionário contendo as respectivas configurações para a execução desta função. Aqui são esperadas as seguintes entradas:
    - **size**: tamanho fixo a ser considerado para a chave da tabela intermediária de hash;
    - **pepper_file**: arquivo csv onde será salva a tabela intermediária de hash;
    - **allow_empty**: Se permite ou não manter valores nulos/vazios no resultado codificado;
    - **in_place**: variável booleana para saber se devemos sobrescrever ou não a coluna original no dataframe de saída. Se não sobrescrevermos,os dados com hash irão para uma nova coluna que será criada com o nome "anon_" (default: True).

Retorno

Dataframe original com a adição da coluna de dados codificados, nomeada como "hash_". A coluna/variável/atributo original é mantida ou removida conforme aconfiguração passada.

6.2.3.2 Heuristic Regex

Descrição

Identifica se uma determinada coluna/variável/atributo se enquadra numa expressão regular.

Itera pelas linhas de um dataframe verificando se o valor ali presente é identificado pela expressão regular passada. Em geral esta função será utilizada para criar uma variável de controle booleana sobre uma outra variável.

Na presente POC é usada como função auxiliar de outros algoritmos, mas pode ser usada sozinha caso convenha.

Argumentos

- **df**: pandas dataframe com os dados de referência/origem;
- **column_name**: nome da coluna de referência no dataframe;
- **configuration**: dicionário contendo as respectivas configurações para a execução desta função. Aqui são esperadas as seguintes entradas:
    - **pattern**: Expressão regular a ser utilizada para match;
    - **in_place**: variável booleana para saber se devemos sobrescrever ou não a coluna original no dataframe de saída. Se não sobrescrevermos,os dados com hash irão para uma nova coluna que será criada com o nome "anon_" (default: True).

Retorno

Dataframe original com uma coluna adicional (heuristic_{column_name}), booleana, indicando se os registros se enquadram ou não na expressão regular passada.

6.2.4 Aplicações dos algoritmos

A seguir estão indicados os **métodos e algoritmos a serem empregados em cada uma das variáveis, de cada base** trabalhada. Cabe esclarecer que os identificadores indiretos são analisados em pequenos conjuntos e a abordagem indicada também deve ser analisada considerando esses pequenos conjuntos de atributos. Com asterisco (*) estão indicados os atributos que também são classificados como sensíveis.

Em relação aos identificadores diretos, à exceção do CPF e da razão social do empregador, a recomendação foi pela exclusão do atributo (variável). No caso dos CPFs foi sugerido utilizar o algoritmo de criptografia hash com pepper e não a remoção porque é uma variável que serve de chave para o cruzamento de diversas bases. No caso de razão social de pessoa jurídica, só é um problema quando se trata de MEI que, por um período, teve como regra de formação do nome a utilizaçãode nome e CPF da pessoa física responsável. Reconhecendo a importância

dessa variável, tanto do ponto de vista da publicidade quanto como ferramenta para cruzamento entre bases, não se recomendou sua exclusão, mas sim o mascaramento dos registros específicos em que a razão social for composta por um texto seguido por um conjunto de 11 algarismos**.**

Nos conjuntos de **identificadores indiretos** foram adotados os seguintes critérios de generalização:

- datas que apresentam-se no formato DD/MM/AAAA para o formatoMM/AAAA;
- idades agregadas em faixas etárias utilizadas pelo IBGE para produção desua pirâmide etária[32];
- salários (em quaisquer tempos ou datas) em faixas salariais (em salários mínimos ou valores proporcionais)[33].

6.2.4.1 BEM

- **Identificador direto individual**

Variável	Método de anonimização	Algoritmo
CPF Requerente	Aplicação de técnicas de criptografia	peppered_hash
CPF Usuário Portal	Aplicação de técnicas de criptografia	peppered_hash
Data Nascimento	Supressão de atributos	column_removal
Data Nascimento RFB	Supressão de atributos	column_removal
Matrícula eSocial	Supressão de atributos	column_removal
NIT Requerente	Supressão de atributos	column_removal
Nome Mãe Requerente	Supressão de atributos	column_removal
Nome Requerente	Supressão de atributos	column_removal
Número Conta DV	Supressão de atributos	column_removal
Número Requerimento	Supressão de atributos	column_removal
Número da Conta	Supressão de atributos	column_removal

- **Identificador indireto**

Variável	identificador indireto	Método de anonimização	Algoritmo
Cód. Cnae 2.0	sim	Supressão de atributos	column_removal
Cód. Município	sim, em conjunto com	-	
Empresa	subclasse/cbo		
Data Acordo	sim, em conjunto com subclasse/cbo ou Classe CNAE2.0	Generalização, disponibilizando apenas mês/ano	stratification_date
Data Admissão	sim, em conjunto com subclasse/cbo ou Classe CNAE2.0	Generalização, disponibilizando apenas mês/ano	stratification_date
Data Finalização Acordo	sim, em conjunto com subclasse/cbo ou Classe CNAE2.0	Generalização, disponibilizando apenas mês/ano	stratification_date
MunicípioEmpresa	sim, em conjunto com subclasse/cbo	-	
Núm Inscrição Empregador	sim, em conjunto com Cód. Tipo Empregador ou Tipo Empregador	Caso o Cód. Tipo Empregador seja 5 ou Tipo Empregador sejaCEI, aplicar criptografia no Núm Inscrição Empregador do registro específico	peppered_hash
Cód. Tipo Empregador	sim, em conjunto com Núm Inscrição Empregador ou Tipo Empregador	-	-
Tipo Empregador	sim, em conjunto com Núm Inscrição Empregador ou Cód. Tipo Empregador	-	-
Subclasse CNAE 2.0	sim, em conjunto com município	Supressão de atributos	column_removal
Valor Antepenúltimo Salário	sim, com cnpjraiz ou cnpjcei	Generalização, estabelecendo faixa salarial	stratification_numeric
Valor Antepenúltimo Salário CNIS	sim, com cnpjraiz ou cnpjcei	Generalização, estabelecendo faixa salarial	stratification_numeric
Valor Penúltimo Salário	sim, com cnpjraiz ou cnpjcei	Generalização, estabelecendo faixa salarial	stratification_numeric
Valor Penúltimo Salário CNIS	sim, com cnpjraiz ou cnpjcei	Generalização, estabelecendo faixa salarial	stratification_numeric

Valor Último Salário	sim, com cnpjraiz ou cnpjcei	Generalização, estabelecendo faixa salarial	stratification_numeric
Valor Último Salário CNIS	sim, com cnpjraiz ou cnpjcei	Generalização, estabelecendo faixa salarial	stratification_numeric

6.2.4.2 CAGED

- **Identificador direto individual**

Variável	Método de anonimização	Algoritmo
nit	Supressão de atributos	column_removal
gpf	Aplicação de técnicas de criptografia	peppered_hash
datanascimento	Supressão de atributos	column_removal

- **Identificador indireto**

Variável	Identificador indireto	Método de anonimização	Algoritmo
município	sim, em conjunto com subclasse/cbo	-	-
subclasse	sim, em conjunto com município	Encobrimento de caracteres com máscara 4 dígitos à direita	mask
cbo2002ocupação	sim, em conjunto com município	Encobrimento de caracteres com máscara 2 dígitos à direita	mask
Grau de instrução	sim, com sexo, raçacor, idade e alguma info de ocupação	-	-
sexo	sim, com raçacor, idade, grau de instrução e alguma informação de ocupação	-	-
idade	sim, com sexo, raçacor, grau de instrução e alguma informação de ocupação	Generalização, estabelecendo faixa etária	stratification_numeric
Tipo de deficiência *	sim, com sexo, raçacor, grau de instrução, idade e alguma informação de ocupação	Supressão de registros combinada com: faixa etária, raçacor, sexo, município	remove_uniques
Tipo empregador	sim	-	-

Variável	Identificador indireto	Método de anonimização	Algoritmo
cnpjraiz	sim	Caso o tipoempregador seja == 2, aplicar criptografia no cnpjraizdo registro específico	peppered_hash
tipoestabelecime nto	sim	-	-
cnpjcei	sim	Caso o tipoestabelecimento seja == 3 ou ==5, aplicar criptografia no cnpjraiz do registro específico	peppered_hash
Indicador aprendiz *	sim, com cnpjraiz ou cnpjcei, sexo ou raçacor	Supressão de registros combinada com: faixa etária, raçacor, sexo e cnpjraiz ou cnpjcei	remove_uniques
salário	sim, com cnpjraiz ou cnpjcei	Generalização, estabelecendo faixa salarial	stratification_numeric
Valo rsalário fixo	sim, com cnpjraiz ou cnpjcei	Generalização, estabelecendo faixa salarial	stratification_numeric
País de nacionalid	sim, com cnpjraiz ou cnpjcei	Supressão de registros	remove_uniques
Idade*	e/ou município	combinada com: com cnpjraiz ou cnpjcei e/ou município	
País de origem*	sim, com cnpjraiz ou cnpjcei e/ou município	Supressão de registros combinada com: com cnpjraiz oucnpjcei e/ou município	remove_uniques

6.2.4.1 RAIS-estabelecimentos

- **Identificador direto individual**

Variável	Método de anonimização	Algoritmo
razãosocial	Encobrimento completo de caracteres se a razãosocial da empresa for composta por "TEXTO" seguido de "CONJUNTO DE 11 NÚMEROS"	mask
emailestabelecimento	Supressão de atributos	column_removal
númerotelefoneempresa	Supressão de atributos	column_removal

- **Identificador indireto**

Variável	Identificador indireto	Método de anonimização	Algoritmo
cnpjcei	sim	Caso o naturezajurídica seja == 4030 e 4049, aplicar criptografiano cnpjcei do registro específico	peppered_hash
cnpjraiz	sim	Caso o naturezajurídica seja == 4030 e 4049, aplicar criptografia no cnpjraiz do registro específico	peppered_hash
Naturezaj urídica	sim	-	-
Cep estab	sim, em conjunto com nomelogradouro e/ou númerologradouro, para MEIs*	-	-
Nome logradouro	sim, em conjunto com cepestab e/ou númerologradouro, para MEIs*	Supressão de atributos	column_removal
Número logradouro	sim, em conjunto com cepestab e/ou nomelogradouro, para MEIs*	Supressão de atributos	column_removal
cnae20subclasse	sim, em conjunto com município	Supressão de atributos	column_removal

[32] Referência das faixas etárias para pirâmide etária do IBGE: https://educa.ibge.gov.br/jovens/conheca-o-brasil/populacao/18318-piramide-etaria.html

[33] Foram adotadas as faixas salariais a partir do Caderno Observatório Nacional do Mercado de Trabalho (2015), Tabela 3, p.55 (disponível em: https://www.dieese.org.br/livro/2016/cadernoObservatoNacionalVol1/index.html?page=56), e tomados os valores de salário mínimo de R$1.320 segundo a Lei 14.663/2023 (disponível em: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2023/Lei/L14663.htm).

6.2.4.2 RAIS-vínculos

- **Identificador direto individual**

Variável	Método de anonimização	Algoritmo
pis	Supressão de atributos	column_removal
datadenascimento	Supressão de atributos	column_removal
númeroctps	Supressão de atributos	column_removal
cpf	Aplicação de técnicas de criptografia	peppered_hash
nometrabalhador	Supressão de atributos	column_removal
razãosocial	Encobrimento completo de caracteres se a razãosocial da empresa for composta por "TEXTO" seguido de "CONJUNTO DE 11 NÚMEROS"	mask

- **Identificador indireto**

Variável	Identificador indireto	Método de anonimização	Algoritmo
Escolaridade após 2005	sim, em conjunto com sexotrabalhador, nacionalidade, raçacor, idade, anochegadabrasil, indportadordefic, tipodefic e dataadmissãodeclarada	-	-
Sexo trabalhador	sim, em conjunto com escolaridadeapós2005, nacionalidade, raçacor, idade, anochegadabrasil, indportadordefic, tipodefic e dataadmissãodeclarada	-	-
nacionalidade*	sim, em conjunto com escolaridadeapós2005, sexotrabalhador, açacor, idade, anochegadabrasil, indportadordefic, tipodefic e dataadmissãodeclarada	Supressão de registros combinada com: raçacor, idade, anochegadabrasil, tipodefic e dataadmissãodeclarada	remove_uniques
Raça cor*	sim, em conjunto com escolaridadeapós2005, sexotrabalhador, nacionalidade, idade, anochegadabrasil, indportadordefic, tipodefic e dataadmissãodeclarada	Supressão de registros combinada com: nacionalidade, idade, anochegadabrasil, tipodefic e dataadmissãodeclarada	remove_uniques

Variável	Identificador indireto	Método de anonimização	Algoritmo
idade	sim, em conjunto com escolaridadeapós2005, sexotrabalhador, nacionalidade, raçacor, anochegadabrasil, indportadordefic, tipodefic e dataadmissãodeclarada	Generalização, estabelecendo faixa etária	stratification_numeric
Ano chegada brasil	sim, em conjunto com escolaridadeapós2005, sexotrabalhador, nacionalidade, raçacor, idade, indportadordefic, tipodefic e dataadmissãodeclarada	Supressão de registros combinada com: nacionalidade, raçacor, idade,tipodefic e dataadmissãodeclarada	remove_uniques
Ind portador defic*	sim, em conjunto com escolaridadeapós2005, sexotrabalhador, nacionalidade, raçacor, idade, anochegadabrasil, tipodefic e dataadmissãodeclarada	-	-
Tipo defic*	sim, em conjunto com escolaridadeapós2005, sexotrabalhador, nacionalidade, raçacor, idade, anochegadabrasil, indportadordefic e dataadmissãodeclarada	Supressão de registros combinada com: nacionalidade, raçacor, idade,anochegadabrasil e dataadmissãodeclarada	remove_uniques
Data admissão de-clarada	sim, em conjunto com escolaridadeapós2005, sexotrabalhador, nacionalidade, raçacor, idade, anochegadabrasil, indportadordefic e tipodefic	Generalização, disponibilizando apenas mês/ano	stratification_date
Vlr em un médiano m	sim, com cnpjraiz ou cnpjcei	Generalização, estabelecendo faixa salarial	stratification_numeric
Vlr em un médias m	sim, com cnpjraiz ou cnpjcei	Generalização, estabelecendo faixa salarial	stratification_numeric
Vlr em un dezembr o nom	sim, com cnpjraiz ou cnpjcei	Generalização, estabelecendo faixa salarial	stratification_numeric
Vlr em un dezembr o sm	sim, com cnpjraiz ou cnpjcei	Generalização, estabelecendo faixa salarial	stratification_numeric
Vl última remunera ção ano	sim, com cnpjraiz ou cnpjcei	Generalização, estabelecendo faixa salarial	stratification_numeric
Vl salário contratua l	sim, com cnpjraiz ou cnpjcei	Generalização, estabelecendo faixa salarial	stratification_numeric
Cei vinculado	sim	Caso o naturezajurídica seja == 4030 e 4049, aplicar criptografiano cnpjcei do registro específico	peppered_hash
Cnpj cei	sim	Caso o naturezajurídica seja == 4030 e 4049, aplicar criptografiano cnpjcei do registro específico	peppered_hash

Variável	Identificador indireto	Método de anonimização	Algoritmo
Cnpj raiz	sim	Caso o naturezajurídica seja == 4030 e 4049, aplicar criptografia no cnpjraiz do registro específico	peppered_hash
Naturezaj urídica	sim	-	-
Mun trab	sim, em conjunto com subclasse e/ou cboocupação2002	-	-
município	sim, em conjunto com cnae20subclasse	-	-
cboocupação200 2	sim, em conjunto com muntrabou município	Encobrimento de caracteres com máscara 2 dígitos à direita	mask
cnae20subclasse	sim, em conjunto com muntrabou município	Supressão de atributos	column_removal
cepestab	sim, em conjunto com nomelogradouro e/ou númerologradouro, para MEIs* (na RAIS estabelecimentos)	-	-
Vlr em janeiros sc	sim, com cnpjraiz ou cnpjcei	Generalização, estabelecendo faixa salarial	stratification_numeric
Vlr emf evereiro sc	sim, com cnpjraiz ou cnpjcei	Generalização, estabelecendo faixa salarial	stratification_numeric
Vlr em março sc	sim, com cnpjraiz ou cnpjcei	Generalização, estabelecendo faixa salarial	stratification_numeric
Vlr em abril sc	sim, com cnpjraiz ou cnpjcei	Generalização, estabelecendo faixa salarial	stratification_numeric
Vlr em maio sc	sim, com cnpjraiz ou cnpjcei	Generalização, estabelecendo faixa salarial	stratification_numeric
Vl rem junho sc	sim, com cnpjraiz ou cnpjcei	Generalização, estabelecendo faixa salarial	stratification_numeric
Vlr em julho sc	sim, com cnpjraiz ou cnpjcei	Generalização, estabelecendo faixa salarial	stratification_numeric
Vlr em agosto sc	sim, com cnpjraiz ou cnpjcei	Generalização, estabelecendo faixa salarial	stratification_numeric
Vlr em setembro sc	sim, com cnpjraiz ou cnpjcei	Generalização, estabelecendo faixa salarial	stratification_numeric
Vlr em outubro sc	sim, com cnpjraiz ou cnpjcei	Generalização, estabelecendo faixa salarial	stratification_numeric
Vlr em novembro sc	sim, com cnpjraiz ou cnpjcei	Generalização, estabelecendo faixa salarial	stratification_numeric

6.2.4.3 SD

- **Identificador direto individual**

Variável	Método de anonimização	Algoritmo
agentedigitação	Supressão de atributos	column_removal
agenterecepção	Supressão de atributos	column_removal
endereçorequerente	Supressão de atributos	column_removal
logradourorequerente	Supressão de atributos	column_removal
cpfrequerente	Aplicação de técnicas de criptografia	peppered_hash
datanascimento	Supressão de atributos	column_removal
matragentedigitação	Supressão de atributos	column_removal
nitpisseguradoespecial	Supressão de atributos	column_removal
nomemãerequerente	Supressão de atributos	column_removal
nomerequerente	Supressão de atributos	column_removal
númeroctps	Supressão de atributos	column_removal
numerosentençajudicial*	Supressão de atributos	column_removal
númerosentençajudicial*	Supressão de atributos	column_removal
pispasepnit	Supressão de atributos	column_removal
telefonerequerente	Supressão de atributos	column_removal
agenteliberaçãoúltimanotiffase2	Supressão de atributos	column_removal
cpfgestorempweb	Supressão de atributos	column_removal
cpfprocuradorempweb	Supressão de atributos	column_removal
códagenteliberaçãoúltimanotiffas	Supressão de atributos	column_removal
códagentedigitação	Supressão de atributos	column_removal
nomegestorempweb	Supressão de atributos	column_removal
nomeprocuradorempweb	Supressão de atributos	column_removal
numeroctps	Supressão de atributos	column_removal
pispasepnitbase	Supressão de atributos	column_removal
pispasepprocuradorempweb	Supressão de atributos	column_removal
razãosocialempregador	Encobrimento de caracteres	mask
telefonegestorempweb	Supressão de atributos	column_removal
telefoneprocuradorempweb	Supressão de atributos	column_removal
usuáriogestorempweb	Supressão de atributos	column_removal
matragenterecepção	Supressão de atributos	column_removal
ceirequerente	Supressão de atributos	column_removal
identidaderequerente	Supressão de atributos	column_removal

Variável	Método de anonimização	Algoritmo
logradouro	Supressão de atributos	column_removal
nitpisrequerente	Supressão de atributos	column_removal
númerorgp	Supressão de atributos	column_removal

- **Identificador indireto**

Variável	Identificador indireto	Método de anonimização	Algoritmo
Cep requerente	sim, em conjunto com endereçorequerente	-	-
Número série ctps	sim, em conjunto com cpfrequerente	-	-
Cod grau instrucao	sim, em conjunto com grauinstrucao, codgenero, genero, raçarequerente e datanascimento	-	-
Grau instrucao	sim, em conjunto com codgrauinstrucao, codgenero, genero, raçarequerente e datanascimento	-	-
Cod genero	sim, em conjunto com codgrauinstrucao, grauinstrucao, genero, raçarequerente e datanascimento	-	-
genero	sim, em conjunto com codgrauinstrucao, grauinstrucao, codgenero, raçarequerente e datanascimento	-	-
Raça requerente*	sim, em conjunto com codgrauinstrucao, grauinstrucao	Supressão de registros	remove_uniques
Códa gente rece pç ão	sim	Supressão de atributos	column_removal
Ocupação cbo atual	sim, em conjunto com municipioresidencia ou codmunicipioresidencia	Supressão de atributos	column_removal
Subgrupo cbo	sim	Supressão de atributos	column_removal
Ocupação cbo	sim, em conjunto com municipioresidencia ou codmunicipioresidencia	Supressão de atributos	column_removal
Cód ocupação cbo	sim	Encobrimento de caracteres com máscara 2 dígitos à direita	mask
Cod municipio residencia	sim, em conjunto com códocupaçãocbo ou ocupaçãocbo	-	-

Variável	Identificador indireto	Método de anonimização	Algoritmo
Municipio residencia	sim, em conjunto com códocupaçãocbo ou ocupaçãocbo	-	-
Data admissão	sim, em conjunto com códocupaçãocbo ou ocupaçãocbo	-	-
Data demissão requerente	sim, em conjunto com códocupaçãocbo ou ocupaçãocbo	-	-
Data sentença judicial	sim, cruzando com base do jusrasil, por exemplo	Generalização, estabelecendo a competência como mês/ano	stratification_date
modalidade	sim, em conjunto com númeroinscriçãoempregador	-	-
Número inscriç o empregador	sim, em conjunto com modalidade	Caso modalidade seja == 5-Resgatado ou 2-Doméstico, aplicar criptografia no númeroinscriçãoempregador do registro específico	peppered_hash
Último salário	sim, com cnpjraiz ou cnpjcei (CAGED ou outras bases)	Generalização, estabelecendo faixa salarial	stratification_numeric
Cód subclasse cnae20	sim, com codmunicipioresidencia	Encobrimento de caracteres com máscara 4 dígitos à direita	mask
Cod subclasse cnae20	sim, com codmunicipioresidencia	Encobrimento de caracteres com máscara 4 dígitos à direita	mask
subclassecnae20	sim, com codmunicipioresidencia	Supressão de atributos	column_removal
Cód município demissão	sim, com subclassecnae20 ou codsubclassecnae20 ou códsubclassecnae20	-	
Município colônia	sim, com subclassecnae20 ou codsubclassecnae20 ou códsubclassecnae20	-	
Cód município naturalidade	sim, com subclassecnae20 ou codsubclassecnae20 ou códsubclassecnae20	-	
Município naturalidade	sim, com subclassecnae20 ou codsubclassecnae20 ou códsubclassecnae20	-	
Cód municípiosuspensão	sim, com subclassecnae20 ou codsubclassecnae20 ou códsubclassecnae20	-	
Município suspensão	sim, com subclassecnae20 ou codsubclassecnae20 ou códsubclassecnae20	-	
Município demiss ão	sim, com subclassecnae20 ou codsubclassecnae20 ou códsubclassecnae20	-	

Variável	Identificador indireto	Método de anonimização	Algoritmo
Dataad missão procurador em pweb	sim, em conjunto com códsubclassecnae20 ou cnpj	Generalização, disponibilizando apenas mês/ano	stratification_date
Data admissão req	sim, em conjunto com códsubclassecnae20 ou cnpj	Generalização, disponibilizando apenas mês/ano	stratification_date
Dia demissão	sim, em conjunto com resto dadata e códsubclassecnae20 oucnpj	Supressão de atributos	column_removal
emalgestorempweb	sim, em caso de MEI e email contendo nome	Supressão de atributos	column_removal
Inscriçãoe mprega dor cei cnpj	sim, em caso de MEI	Supressão de atributos	column_removal
Nome fantasia em pregador	sim, caso o nome fantasia seja o nome do empregador	Supressão de atributos	column_removal
Numero comunica do dispensa	sim, cruzando com outrabase interna do MT	Supressão de atributos	column_removal
Numero protocolo	sim, cruzando com outrabase interna do MT	Supressão de atributos	column_removal
Numero requerime nto	sim, cruzando com acordos de pagamentos de outras bases	Assegurando que as outras bases com numerorequerimentonão contenham dados pessoais identificadores, ainda que individualizem o registro	-
Numeros érie ctps	sim	Supressão de atributos	column_removal
Valor ante penúltim o salário	sim, com cnpjraiz ou cnpjcei (CAGED ou outras bases)	Generalização, estabelecendo faixa salarial	stratification_numeric
Valo rantepenúltim os alário cnis	sim, com cnpjraiz ou cnpjcei (CAGED ou outras bases)	Generalização, estabelecendo faixa salarial	stratification_numeric
Penúltimo salário	sim, com cnpjraiz ou cnpjcei (CAGED ou outras bases)	Generalização, estabelecendo faixa salarial	stratification_numeric
Valor penúltimo sal ário	sim, com cnpjraiz ou cnpjcei (CAGED ou outras bases)	Generalização, estabelecendo faixa salarial	stratification_numeric
Valor penúltimo sal ário cnis	sim, com cnpjraiz ou cnpjcei (CAGED ou outras bases)	Generalização, estabelecendo faixa salarial	stratification_numeric
Valor último salário	sim, com cnpjraiz ou cnpjcei (CAGED ou outras bases)	Generalização, estabelecendo faixa salarial	stratification_numeric
Valor último salário cnis	sim, com cnpjraiz ou cnpjcei (CAGED ou outras bases)	Generalização, estabelecendo faixa salarial	stratification_numeric
Data admissão req uerente	sim, com cnpjraiz ou cnpjcei (CAGED ou outras bases)	Generalização, disponibilizando apenas mês/ano	stratification_date
Ocupação cbo pret endida	sim, em conjunto com municipioresidencia ou codmunicipioresidencia	Supressão de atributos	column_removal

6.3 Testes e validações

A fim de avaliar o resultado final das técnicas aplicadas, foi aplicada a técnica de verificação do k-anonimato, já citada no item 5.3.2 do presente relatório.

Algumas bibliotecas que implementam tal técnica foram testadas, como a pycanon[34]. Porém, nenhuma das ferramentas prontas testadas foi suficiente para executarmos os relatórios nos equipamentos disponíveis sem que houvesse escassez de recursos (memória RAM). Desta forma, tendo em vista o alcance limitado desta POC, optou-se por fazer uma implementação própria e simplificada do algoritmo k-anonimato também disponível no GitlLab do projeto.

Vale destacar que a Google possui um serviço[35] que permite a execução desta avaliação (e outras). Entretanto, este tipo de serviço não foi utilizado por conta dos termos contratuais firmados no âmbito do desenvolvimento desta POC: seja por não termos autorização para subir as bases de dados em serviços de terceiros, seja porquestões de custos extras associados.

O resultado da avaliação pelo k-anonimato pode indicar necessidades de modificações nas técnicas aplicadas. Como exemplos de modificações podem-se citar: alterações nas faixas de estratificação de algumas variáveis; remoção de “registros únicos” de acordo com determinados filtros/agrupamentos ou ainda a remoção de algumas variáveis. Tudo isso precisa necessariamente ser avaliado levando-se em consideração os resultados do processo de anonimização numa determinada base (conjunto de colunas/variáveis/atributos e linhas/registros). Ou seja, não se pode transpor o resultado da POC com uma amostra para a base completa sem uma reavaliação do resultado da aplicação dos métodos indicados.

No contexto desta POC, para apresentar o processo de avaliação de resultados da aplicação das técnicas indicadas, usou-se a implementação própria do algoritmo dek-anonimato na base do CAGED. Isso se deu em 2 etapas: (1) na amostra original dabase de dados e (2) após a aplicação das técnicas indicadas. A tabela a seguir traz uma síntese dos resultados obtidos. Cada linha da tabela contém uma aplicação da técnica K-Anonimato. A primeira coluna contém qual coluna/variável/atributo ou conjunto de colunas/variáveis/atributos foram usadas no teste. A segunda coluna elenca quantos itens na base original não passaram no teste. A terceira coluna indica quantos itens na base anonimizada não passaram no teste. A quarta e última coluna apresenta a melhora percentual antes e depois do processo de anonimizaçãodesta POC.

[34] https://pycanon.readthedocs.io/
[35] https://cloud.google.com/dlp/docs/compute-k-anonymity#console

Quasi-identificadores testados (unicamente ou em combinação)	Itens base bruta	Itens base anonimizada	Melhoria
município	0	0	-
subclasse	40	3	92,50%
cbo2002ocupação	168	15	91,07%
graudeinstrução	0	0	-
sexo	0	0	-
Idade	7	0	100,00%
tipodedeficiência	0	0	-
tipoempregador	0	0	-
tipoestabelecimento	0	0	-
indicadoraprendiz	0	0	-
salário	14003	0	100,00%
valorsaláriofixo	13560	0	100,00%
paísdenacionalidade	17	3	82,35%
paísdeorigem	14	0	100,00%
graudeinstrução, sexo, idade, tipodedeficiência	591	62	89,51%
município, subclasse, cbo2002ocupação, graudeinstrução, sexo, idade, tipodedeficiência, tipoempregador, tipoestabelecimento, indicadoraprendiz, salário, valorsaláriofixo, paísdenacionalidade, paísdeorigem	196563	55766	71,63%

Destaca-se aqui a grande melhora apresentada em todas as análises. Onde a melhoria indicada foi de 100%, para a amostra entregue, as técnicas de anonimização desta POC foram suficientes. Onde a melhoria indicada foi inferior a 100%, cabe reavaliação. Essa reavaliação pode significar manter ou não a variável como quasi-identificadora ou não, ou ainda aplicar técnica adicional de anonimização.

Por exemplo, a variável ‘subclasse’ (correspondente à subclasse CNAE) contém 7 dígitos, sendo que os 3 primeiros correspondem ao GRUPO. Mascarar os 4 últimos dígitos é equivalente a trocar SUBCLASSE por GRUPO, generalizando. Feita essa generalização, esta variável deixa de ser um quasi-identificador. Situação semelhante ocorre com ‘cbo2002ocupação’, que contém 6 dígitos, sendo os 4 primeiros correspondentes à FAMILIA. Ao mascarar os 2 últimos dígitos, entrega-se a informação generalizada da família e a variável deixa de ser uma uma quasi-identificadora. E ao

reclassificar as variáveis, muda-se também o grupo a ser avaliado pelo algoritmo de k-anonimato.

No caso da variável ‘paísdenacionalidade’ é provável que, com uma base de dados maior, as 3 ocorrências de registro único sumam. Mas seja para esta POC ou para a necessidade de fazer um outro extrato da base completa, é possível empregar o algoritmo de remove_uniques sobre estes registros.

Na penúltima linha da tabela encontramos uma combinação de indicadores que inicialmente parecia muito preocupante, e para a qual foi obtida uma melhora de 89%. Aqui os critérios de generalização poderiam ser reavaliados. Por exemplo: ao transformar de ‘idade’ para a ‘faixa_etaria’ os intervalos podem ser alterados, em especial no que tange aos valores mais extremos, como os que correspondem às pessoas muito idosas.

Já na última linha, contendo uma quantidade muito maior de variáveis, a melhora foi de 71%, reduzindo a possibilidade de identificação de mais de 140 mil registros (dentre os mais de 326 mil registros da base original). Neste caso, o indicado é avaliar os critérios anteriormente mencionados (reclassificação de variáveis quasi-identificadoras, adoção de tamanho de base de dados mais realista, alterações dos critérios de generalização se necessário, adoção iterativa do remove_uniques se preciso) e rodar novamente o k-anonimato. A melhora do grau deanonimidade certamente aumentará. Adicionalmente podem ser empregadas novamente técnicas aqui desenvolvidas ou ainda outras listadas no item 5 - indicando que trata-se de um processo iterativo, em que a quantidade de iterações depende do grau de risco tolerado.

É importante frisar a classificação do que é “quasi-indentificador” (indiretos), dos intervalos das faixas usadas para generalização e o tamanho mínimo de uma amostra de dados a ser tornada pública são **questões arbitradas pela organização que detém e gerencia tais dados**.

6.4 Principais Avanços

O principal avanço advindo desta POC é a demonstração que é possível abrir e distribuir bases com dados considerados pessoais e sensíveis, desde que devidamente tratados, devido à sua importância para a sociedade brasileira. Apesar dos algoritmos da POC terem sido executados numa amostra pequena das bases, não foi necessária nenhuma tecnologia extremamente avançada para promover anonimização ou pseudonimização dos dados inicialmente categorizados como identificadores (diretos ou indiretos). Isto é

uma evidência da factibilidade de pôr este código em produção, ainda que sejam necessárias algumas adaptações.

Um outro avanço deste processo foi a criação de um algoritmo para anonimização de CPFs baseado em técnicas normalmente usadas para guarda de senhas em sistemas computacionais, aplicando algoritmos tradicionais de criptografia numa combinação do CPF e de uma sequência de caracteres aleatória. Este processo traz três características que não estão presentes nos métodos atual e comumente usados pelos governos no Brasil:

1. **Unicidade dos CPFs anonimizados nas bases**: O processo garante que para cada CPF, o resultado anonimizado é também único, o que impede ambiguidades no registro.

2. **Controle do cruzamento dos dados com outras bases**: O processo proposto permite ao órgão decidir se permite ou não o cruzamento de diferentes basesa partir de um CPF que passou por hash com pepper. Caso se queira permitiro cruzamento entre bases, basta usar a mesma tabela de pepper, caso se queira impedir, basta usar diferentes tabelas de pepper.

3. **Resistência a ataques de desanonimização**: Um ataque para desanonimizar os dados sem acesso ao CPF original são computacionalmente muito caros, diferente de ataques contra os métodos atualmente usados.

6.5 Recomendações

A seguir, ficam registradas as 5 principais recomendações que advêm tanto de lições aprendidas no processo, como do reconhecimento das limitações inerentes auma prova de conceito.

- **Necessidade de elaborar Dicionário de Dados bem definidos e atualizados:** Foram recebidos layouts das bases CAGED, RAIS - estabelecimentos e RAIS -

vínculos, mas não das bases BEM e SD. Houve dúvida inicial em relação a 119 variáveis (19% do total), o que sinaliza o quão importante é a documentação das bases para que seja possível um tratamento de dados adequado. Tais dúvidas foram pormenorizadas em tabelas e foram esclarecidas em reuniões online e por correspondência eletrônica. Isso, entretanto, não invalida a necessidade de um esforço interno do Ministério em melhorar sua documentação e prever um dicionário de dados para cada base produzida. Se num primeiro momento isso demanda esforço concentrado, no futuro, economiza recursos internos e externos.

- **Necessidade de ajustes dos algoritmos para um ambiente de produção:** A POC implementada não foi desenvolvida para um cenário de produção seja por conta do acesso a um universo reduzido de dados (amostra) ou pelo não acesso à infraestrutura de dados do órgão responsável pela coleta, tratamento e armazenamento das bases. Portanto, ao internalizar e escalar apresente POC é necessário considerar a utilização de recursos mais elaborados como sistemas gerenciadores de bancos de dados[36], caches em memória[37], sistemas de processamento distribuídos como o Apache Spark[38], ou mesmo técnicas de programação paralela/concorrente[39] e multiprocessamento. Esses recursos precisam ser avaliados tendo em vista:o ambiente de produção disponível, as tecnologias de ETL[40] já empregadas e as possibilidades e necessidades de CI/CD[41] do órgão.

[36] https://pt.wikipedia.org/wiki/Sistema_de_gerenciamento_de_banco_de_dados
[37] https://kinsta.com/blog/in-memory-database/
[38] https://spark.apache.org/
[39] Computação paralela é uma forma de computação em que vários cálculos são realizados ao mesmo tempo, operando sob oprincípio de que grandes problemas geralmente podem ser divididos em problemas menores, que então são resolvidos concorrentemente.
[40] ETL é o acrônimo para extract, transform, load (extração, transformação, carregamento) que são três etapas usadas para combinar dados de diversas fontes de dados.
[41] CI/CD é o acrônimo para Continuous Integration/Continuous Delivery (integração e entrega contínuas) e trata de uma prática de desenvolvimento de software que visa tornar a integração de código mais eficiente por meio de builds e testes automatizados.

- **Transparência em relação aos processos de anonimização e**

pseudoanonimização adotados quando da publicação das bases: É importante, quando adotados os algoritmos em ambiente de produção, o registro e a transparência acerca de quantos registros foram removidos e o respectivo motivo. Isso é importante para não perder rigor no controle estatístico por parte de quem consumirá essas bases (não perder a noção dotamanho do universo).

- **Necessidade de contar com definições legais/jurídicas com relação a alguns dados** (e.g.: MEI): O tratamento técnico conferido às MEIs (microempreendedores individuais) na POC já foi explicitado anteriormente. Entretanto, foi sentida uma lacuna de definição jurídica com diretrizes do Ministério do Trabalho e Emprego de como tratar dados de MEI. Afinal, trata-se de uma pessoa física que pode ter ou não nome/CPF na razão social (porque nunca o foi ou porque mascaramos), mas que pode ter endereço jurídico ou telefone comercial coincidente com residência ou telefone pessoal - como sabemos muitas vezes ser o caso.

- **Necessidade de consideração de aspectos jurídicos no processo de internalização da POC**: A POC se ateve a conceitos atuais de criptografia, anonimização e identificação pessoal, bem como a legislação geral vigente. Entretanto, aspectos jurídicos de interesse específico, assim como normas internas, devem ser considerados no processo de internalização da POC pelo Ministério do Trabalho e Emprego. Por exemplo, recomenda-se ao menos observar o modelo analítico de entropia da informação apresentado na Figura 04 que agrupa logicamente os sete critérios normativos prescritos pela LGPD e lista uma série de fatores que ajudam na identificação do quão tolerável (razoável)são os riscos de reversão das técnicas de anonimização aplicadas.

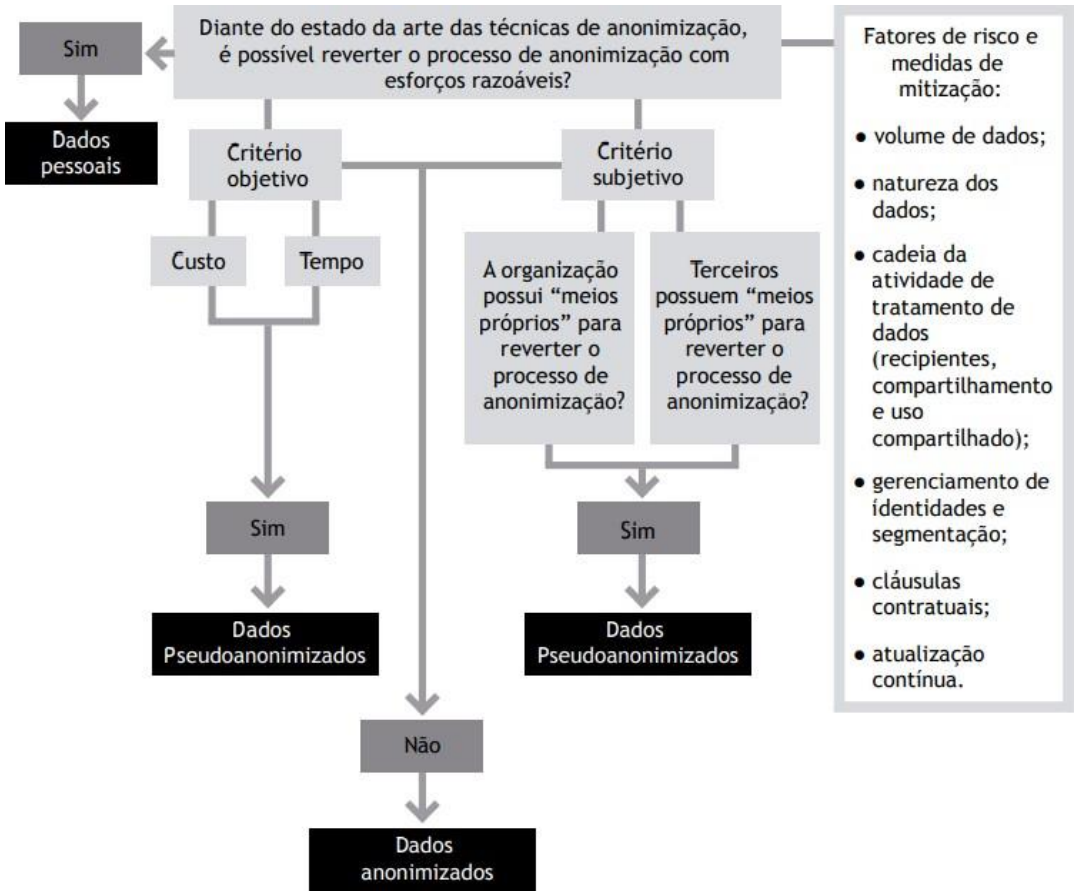


Figura 04 - Entropia da Informação (Bioni, 2020[42])

7 Considerações Finais

Considerando que o presente trabalho foi desenvolvido dentro de parâmetros de uma Prova de Conceito, por meio da qual valida-se ou não uma hipótese inicial, o resultado esperado foi hesitoso. Nesse sentido, a pergunta **"É possível anonimizar as cinco (05) bases apresentadas, garantindo a necessária publicização destas e, ao mesmo tempo, a privacidade de cidadãs e cidadãos?"** foi respondida positivamente pelo processo.

Além de validar a hipótese, o processo de desenvolvimento desta Prova de Conceito, também promove ganhos secundários ao oferecer um diagnóstico abrangente da proposta, que envolve desde orientações para a estruturação das bases (Dicionário de dados) até recomendações

[42] Bioni, B. R. Compreendendo o conceito de anonimização e dado anonimizado. Cadernos Jurídicos, São Paulo, v. 19, n. 53, p. 191-202, jan./mar. 2020. Bimestral. Disponível em:

https://www.tjsp.jus.br/download/EPM/Publicacoes/CadernosJuridicos/ii_9_anonimiza%C3%A7%C3%A3o_e_dado.pdf?d=637250349860810398

presentes na documentação sobre como aprimorar ouso das técnicas de anonimização (resultados dos testes), garantindo a máxima segurança.

Por fim, destaca-se a relevância de, em um próximo ciclo deste trabalho, incorporar a competência jurídica especializada em privacidade de dados públicos, de forma ampla. Tanto no que tange decisões a serem tomadas na avaliação de disponibilização de dados específicos (como é o caso das MEI's), como também noque tange a própria natureza do trabalho, uma vez que determinadas decisões devem levar em consideração o interesse público em intersecção com a proteção dos dados pessoais - uma dinâmica que ultrapassa a técnica e incorpora decisões políticas agregadas em casos análogos já pesquisados e consolidados ao redor do mundo.

Anexo I - Relatório da Oficina com especialistas: Desafios da anonimização de grandes bases de dados.

20/06/2022

Abertura

14h00 - 14h15 - Patrícia Pelatieri - Diretora Técnica Adjunta do DIEESE

Demandas para disponibilização das bases de dados do Ministério do Trabalho e Previdência

14h15 - 14h45 - Felipe Vella Pateo (Coordenador-Geral de Cadastros, Identificação Profissional e Estudos do Ministério do Trabalho e Previdência)

Roda de Conversa: Desafios e Caminhos possíveis para anonimização de grandes bases

14h45 - 15h15 - Augusto Fadel (IBGE) - Uma visão geral de técnicas de preservação de privacidade aplicadas ao compartilhamento seguro de microdados
15h15 - 15h45 – Erivelton Pires Guedes – (IPEA) A experiência de anonimização da RAIS
16h00 - 16h30 – Thais Paiva (UFMG) - A experiência da UFMG com anominização de bases de dados

Debate

16h30 - 17h30

Encerramento

17h30 - 18h

Participantes:

Patrícia, Márcia, Celi, Eloá, Geovana, Jeniffer, Laís, Tatiana, Thais Paiva,Thais Cruz, Fernando, Amilton, Augusto Albuquerque, Augusto Fadel, César, Clóvis, Daniel, Diego, Felipe, Matheus, Raigner, Rodrigo, Sérgio, Welton, Erivelton

Abertura: Patrícia Pelatieri – DIEESE

O objetivo da oficina é a troca de ideias, troca de experiência para subsidiar a escolha de caminhos possíveis para anonimização e integração de grandes bases de dados.

Quero começar agradecendo todos e todas que se disponibilizaram a estar aqui, em aceitar o desafio dessa conversa, em especial o Augusto Fadel, o Erivelton e a Thais Paiva, que toparam trazer aqui as suas experiências nesse assunto e também quero agradecer ao Edgard do Dieese que fez a primeira pesquisa e contato com os especialistas no assunto. As agendas são bem complicadas, então foi um esforço grande juntar e realizar essa nossa primeira conversa. Esse trabalho, está inserido em uma parceria que o DIEESE tem com o Ministério do Trabalho, uma parceria de longa data. No início da parceria, muitos anos atrás, o Dieese ajudou no aprimoramento dos registros administrativos do Ministério, e ao longo desse tempo, uma demanda que é muito constante, e se intensificou nos últimos anos, são as solicitações dos microdados dessas grandes bases de registros administrativos, pelas mais diversas entidades. Em seguida a abertura o Felipe, do Ministério do Trabalho vai explicar melhor o que é essa demanda e o desejo de disponibilizar e as dificuldades do Ministério em disponibilizá-las, uma vez que elas identificam o indivíduo. Então, é sempre um dilema. O desafio que se coloca é: como desidentificar as bases de tal modo a poder ofertar para a sociedade um volume de informações para além daquelas que o Ministério já disponibiliza para estudos, para políticas públicas, enfim, para a sociedade de uma maneira geral.

Demandas para disponibilização das bases de dados do Ministério do Trabalho e Previdência – Felipe Ministério do Trabalho

Obrigado, Patrícia. Acho que, em primeiro lugar, agradecer ao DIEESE por ter topado esse desafio. Nessa parceria que a gente vem construindo e repetindo sempre a partir do esforço do DIEESE na captação de emendas e a gente na operacionalização e construção de resultados para elaboração de estudos e estatísticas de trabalho. E foi interessante o Dieese, mesmo sem termos falado, chegar ao Augusto Fadel, que já ajudou a gente nesse tema há um tempo atrás. É uma satisfação vê-lo novamente. O Augusto está sempre disponível para ajudar a gente nessa discussão que é muito difícil. O Ministério tem uma equipe técnica limitada, comparado com institutos de estatística propriamente ditos, para o tamanho do desafio, e por isso que é importante a gente contar sempre com as parcerias, porque a gente sabe que sozinho a gente não dá conta do desafio. O desafio grande para gente foi colocado a partir da publicação da Lei Geral de Proteção de Dados, ela gera uma insegurança jurídica grande dentro do Ministério. E também acho que vem daí o interesse do pessoal do

Ministério da Cidadania, que a gente está superfeliz de começar essa conversa junto com a ponte do Amilton, porque acho que lá no fim desse trabalho a gente pode até construir uma integração das bases. Acho que seria genial ter uma integração da base de mercado de trabalho, RAIS, CAGED com o Cadastro Único e acompanhar essa trajetória em diversos aspectos, para além do trabalho formal que a gente tem aqui.

Quais as bases de dados que existem hoje no âmbito da Secretaria, pelo menos as principais com as quais a gente trabalha aqui? A RAIS e o CAGED, que são as bases que têm controle estatístico da movimentação do mercado de trabalho, todo mundo conhece. E a gente tem também algumas bases de gestão de políticas públicas, a base de gestão da política do seguro-desemprego, a base de gestão do benefício emergencial. Essas quatro primeiras aqui, elas já estão contempladas em algumas normativas internas do Ministério, porque elas já eram de responsabilidade da Secretaria de Trabalho antes da volta da reintegração do Ministério do Trabalho. A gente tem uma base de gestão do abono que é um processo em construção ainda, porque não existia essa governança dentro do Ministério, já que o pagamento era operacionalizado diretamente pelos bancos. E agora a gente está puxando isso para dentro do Ministério, o abono salarial. Tem a base de gestão já mais antiga da intermediação de mão de obra, do Proger. Do Proger ainda tem alguma peculiaridade ali, porque envolve o sigilo fiscal também, mas são bases hoje de posse da Secretaria de Trabalho. E o nosso objetivo é começar o trabalho com aquelas bases mais demandadas, que é RAIS e CAGED, mas que a gente possa paulatinamente integrando outras bases que permitam mais pesquisadores e usuários fazerem as suas pesquisas sem necessariamente passar por um processo de acesso a dados identificados.

Quais são as formas de disponibilização dos dados hoje? Como a Patrícia falou, a gente já disponibiliza diversos dados com acesso livre. A gente tem dados em painéis estruturados, painéis de consulta online, que tentam ser mais intuitivos para consulta por repórteres, outros usuários, a gente já tem RAIS, CAGED, seguro de desemprego, alguns dados da intermediação de mão de obra que a gente já incorporou no painel da classificação brasileira de ocupações, então esses são dados consolidados, que o Ministério divulga, mas que também tem alguma preocupação com a anonimização, à medida que esses dados permitem filtro por municípios, municípios pequenos, aí já apresenta um risco de identificação de trabalhadores, por exemplo, nesses painéis. A gente tenta colocar alguns controles para não apresentar dados de remuneração para municípios com pouca quantidade de trabalhadores. Então passa por aí uma preocupação também com a não identificação do trabalhador. Além

disso, o Ministério disponibiliza também, que já vem de muito tempo, microdados não identificados. São microdados públicos, da RAIS e do CAGED.

Então a RAIS é uma base anual, com alguns milhões, 60 milhões de linhas, mais ou menos. Cada linha é um vínculo. E a gente disponibiliza isso em formato TXT, retirando variáveis que possibilitem identificação do trabalhador, como CPF e nome. Já o CAGED é uma base mensal, onde tem só as movimentações mensais do trabalhador, dos diferentes trabalhadores, e a gente faz o mesmo processo, seria o mais simples, que é só tirar as variáveis de identificação e aí isso fica público já no site do Ministério, nesse formato - pdet.mte.gov.br, faz parte do que chamamos no Ministério de Programa de Disseminação das Estatísticas de Trabalho. Além disso, temos um nível intermediário de dado que não chega a ser identificado por pessoa, por trabalhador, mas é um microdado com identificação de CNPJ, que costumamos disponibilizar quando chegam processos fundamentados, de usuários que justifiquem a sua utilização, com assinatura de termos de compromisso de confidencialidade. Apesar de não ser um dado pessoal direto, uma vez que é só o dado do estabelecimento da empresa, a gente impõe um nível já um pouquinho maior de restrição também pensando que ele permite mais facilmente um processo de desidentificação. Prioritariamente só atendemos as demandas com dados mais consolidados, a base mesmo de microdados com identificação de CNPJ do estabelecimento. E, por fim, a gente tem os microdados identificados propriamente ditos. São de acesso restrito, contém as informações pessoais. E a gente tem dois tipos de procedimentos estabelecidos para esse acesso a esses microdados identificados. Tem um procedimento simplificado que é baseado no decreto de compartilhamento de dados, que é esse decreto 1046 de 2019. Então, para usuários do governo federal, o processo é mais simples e aí inclui também universidades federais, que entendemos que são abrangidas pelo decreto, órgãos de controle, outros ministérios finalísticos que requeiram um dado para a execução das suas políticas públicas. O procedimento consiste em só um ofício com solicitação e a gente se baseando no decreto, costuma disponibilizar esses dados de forma um pouco mais imediata. Temos um outro processo mais complexo, que foi estabelecido nessa portaria do Ministério, a portaria 671 de 2021, mas que ainda é um processo que passa por muita insegurança jurídica. A gente começou a discutir essa portaria ainda no final de 2019, começo de 2020, a partir da LGPD, a portaria estabeleceu uma série de critérios, foi publicada, e depois de ser publicada houve mais um questionamento da consultoria jurídica do Ministério, então ficou um tempo mesmo com a portaria vigente, não estávamos disponibilizando o acesso a esse tipo de requisitante. Fizemos uma revisão da portaria, que foi esse ano ainda, e conseguimos disponibilizar alguns acessos para algumas

solicitações, só que sempre que acaba tendo alguma mudança de gestão, volta a insegurança, a gente rediscute com a consultoria jurídica e o processo acaba sendo bem vagaroso.

Por isso a necessidade de buscar outras soluções, prioritariamente no espírito da LGPD, construir soluções que não envolvam a transferência de dados identificados. E quais são as justificativas? Já que a gente disponibiliza aquelas bases sem identificação, por que usuários pedem ainda acesso a dados identificados? Tem diversos tipos de órgãos de controle que apontam a necessidade dos dados para fazer seus controles, seus mecanismos de finalização. Eu chamo de diversos tipos, porque são não só TCU, CGU, mas às vezes aparece uma controladoria municipal de um município, uma controladoria estadual, às vezes a polícia militar de um determinado estado pede acesso a esse dado para a gente. E tudo que é no nível estadual e municipal não é abrangido pelo decreto de compartilhamento, então é submetido às restrições da LGPD.

A necessidade de avaliação de resultados de ações ou políticas públicas, que muitas vezes só com a base de dados não identificados, que a gente disponibiliza hoje, não é suficiente para fazer esse tipo de avaliação, que por vezes envolve cruzamento com listas de beneficiários de uma política específica, seja de qualificação, microcrédito, diversas outras possibilidades, necessidade de cruzamento com outras bases. A gente tem muitos pesquisadores que, por exemplo, falam que precisam do dado identificado, porque eles querem fazer, ou já estão fazendo cruzamentos com o cadastro único. E a gente tem estudos que simplesmente precisam acompanhar a trajetória dos trabalhadores ao longo dos anos. Então, como a RAIS é anonimizada, ela é uma base anual, se você precisa saber o que aconteceu com o trabalhador ao longo de um período maior de anos, não é possível fazer sem os dados identificados.

Então, nosso objetivo geral frente é esse tipo de justificativa para solicitação de dados é diminuir a disponibilização de dados identificados, mas para isso é preciso a criação de bases anonimizadas, vinculadas, seja vinculada entre as diversas bases do Ministério, seja vinculada longitudinalmente, e quem sabe num dia a gente conseguir vincular com outras bases. Só um pouco mais de detalhe de quais são as modalidades de procedimento para acesso a dados pessoais identificados hoje:, tem um processo simplificado pelo decreto de compartilhamento para órgãos federais. Tem processos para entidades públicas fora do âmbito federal, que são um pouco mais simples também. E quando a solicitação é feita por uma organização da sociedade civil, que além da portaria 671, com procedimentos específicos, temos ainda que se referendar no MROSC, no Marco Regulatório das

Organizações da Sociedade Civil, porque é uma parceria entre o governo e uma organização da sociedade civil. Então, envolve mais uma série de critérios de comprovante de adimplência, de detalhamentos do plano de trabalho, de cumprimento de tudo, como se fosse, tem todos os mesmos critérios, o repasse de um dinheiro para a entidade da sociedade civil, só que no caso é um repasse de informações pessoais identificadas. Então, um processo, quando chega para a gente, ele cumpre todas essas etapas, tem alguns procedimentos para envio da solicitação, uma série de documentos comprobatórios, é feito uma minuta para estabelecer uma nota técnica justificando o processo, uma minuta de um acordo de cooperação técnica para ser celebrada entre o dirigente do Ministério, que hoje está delegado pelo secretário-executivo do Ministério, e tem que ser assinado pelo dirigente da entidade solicitante. Depois disso, a gente vai ter um mecanismo de disponibilizar periodicamente os dados por um arquivo que é preparado pela equipe, disponibilizado com uma extração customizada, dependendo do seu usuário é de um município ou de uma categoria, é um sindicato, às vezes, de uma categoria específica, a gente vai passar só os dados do município ou da categoria específica a qual o usuário está se referindo, dependendo aí, variando com o tipo de demanda.

Isso dá a vocês uma ideia da dimensão da situação, fora os processos de acordo de compartilhamento que acontecem um pouco mais fluidamente, para o governo federal, a gente tem hoje 43 organizações fora do governo com acordos, a maioria delas são acordos antigos, estabelecidos em 2019, antes da mudança decorrente da LGPD, que recebem dados identificados, muito com o acordo vencendo. E temos um estoque de solicitação de dados de 146 processos abertos de instituições variadas. Então, dentro dessas categorias, a gente tem sindicatos, entidades privadas, universidades privadas também, algumas delas grandes, FGV, PUC, etc. A gente tem universidades estaduais, órgãos públicos, municipais, estaduais, a gente tem conselhos profissionais também, que é uma demanda bastante significativa, organismos internacionais, BID, OIT, Banco Mundial, universidades internacionais, então alguns desses que eu mencionei tem processos abertos que estão vencendo, alguns estão com processos esperanto, totaliza assim 146 processos, que passam por todas aquelas etapas.

Por isso precisamos construir soluções mais estruturantes, mais seguras juridicamente. Garantir essa segurança jurídica, essa adequação a LGPD, mas sempre com a preocupação de não prejudicar a realização de pesquisas, a produção de conhecimento sobre o mercado de trabalho, as políticas públicas de trabalho. Existe uma cultura de acesso aos dados identificados do CAGED e da RAIS, principalmente, que já vem de muitos anos. Nosso objetivo é não ocasionar uma descontinuidade, não causar um problema de, ter toda essa

comunidade de usuários que está acostumado a acessar dados identificados e, de repente, todo mundo ficar no escuro. Queremos construir uma transição, uma solução que diminua, mas que tenha outras alternativas. Então, acho que o desafio aqui é construir bases vinculadas e desidentificadas. Tem o caminho da construção de salas de sigilo, tem outros órgãos que já têm nossos dados em salas de sigilo, que é um caminho. Tiveram a importância de criar uma sala de sigilo própria. Tem esses outros caminhos de construção de APIs que permitam realização de consultas online, com a especificação do usuário colocar sem acessar dados identificados. Então, são caminhos possíveis que precisamos discutir e esperamos poder com esse apoio avançar.

Roda de Conversa: Desafios e Caminhos possíveis para anonimização de grandes bases

Augusto Fadel - IBGE

Augusto Fadel: Bom, antes de mais nada, obrigado, Patrícia, pelo convite, Edgar, também. O tema não é fácil, de profunda importância, é um tema complicado em vários sentidos, a gente vai falar bastante sobre isso aqui hoje, mas acho que o IBGE também talvez tenha um pouco mais de experiência com o usuário geral, com o usuário externo, por conta da natureza da instituição, de realmente ter esse aspecto de Instituto Nacional de Estatística e acaba sendo procurado nesse tipo de situação. Muita coisa mudou recentemente em termos de legislação e tecnologia. E é muito difícil acompanhar esse progresso, como instituição só. Eu acho que se a gente reunir esforços, vamos conseguir caminhar muito melhor. Então, obrigado pela oportunidade, vai ser ótimo ouvir de vocês aqui também. Vou compartilhar um pouquinho do que a gente tem visto sobre isso no assunto, que assim, é superficial ainda, o assunto é muito profundo, mas essa troca é absolutamente essencial, assim, muito bem-vindo.

O que eu trouxe, na verdade, é compartilhar algumas das técnicas que estudamos, testamos, e temos avançado um pouco no assunto. Mas antes de falar disso, é essencial o contexto: Por que compartilhar microdados? É muito mais fácil a gente simplesmente colocar num servidor seguro, cheio de proteção e trabalhar para que ninguém tenha acesso, já que ele contém informações sensíveis. Não é bem assim. Então, puxando um pouco para a questão da estatística pública, não é? Pensando nos princípios fundamentais das estatísticas públicas da ONU tem o primeiro princípio, que é o de relevância em parcialidade e igualdade de acesso, que ressalta a importância de um sistema de informação numa sociedade democrática, tanto para o governo quanto para o cidadão, etc. Então tem uma consideração

importante por trás. Enquanto sociedade nunca tivemos um potencial tão grande de capitar, armazenar e processar dados. Então eu acho que essa demanda, ela vem muito daí também, a gente hoje pode fazer isso. Há bem pouco tempo atrás, uma base de 60 milhões de linhas, você precisava de um equipamento muito específico, poucas instituições tinham acesso, , talvez há 20 anos, isso ainda fosse realidade. Não é mais realidade, 60 milhões de linhas, qualquer computador pessoal, dependendo do que você vai fazer, já é capaz de processar. Então, essa questão tem toda uma relação de compartilhamento de microdados com a de desenvolvimento e pesquisa.

Então, tem a questão das fontes de dados. Todos nós como produtores de estatística pública, de dado público, nós também somos potenciais consumidores. Há um aumento expressivo de não resposta às pesquisas, da dificuldade de coleta de informação, aspectos relacionados à eficiência, à tempestividade, principalmente falando de dado econômico. A gente leva um ano, um ano e meio para conseguir divulgar um dado econômico, em termos históricos ok, mas em termos de análise conjuntural é um problema. A qualidade que precisa ser preservada, se a gente conseguir trazer mais, ter um envolvimento melhor com o setor privado, informações como telefonia móvel, por exemplo, e todo o impacto que isso tem para a sociedade, pode ajudar na tempestividade das pesquisas. Enfim, tudo isso precisa de uma coordenação nacional.

Tem essa iniciativa do ConnectaGov, que é muito interessante, é uma coleção de APIs que você consegue reunir para atualizar bases, enfim. Mas é uma iniciativa, todas essas iniciativas têm um foco muito grande na administração pública, não necessariamente na produção de estatística, e menos ainda no compartilhamento de microdados com um público fora da esfera governamental. E chega finalmente na questão da confidencialidade. Então, temos um princípio fundamental também sobre isso. Os dados coletados para fins estatísticos devem ser usados exclusivamente para esse fim. A LGPD, que é uma novidade um pouco mais recente, em pleno vigor desde agosto do ano passado. A PEC, que aprovou a proteção de dados pessoais, inclusive digitais, como direito fundamental na construção brasileira no início do ano.

Tem esse caso de 2012 que ficou bastante famoso do mais gigante do Varejo nos Estados Unidos que analisava o perfil de compra dos usuários para identificar alguns grupos, alguns nichos, dentre eles mulheres grávidas e fazendo uma campanha bastante intensiva de marketing direcionado, mandou uma correspondência para a casa de uma adolescente que estava grávida, a família não sabia, descobriu assim. Isso é a história que veio a público, foi

capa da New York Times Magazine em 2012. Porque eu estou trazendo esse assunto de dez anos atrás, porque esse assunto voltou a pauta com muita força, com o retrocesso na região de direito ao aborto nos Estados Unidos, vários dos gigantes do comércio voltaram, ou aliás nunca deixaram de se interessar por esses perfis de consumidores, mas voltaram particularmente ao perfil das gestantes, voltou a dedicação ao perfil das gestantes. Só que mais do que isso, tem grupos, enfim, antiaborto e tal, que estão usando esses dados ou trabalhando esses dados para monitorar o comportamento das mulheres com relação a esse tema. A questão é que relações confiáveis, relações não confiáveis, elas sempre vão existir. Então, esperar estabelecer uma relação confiável para que se possa compartilhar dados, ficaremos na situação onde estamos hoje, que quase ninguém consegue compartilhar dado com ninguém. Isso é seguro, o seu dado é seu, mas é ruim para quem precisa do dado. O Ministério de Trabalho é realmente uma instituição com bastante relevância. O cadastro de empresa do IBGE, de onde são retiradas, selecionadas as pesquisas econômicas das pesquisas por empresa do IBGE, é atualizado pela RAIS. Então, se em algum momento se estabelecer que esse vínculo não é mais confiável, vai ser um problema bastante grave em termos do levantamento da situação econômica no Brasil por parte do IBGE. A questão é que a legislação não impede o compartilhamento. Na verdade, ela estimula o investimento no método de seguros, que é bom para todo mundo. É bom para o proprietário do dado, é bom para quem está representado no dado, é bom para quem vai usar o dado.

Sobre as técnicas. A primeira coisa que você pensa é a anonimização de base de dados. Quando eu falo anonimização é simplesmente você remover o identificador dos indivíduos que estão representados. E tem um caso que ficou bastante famoso também de uma competição da Netflix para mostrar que esse problema dá rasteira em gente bem grande também. Uma gigante da tecnologia que não pensou direito no assunto, publicou em 2006 um conjunto de dados com formação de 480 mil usuários anonimizados. Eles eram identificados simplesmente por um ID sequencial. E esses usuários tinham dado avaliação a 18 mil filmes diferentes. Ao todo, havia 100 mil avaliações com a data em que essas avaliações foram feitas, desses usuários. A questão é que 480 mil usuários com 18 mil filmes dariam alguma coisa na casa de bilhão. E tinham 100 milhões de avaliações nessa base. Isso significa que a combinação de filmes para cada usuário era praticamente uma impressão digital, ela era única. E um grupo acessando a base, aliás, acessando o site, não foi uma base, eles fizeram uns créditos do site, do IMDB, da internet no Movie Data, que é um site que contém informações de filmes e avaliação de usuários, eles conseguiram reidentificar muitos usuários, principalmente as personagens públicas, que são um pouco mais fáceis de você

reconhecer. Enfim, essa competição que a Netflix promoveu, a primeira edição se encerrou em 2009, eles planejaram uma segunda edição em 2010 que foi cancelada por conta dessa questão. Esse caso, como é que você vai imaginar que, não é trivial reconhecer que existia uma brecha ali, existia uma brecha enorme naquele conjunto.

O que eu queria propor em termos de abordagem, é um pouco difícil classificar esse assunto, está se modificando o tempo inteiro, entrando muita informação nova, é difícil conseguir organizar em termos de conhecimento o que você coloca onde, mas uma divisão que eu tenho gostado bastante é uma que separa os métodos entre Output e Input Privacy.

Output Privacy, que é uma coisa muito mais próxima para a gente. A gente já faz em parte, muito mais próxima também de implementar. E esses métodos, em geral, eles produzem um arquivo de dados autos contidos daquele que ou você vai compartilhar com um usuário a partir de um termo de confidencialidade, responsabilidade, ou mesmo tornando público no website.

Há uma diferenciação entre esse sigilo de registro de atributo. E esses métodos, uma classificação que também é comum, acho que é questionável como as demais, mas é separar esses métodos entre não perturbativos, perturbativos e geração de microdados sintéticos. Os não perturbativos são aqueles em que você altera o dado, por exemplo, fazendo uma agregação de idade em faixas etárias, por exemplo, esse seria um método não perturbativo que não induz ruído. O perturbativo é um método que induz ruído. Ele altera a informação induzindo algum tipo de erro controlado, claro, naquela informação. E os microdados sintéticos são produzidos a partir de uma modelagem, gerado um novo conjunto de dados que preserva as informações estatísticas, mas o conjunto de dados que é disponibilizado não tem nenhuma informação do conjunto de dado original.

O que a gente tenta decidir nesse caso é o quanto está disposto, a comprometer a confidencialidade para maximizar a utilidade ou o inverso? O que eu posso comprometer da utilidade para maximizar a confidencialidade? Com esses objetivos, claro, eles são divergentes. Você tem que fazer um *straight-off* de que quanto mais por um lado menos para o outro. Bom, então, para ilustrar o que seriam esses métodos: tem esse conjunto de dados hipotéticos com um atributo identificador, que seria o CPF, por exemplo. Três atributos que não contêm nenhuma informação sensível, que eu estou chamando de semi identificadores, porque eles combinados podem reidentificar um indivíduo. E, por fim, um atributo sensível que, pode ser mais de um, mas normalmente são mais de um atributo, mas que trazem uma informação que a gente não quer que ela se torne pública. A primeira coisa a fazer é remover

o identificador, ainda que seja a primeira etapa, mas isso não é suficiente, remover o identificador, porque se a gente tiver uma fonte externa como esse caso da Netflix, ela não precisa estar pronta assim como eu trouxe nesse exemplo, mas com uma fonte externa a gente consegue, mesmo que parcialmente, começar a reidentificar parte desses registros. Então, a redução da informação pode ser feita em cima do conjunto de dado original, para fazer com que essa reidentificação não seja mais possível, ou pelo menos não seja mais um nível que ali registro a registro. Então, eu modifiquei o dado ali no primeiro registro de um indivíduo de 37 anos que vivia no Brasil e um de 38 no México. Agora, ambos estão no mesmo grupo, de 31 a 40 anos, vivendo na América. Eu tive uma perda de informação gigantesca aqui. Eu praticamente não utilizei esse dado, mas é justamente para evidenciar essa questão da brincadeira entre a utilidade do dado e o que eu consigo de confidencialidade. Em geral, você não vai implementar um exemplo tão extremo. Então, comentando sobre a situação de dados amostrais, em algum momento havia uma tranquilidade em divulgar dados amostrais porque considerava-se que dados amostrais, como não tinha a estação completa da população, você tinha alguma segurança em termos de sigilo. O que não é verdade. Em geral, você tem na população alguns indivíduos com características muito particulares. Se você faz um desenho amostral para melhorar a tua eficiência em algum aspecto com mais estratificação, o que quase, em geral, necessariamente você vai fazer, esses indivíduos vão estar alocados nesses estratos conglomerados, enfim, nessas unidades de análise, e o momento que você seleciona a amostra, nesse primeiro extrato, por exemplo, nesse exemplo que eu fiz, você tem, vamos dizer que esse indivíduo que está identificado na cor verde, ele seja, por exemplo, tenha respondido como cor, raça indígena, um descendente dos povos originários e uma característica que é visualmente identificável. Então, se esse extrato com uma escola, uma empresa, pessoas que se veem, e só existe um indivíduo com essa característica, todo mundo vai saber quem ele é. Se nesse microdado tem alguma informação sigilosa relacionada ao uso de drogas, em violência, algum tipo de aspecto relacionado a uma questão sensível como essa, todos os colegas de trabalho dele agora sabem quais são as características dele nesse sentido.

No segundo extrato, embora a gente tivesse dois indivíduos, um do sexo feminino e outro masculino nessa representaçãozinha binária do Bunday que eu fiz, a mesma coisa, só trazendo mais um atributo, além do color raça, estou acrescentando o atributo de sexo, agora eu sei quem é quem, eu sei que foi do sexo masculino que saiu na amostra, ou mesmo caso. Isso tudo para dizer que também existe uma abordagem de amostragem para dados sem citar os registros administrativos, é que ela também não vai ser muito efetiva, porque para você

eliminar os casos em que você tem um risco de reidentificação ou um risco de violação da confidencialidade alta. Você teria que selecionar uma amostra priorizando os indivíduos com risco baixo. Isso significa que você vai produzir uma amostra representativa dos indivíduos sem risco de reidentificação. Vai produzir uma amostra representativa da sua população original, você vai excluir, por exemplo, no exemplo que a gente está conversando aqui, os indígenas. Então é também uma abordagem muito limitada. Enfim, então isso seria meio como não fazer, não é?

E o que a gente tem inicialmente, o que são as abordagens mais, vamos dizer, tradicionais de se fazer? A primeira delas é a canonização. É um método bem conhecido, bem falado, que faz justamente aquilo que eu tinha comentado no exemplo. Eu converti aquele grupo de indivíduos entre 30 e 40 anos que viviam no México ou na América em um grupo de pessoas de 31 a 40 anos que vivem na América. Então, eu não consigo, mesmo com uma fonte externa, identificada com as mesmas variáveis, os mesmos detalhes, os mesmos indivíduos representados, eu não consigo mais saber quem é quem. Eu consigo saber que tem quatro indivíduos do lado direito que estão presentes em algum grupo desses quatro indivíduos do lado esquerdo, mas eu não sei exatamente quem é quem. Isso me dá um sigilo absoluto? Não, não me dá confidencialidade absoluta, mas acho que nenhum método dá. A brincadeira é o quanto eu estou disposto, que risco eu acho que faz sentido correr. O fato é que, em princípio, eu não conseguiria violar ou identificar registro a registro as duas bases. Mas tem um problema nesse caso aqui, e às vezes a canonização não é suficiente, mesmo que esse risco de reidentificação esteja adequado para você, que é, nesse exemplo, os quatro indivíduos que eu reuni no primeiro grupo e os quatro que eu reuni no segundo grupo, eles têm o mesmo atributo sensível, o mesmo valor do atributo sensível. Então não importa, se eu conseguir identificar quem são esses quatro, mesmo sem saber quem é quem, mesmo sem fazer um pareamento unívoco dos registros, eu sei que todos aqueles indivíduos tiveram câncer e do outro grupo todos eles tiveram zika. Então eu revelei de qualquer forma o atributo sensível, mesmo sem conseguir efetivar o pagamento. O método que trata isso é a diversidade, tem métodos, esses são os básicos, tem métodos muito mais sofisticados que tratam esse problema de uma maneira muito mais automática, inclusive os dois aspectos juntos. Mas eu acho que é um conceito importante, um caminho interessante para onde se começa a relacionar esse problema. Então, uma maneira de resolver isso seria incorporar a diversidade na canalização, e agora eu transformei aquele grupo, que era de 31 a 40, em 31 a 50, porque quando eu faço isso, eu junto o primeiro grupo com o último e agora eu não sei mais quem decidiu que teve câncer ou zika. Então, eu tive de novo uma perda de informação.

Então, para salientar o fato de que esse conjunto de dados está com uma utilidade muito baixa, só por eu agregar a idade, eu já inviabilizo algum outro trabalho, por exemplo, eu já dificulto um trabalho de regressão. Estou o tempo todo perdendo utilidade no dado, não necessariamente precisão, perdendo utilidade nesse dado.

Uma outra abordagem que, eu falei de atributos categóricos, uma outra abordagem bem comum também, bem explorada na literatura, é a da agregação, que ela não serve só para atributos numéricos, mas nasceu para isso e é bastante utilizada para isso. Eu coloquei mais dois exemplos de atributo, altura e peso, imaginando que eles fossem sensíveis. O que eu faço é tentar identificar os indivíduos que são de alguma forma próximos segundo esses dois atributos. Como são só dois, eu consigo fazer esse diagrama, esse quadrinho, e a partir desse gráfico eu vejo quais são os indivíduos que estão próximos uns dos outros, não é? E aí eles passam a ser representados pela média do grupo, que é esse asterisco na mesma cor do grupo. Então você pode ver que tem um indivíduo aqui na primeira linha, que ele tinha 1 metro e 79 de altura e 93 quilos. Ele está muito próximo da média do grupo, ele passou a ser representado por um indivíduo de 1 metro e 79 com 94 quilos. Mas o último indivíduo desse grupo, na quarta linha dessa tabela, ele tinha 1 metro e 68 e 98 quilos. Agora ele tem 1 metro e 79 e 94 quilos. A perda de informação foi enorme nesse caso. Mas é uma abordagem que também pode ser combinada com as demais.

Bom, isso foi o Output Privacy. Agora eu vou falar de novo. Aquilo foi bem superficial. Eu acho que tem o universo, além daquilo, tem décadas de desenvolvimento acadêmico em cima desse assunto, que eu falei, foram os métodos realmente bem elementares, mas todos são, de alguma forma, uma variação dessa lógica, dessa ideia, desse raciocínio de reduzir a utilidade para trazer a confidencialidade ao vice-versa. O Input Privacy, esses métodos, eles têm um olhar um pouco diferente. Eles são, de certa forma, um compartilhamento de microdados sem compartilhamento. Na verdade, você faz um processamento remoto. Então, existe alguma maneira de você processar o dado, sem que ele precise necessariamente ser compartilhado. Tem uma referência aí do grupo de privacidade, do Comitê de Especialistas em Dignidade e Ciência de Dados para Estatísticas Oficiais da ONU. O primeiro método, eu vou comentar rapidamente, o Multiparty Computation para dar um exemplo de como esse método funcionaria, para tornar ele um pouco mais palpável. Imagina que a gente vai calcular a idade média dos indivíduos no ambiente, só que eles não querem revelar sua idade. Então vai ter um outro indivíduo neutro, e ele vai pegar uma calculadora, inicializar com valor aleatório qualquer que ele selecionou lá na cabeça dele, passar essa calculadora para o primeiro indivíduo que vai somar aquele valor à idade dele, o segundo indivíduo vai fazer a

mesma coisa, o terceiro indivíduo a mesma coisa, e por último essa calculadora volta para aquele indivíduo neutro, que a gente está chamando de Computing Party, que vai subtrair o valor de inicialização aleatória e dividir pelo número de indivíduos que estão na sala, obtendo a média da idade sem que ninguém saiba a idade um do outro. Isso, na realidade, no Multipart Computation, nem a Computing Party recebe o resultado final. A Resulting Party é só a parte interessada no resultado que solicitou a computação, só ela vê o resultado. Nem a Computing Party veio, mas esse exemplo é mais para ilustrar o conceito como funciona, a implementação é bem mais complexa do que isso, mas a essência basicamente é essa. Você tem a fonte do dado, que é Input Party, você tem a Computing Party, que vai fazer de forma distribuída esse cálculo sem que nenhuma delas tenha nenhum aumento acesso ao dado completo, e pôr fim a Result Party recebe o resultado da computação.

Um outro método que também é bem falado é a encriptação ou a criptografia homomórfica, em que, para resumir bastante, você tem um dado, esse dado é criptografado, ele é enviado para um servidor de processamento que faz a computação sem descriptografar esse dado e ele retorna, só quando ele volta para a parte solicitante é que ele é novamente descriptografado e o resultado é obtido. Na verdade, existe uma variação disso, que é realmente a que se considera para compartilhamento de microdados, que é a Full Homomorphic Encryption, que é um pouquinho mais complexa do que isso, com o problema de que isso tem um custo computacional bastante alto. A produção desse método não é uma solução, nenhum desses métodos talvez seja uma solução geral. Cada um deles tem uma adequação que faz mais sentido em determinados aspectos. Sendo que essas abordagens de Input Privacy, como vocês já devem ter percebido, elas exigem uma questão de infraestrutura diferenciada. Você precisa ter essa estrutura de servidores de processamento, essas redes seguras de envio de arquivos, ou de parcela de arquivos, enfim. E fora o usuário que vai ter que ter uma especialização para tratar isso. Acho que é um método de compartilhamento muito seguro, muito efetivo, mais custoso e que talvez não seja a solução para tudo. Alguns usuários mais avançados vão poder adotar essa solução, outros não. Com a vantagem que, em geral, o usuário mais avançado é aquele que tem uma aplicação que exige um dado mais detalhado que, eventualmente, vai poder ser disponibilizado por esse método. Mas, enfim, essa discussão é bem profunda. Aqui eu só mencionei algumas aplicações que estão sendo feitas, na verdade, por instituições públicas. Então, usando essas técnicas de Input Privacy, isso já é realidade, embora não seja uma realidade tão próxima para a gente. Bom, só para encerrar essa frase do Gabriel Garcia Marques, que diz que todo escritor tem três vidas, uma pública, uma privada e uma secreta, para salientar essa diferença entre o aspecto da

privacidade e do segredo. A gente enquanto produtor de estatística, detentor de dado, a gente está na linha de frente da garantia dessa questão da privacidade. Isso não é só por respeito ao usuário, não é só por respeito ao informante que cedeu informação e confiou na instituição. É uma questão de sobrevivência, porque se a gente está falando que esse usuário precisa confiar cada vez mais na instituição, porque cada vez mais ele vai ter autonomia através de uma LGPD, por exemplo, para decidir que instituição recebe o dado dele ou não, a instituição tem que cada vez mais estar preparada para transmitir essa confiança e também para ser capaz de entrar nesse grupo de compartilhamento de dados com outras instituições, porque outras instituições também reconheçam ela como um parceiro confiável.

Erivelton Pires Guedes - IPEA

Vou falar da experiência que a gente teve de RAIS no Ipea, porque não é a mesma RAIS que todo mundo conhece, mas fizemos uma adaptação aqui para o uso que a gente se propôs. O foco desse trabalho foi no que nós chamamos de Atlas do Estado Brasileiro, depois eu convido a todos para olhar no site e ver o trabalho, fizemos um raio x do funcionalismo público brasileiro nos últimos 36 anos, de 1985 a 2020. Então, a RAIS do Ipea tem um histórico de mais de 30 anos, e ao longo desses 30 anos a gente viveu vários desafios e mudanças tecnológicas. Então, há 20, 30 anos atrás, para trazer o dado aqui para o Ipea, era uma dificuldade imensa, eram arquivos pequenos, arquivos zipados, com qualidades diferentes entre cada arquivo. Então, na prática, nesses 36 anos de RAIS que a gente tinha aqui, eram 36 arquivos gigantes, eles eram gigantes em CSV, e internamente a gente os transformava no formato size, horas e horas de processamento, de leitura de tráfego, de rede, processos repetitivos, erros persistentes, porque você começava a ler o arquivo, aí ele parava no meio do caminho, aí você voltava lá do início, e além disso era difícil compartilhar esses arquivos. Bom, o que nós fizemos? Como íamos mexer com 36 anos de RAIS, a gente começou fazendo um teste e montamos um cluster com cinco computadores. Pegamos cinco computadores velhos, estavam desligados, juntamos os cinco com um software especializado nesse assunto e jogamos todos os arquivos CSV lá para dentro. Fizemos também um teste trazendo esses arquivos CSV para um banco de dados relacional. Bom, o resultado prático diz o seguinte: a gente levava seis horas só para arquivo de um ano, trafegar da rede de Brasília para o Rio, que foi o trabalho que a gente fez, e para cada ano a gente levava mais de 18 horas para importar o arquivo para um servidor PostgreSQL. Num servidor aqui de Brasília ainda era mais demorado um servidor pago, que é o SQL Server.

Bom, no fim a gente conseguiu ter a seguinte estrutura, 36 anos de RAIS numa tabela única. Como é que a gente fez isso? Pegamos a estrutura dos 36 anos e fomos estudando para ver qual era a estrutura comum que tinha naqueles 36 anos. Então, de um ano para o outro, mudava o nome de campo, às vezes mudava só uma letrinha, às vezes tinha um erro de grafia, ou às vezes mudava o nome, mas era a mesma variável, ou às vezes mantinha o nome e a variável tinha mudado algum conceito. Então, foram várias nuances ao longo desse período. A gente também georreferenciou os endereços dos CNPJs, dos 10 milhões de CNPJs de empresas ativas naquele momento de 2020, eu acho, pegamos tabelas do CIAP e juntamos com tabelas de eleições para fazer algumas comparações, mais as tabelas auxiliares, CBO, CNAI, etc. E também as tabelas geográficas que importamos do IBGE, países, municípios, UF, etc. Com isso, chegamos na seguinte situação: temos um banco de dados único. Ele está em PostgreSQL. É um banco de dados relacional gratuito. Podemos conectá-lo por RStudio, por Python, com o Debiver, que é uma ferramenta que usamos para gerenciar. No QuantumGins, que é um software de geoprocessamento. Com isso, conseguimos gerar uma maior consistência dos dados, porque principalmente compartilhamos com vários colegas e ao compartilhar com os colegas, várias equipes diferentes usavam os mesmos dados, e isso começou a gerar um pouquinho mais de consistência, porque alguém achava, olha, tal variável está com alguma coisa estranha. A gente verificava a importação, o tratamento que foi feito, e a gente acabava consertando ou melhorando aquele dado. Para vocês terem ideia, a gente fez isso com uma máquina antiga. Era um servidor que estava para ser descartado, tem 20 gigas de memória, mas ele teve um desempenho muito bom. No momento atual, estamos com novas estruturas tanto no Ipea Rio quanto Brasília, tem máquinas mais novas, mas eu estou colocando essa antiga aqui só para vocês terem uma referência, de tamanho, para ver que 36 anos de RAIS cabe numa máquina mais ou menos normal para os dias atuais, desde que bem organizado. Bom, na situação atual, a gente está com rotinas automatizadas de importação, está integrado com R e com Python. O Deliver, que é uma ferramenta de visualização de banco de dados, de visualização e gerenciamento. Ela é gratuita. Ela permite a gente ver o banco de dados de forma fácil. E uma coisa importante, a gente fez o controle de código fonte no GitLab. Quem programa deve conhecer o Git. Git é uma ferramenta muito importante para você gerenciar códigos. Uma outra coisa que a gente aprendeu também fazendo foi a criação de views materializadas. O que acontece? Eu tenho uma base gigantesca, são quase dois bilhões de linhas. Se eu for fazer uma consulta toda hora nessa base, simplesmente vai levar algum tempo. Por mais que esteja otimizada, demora. Então, para aqueles dados que a gente já sabe que vai consultar várias vezes, criamos, no possível,

o que chama - se view materializada. Na prática, ele transforma uma view numa tabela e já está pronta. Então, quando eu a consulto a segunda vez, ela já levou aquela, hora ou duas que tinha para se construir, mas na segunda vez ela me mostra em alguns segundos. A desvantagem da view materializada é que, se eu atualizar o dado, ele não atualiza a consulta. No caso dessa RAIS aqui, que a gente está falando de atualizações anuais, funciona perfeitamente. Eu crio a view materializada, ele leva uma hora fazendo o processamento ali, e a partir daí tudo o que eu fizer nessa view, eu vou levar alguns segundos. Bom, até hoje, a gente ainda tem arquivos texto na nossa rede. Não conseguimos convencer toda a equipe do IPEA a usar os bancos de dados. E a gente acabou tendo dois bancos separados, um do SQL Server, que é uma ferramenta paga da Microsoft, que está aqui em Brasília, e a PostgresSQL, que é uma ferramenta gratuita que está lá no Rio de Janeiro. O servidor de Brasília, apesar de mais poderoso, ele tem respondido mais lento. Eu não sei dizer se isso é uma inexperiência dos programadores, que sou eu e outros aqui, eu sou engenheiro civil, então a gente acabou aprendendo a programar e usar bancos de dados meio na força bruta e na prática, mas é o que aconteceu. O servidor de Brasília, por outro lado, por ser SQL Server, está melhor integrado à rede Windows. Então, se eu quiser fazer coisa em Excel, esse servidor SQL Server é muito mais prático, muito mais intuitivo. Outro detalhe importante é que o SQL Server que temos aqui em Brasília tem a ferramenta R embutida nele. Na verdade, eu confesso que não conseguimos usar esse R embutido ainda. Mas aparentemente esse R embutido talvez traga um poder maior de processamento para esse SQL Server. É algo que nós vamos ter que explorar ainda. Apesar de não ter o R embutido, a gente conseguiu instalar o R Server no mesmo servidor do PostgreSQL e acabou que isso ficou uma coisa muito rápida. O que aconteceu com esse R Server embutido no servidor? Pense na seguinte situação. Eu tenho uma única máquina que tem o banco de dados e tem o software estatístico. Então, quando eu faço qualquer manipulação estatística ali, eu não estou trafegando o dado em rede nenhuma e nem saindo daquele mesmo HD. Ou seja, eu estou ficando na mesma máquina, eu estou no mesmo HD, no mesmo servidor de banco de dados e na mesma memória. Então, isso fez com que esse processo de servidor ficasse bastante rápido. A gente vai futuramente tentar explorar esse R que está embutido na SQL e se é para ver se ele tem alguma melhoria em relação a isso que a gente acabou de fazer. E a gente está tentando pensar aqui em novos clusters. A gente montou um cluster novo aqui em Brasília só com máquinas virtuais. Veja, o cluster que eu montei antes tinha 20 Gigas de memória. A gente vai montar um agora com 3 terabytes. Ele não tem garantia física da máquina, porque é uma máquina velha, mas o tempo que funcionar para nós é longo. Esse aqui é um Shiny que a

gente fez, ele está hospedado na mesma máquina, então a gente buscava aqui um CNPJ específico, ele fazia uma conta lá em 2 milhões de linhas e levava cerca de 30 segundos para me dar um raio X de um único CNPJ. Claro, é um único CNPJ, então ele restringiu a consulta, mas ele faz isso de forma muito rápida. Uma coisa que a gente aprendeu também na modelagem do Banco de Dados é a criação de índices. Então, nesse exemplo aqui, se eu busco na base da RAIS por CNPJ, eu tenho que ter previamente criado um índice lá no Banco de Dados para dizer para o banco, olha, eu quero frequentemente usar CNPJ com busca. Se eu não faço isso, essa consulta que hoje leva 30 segundos, ia levar meia hora a uma hora. Foi a experiência que fizemos antes. E quais índices a gente criou? Bom, a gente foi criando conforme a necessidade. Então, frequentemente a gente precisava consultar por UF, por município, por CNPJ, por CPF. Na prática, quando eu crio um índice no banco de dados, eu estou aumentando o tamanho do banco de dados e diminuindo o tempo de processamento. Então, você perde no espaço em disco e ganha no processamento. Cada um vai ter que fazer a sua própria conta. O nosso banco de dados começou com 1 terabyte de espaço em disco em arquivo CSV, ele virou 1 terabyte de novo lá quando eu passei para o banco de dados, e depois que a gente criou os índices, ele duplicou de tamanho. Então, eu chutaria nossos índices aqui, eles custam, em termos de espaço em disco, o mesmo tamanho do banco. Mas o desempenho é muito bom. Tudo aqui é feito de forma bastante rápida. E aqui é uma pequena propaganda desse software, DBeaver, o site dbeaver ponto io, ele é um gerenciador universal de bancos de dados. Eles se conectam em arquivos CSV, DBF, Access, Oracle, MySQL, SQL Server, PostgreSQL. Então, eles se conectam em muitos bancos de dados com a mesma interface, com a mesma facilidade. Então, para quem usa bancos de dados, recomendo fortemente a ferramenta fantástica, 100% gratuita. E a versão paga, mas a versão gratuita tem 99% das coisas que a gente precisa. Para conectar no PostgreSQL é três linhas de comando. É uma linha que eu chamo a biblioteca, uma linha que eu passo o endereço do banco e o usuário e pronto. Daí para frente eu digo me mostra a tal tabela, e daí para frente ele é um data frame lá no R. Estou dando um exemplo só do R aqui do Python, é bem semelhante, tá? Aqui tem uma carinha do R rodando aqui dentro do sistema, esse R aqui também, ele está um RStudio integrado lá no servidor, ou seja, o fato de ele estar na mesma máquina multiplica em muito desempenho, tá gente? Tem uma carinha do banco de dados, e, por fim, a gente está em andamento hoje a desidentificação. A gente começou a fazer isso um ano atrás, e o que aconteceu? Esse 1,7 bilhão de registros de pessoas chegou à conclusão que são 200 bilhões de brasileiros que estão ali dentro. No entanto, se você junta CPF com PIS, tem parte de CPF com PIS que tem 2 mil pessoas diferentes ali dentro. Então, CPF e

PIS não é suficiente para identificar uma única pessoa. Tem muita mudança de nome, tem grafia de nome, tem grafia errada, nome que é com S e que é com Z, no outro ano é S ponto, a pessoa casa ou separa, troca de nome, e tem erros também e alterações nos outros dados, tipo data de nascimento, às vezes a data de nascimento se altera de um ano para o outro, mas você percebe que é a mesma pessoa. Para você pegar uma tabela com 1,7 bilhões e transformar em 200 milhões, está dando um certo trabalho. A gente está com uma rotina que, da forma como fazemos, ia levar 90 dias. Então paramos de rodar, fizemos só um teste e vamos esperar montar esse cluster novo com mais memória para a gente rodar rapidinho. Até porque a gente sabe que não, quando a gente faz um trabalho desse você roda a sua rotina. Depois que ela termina, você percebe que, eventualmente, tinha algum erro na sua própria ideia. Aí a gente vai esperar 90 dias para descobrir o erro. Não, vamos fazer um testezinho aqui, preparar a máquina melhor. Então, nesse momento, estamos nessa situação.

Thais Paiva - UFMG

Primeiro, eu queria agradecer o convite. Eu trabalho muito com o método que o Augusto falou um pouco no início da sua apresentação. Eu fiz o doutorado nos Estados Unidos entre 2010 e 2014, foi quando eu conheci essa área de pesquisa, fui fazer o doutorado em estatística, e aí comecei a trabalhar com o meu orientador lá, o Jerry Reiter, que faz muito trabalho sobre métodos estatísticos de privacidade, é um nome bem conhecido. E desde então tenho trabalhado com isso, desde 2016 na UFMG. Dentro daqueles métodos que o Augusto mencionou, eu trabalho mais especificamente com métodos de geração de dados sintéticos.

Vou falar um pouco de um trabalho que estamos aplicando a um banco de dados de Covid aqui de Minas Gerais. Trata-se de uma extensão do método que eu comecei a desenvolver lá no meu doutorado para gerar coordenadas geográficas sintéticas. E a nossa motivação é bem o que o Augusto falou, como divulgar bases de dados para que elas sejam utilizadas, então nós como estatísticos queremos sempre bases de dados interessantes, a gente quer explorar toda a informação das bases de dados, mas que isso seja feito de maneira segura, protegendo a privacidade dos indivíduos. A gente quer sempre controlar essas duas coisas, a utilidade e o risco de identificação de informações pessoais. Então, como já foi falado bastante, a gente tem dados hoje sendo coletados sobre tudo que a gente faz o tempo todo, dados georreferenciados que até um pouco tempo atrás era mais difícil, mas hoje qualquer pessoa tem acesso a um equipamento que vai te dar coordenadas geográficas bem precisas sobre vários eventos, vários sistemas que comunicam entre si e conseguem juntar essas informações, e isso vai trazendo desafios, tanto da parte de armazenamento e análise de bases

cada vez maiores, como a manutenção da privacidade e do sigilo dos dados, que é o que eu tenho trabalhado mais especificamente. Então, a gente pode revelar a identidade dos indivíduos, também divulgar, mesmo que se tire identificadores, alguns atributos sigilosos, tem vários riscos, além de violar princípios éticos, isso fere também a confiança que os indivíduos vão ter nas instituições que estão coletando dados, não vão querer participar de pesquisas posteriormente. Então, todas as instituições que coletam dados têm que tomar medidas para proteger a privacidade. A gente tem várias legislações específicas, como foi falado também, LGPD, algumas nos Estados Unidos, essa que eu coloquei específica de dados de saúde, lá um pouco mais pulverizado, mas também tem um equivalente na Europa. E aí alguns dos métodos para a gente mascarar os dados que eu costumo ver mais na literatura de, posso agregar os dados, suprimir algumas observações, trocar dados entre indivíduos e adicionar ruído aleatório, mas a maioria desses métodos vão distorcer as relações dos dados e afetar qualquer análise estatística, então a gente vai acabar jogando fora muita coisa do que a gente gostaria de ter, ao falar assim, olha, eu tenho um dado aqui super rico, uma informação super rica, mas está todo, misturado, está tudo distorcido, então todas as análises que eu vou fazer vão ir propagando essas distorções. E aí que entra a ideia dos dados sintéticos, que são bem interessantes, onde a gente substitui os dados originais por dados simulados, que chamamos de dados sintéticos, gera-se parte de distribuições de probabilidade que vamos ajustar aos dados originais. Então, desde que se incorpore essas informações, a gente vai conseguir, ao analisar os dados sintéticos, fazer as mesmas análises, as mesmas inferências que eu faria com os dados originais, com a vantagem de que eu estou divulgando dados simulados, então não é dado original. Não é dado verdadeiro. Então a gente consegue um controle muito maior, ainda existe o *trade-off* de privacidade e utilidade que a gente precisa controlar, mas existe um controle muito maior de manter essa privacidade e de preservar a privacidade e manter a utilidade das análises. E aí, especificamente, que eu tenho trabalhado bastante, são dados georreferenciados, quando a gente está falando de localização, que a maioria das vezes nem se pode divulgar. Então, quando se está falando de alguma coisa que está relacionada à localização espacial, pode ter um padrão espacial, alguma coisa que está acontecendo ali naquela cidade, naquele estado. E se a gente simplesmente falar, não posso divulgar, joga fora de novo toda a informação que o banco de dados teria. A maioria das agências vai agregar esse dado, não é? Não vai divulgar, ou vai ter restrições, como a gente viu lá no comecinho da oficina. A gente pode ter todas as demais estratégias, que eu falei antes também, de adicionar o ruído aleatório, trocar as localizações, e aí, de novo, a gente vai torcendo todas as distribuições originais, e também simular

localizações sintéticas, que é o que eu tenho trabalhado mais de maneira mais aprofundada. A gente propôs esse método de simulação de coordenadas geográficas sintéticas lá atrás, e desde então temos trabalhado em algumas extensões para ele. Acho que eu não vou falar muito da parte da metodologia - a gente pega os dados originais, então o que eu estou chamando aqui de X e Z seriam os atributos não espaciais, o S as coordenadas, latitude e longitude nesse caso, o X vão ser os atributos discretos e a gente incorporou também a possibilidade de incluir uma variável contínua, por exemplo, idade. E a gente vai usar esse dado original, a gente não está falando especificamente de como mascarar a parte do X e do Z, só do S. Então teriam outros métodos para variáveis discretas e contínuas que você poderia aplicar nessa partezinha cinza dos dados. Aqui a gente só vai usar as coordenadas geográficas. A gente vai fazer um modelo que vai olhar como que as coordenadas S estão distribuídas considerando as características X e Z. Então, tem mais mulheres morando nessa região, pessoas mais jovens, mais velhas dessa raça e assim por diante, dependendo dos atributos que estiverem contidos no X. Então, a gente está modelando aqui a distribuição do S dado o X e o Z. Aí a gente seleciona algumas amostras e para cada uma delas vai sair um banco sintético, de mesmo tamanho do original, a gente está fazendo aqui completamente sintético, e a gente tem M-cópias, isso é o princípio que vem lá da imputação múltipla, Ao invés de substituir por um único banco, a gente substitui por M bancos. Então, eu vou ter que repetir qualquer análise que eu faça nos M bancos separadamente e depois combinar isso. E isso me permite estimar a variabilidade que vem do processo de imputação. Se eu substituo por uma única cópia, eu falo que com certeza é essa daqui. Quando eu tenho esta ação múltipla, eu tenho a variabilidade entre esses M cópias. E aí os dados sintéticos que eu vou divulgar, aqui nesse caso X e Z permanecem os originais, e eu vou substituir o S lá. O S vai ficar escondido, então ninguém vai saber dele, ninguém vai ver ele. São as minhas coordenadas confidenciais. E a gente controla, nesse caso específico do meu método, o tamanho da grade que a gente usa para estimar, de novo, controlando o risco e a utilidade. Lá no meu artigo tem algumas medidas, então tem uma literatura grande, específica sobre isso, que tem medidas para a gente tentar mensurar o risco de identificação. A gente assume, por exemplo, ao divulgar as coordenadas sintéticas, se eu tiver um invasor querendo descobrir onde que uma pessoa específica mora. Então, qual que vai ser essa probabilidade? Então, tem várias medidas nesse sentido. E a gente também tenta mensurar a utilidade daquele banco sintético para que a instituição decida se aquilo ali faz sentido ou não para aquele banco. Aqui são alguns resultados preliminares, como eu falei esse trabalho ainda está em andamento, mas a gente teve acesso a um banco de dados de Covid entre 2020 e

2021 na cidade de Montes Claros. Então, a gente teve que fazer um processo de tratamento bem grande do banco, só mostrando um pouco dos bancos. Esses aqui seriam o próximo do banco completo. A gente teve que tirar algumas observações isoladas, a gente teve que adicionar ruído, tirar as coordenadas para que a prefeitura concordasse, que a gente divulgasse. A gente teve que assinar vários termos para poder divulgar esse banco. Como tem um banco bem grande, bem denso, fica um pouco mais fácil de poder divulgar isso. Tinha algumas observações aqui que a gente teve que excluir para não identificar. E aqui só o que seria o resultado final, mais ou menos. Hoje em dia eu acho que essa informação ficou menos sigilosa, mas lá em 2020 tinha várias pessoas que ainda tinham preocupação de querer divulgar essa informação ou não. E junto com esse banco, por exemplo, eu sei que tem uma pessoa que mora sozinha em um determinado bairro ali, eu conseguiria ver várias informações sobre aquela pessoa além da situação dela de uma outra doença. Poderia ser o Covid, mas poderia ser câncer, ou AIDS, ou uma outra doença que a gente não gostaria que as pessoas pudessem identificar os moradores. Então aqui a gente selecionou mil coordenadas, mil observações aleatoriamente do Banco Original, todas passaram por aquele critério de se são observações isoladas, então não são plotadas por negativo e positivo. Se elas estavam com Covid ou não na época do atendimento. E aqui seriam quatro exemplos de dados sintéticos. Em dados sintéticos são quatro bancos diferentes, mas que os padrões espaciais são bem parecidos com os originais. Então, se eu fizer alguma inferência aqui no banco sintético, eu volto para os resultados que eu teria, assim, são próximos com os resultados que eu teria com os dados originais. De novo, também plotados por negativo e positivo. Então, se eu quiser contar quantas pessoas com teste positivo estão num bairro tal, eu posso fazer essa análise no banco sintético. E o banco original fica escondido, ninguém vai ter acesso a esse banco original. Só os dados sintéticos. Então, ele abre muitas portas para a utilidade de várias análises, porque a gente está possibilitando, inclusive, de divulgar o microdado. Não precisa ser só dado ou agregado nesse caso. Então, se você tiver um banco de dados qualquer, você consegue gerar suas próprias coordenadas sintéticas e fazer todas essas análises essas mensurações do risco utilidade A gente está trabalhando também na melhoria da eficiência computacional, porque agora a gente tem coordenadas discretas e contínuas também que podem ser utilizadas para modelar aquela parte do X e do Z. E finalizar, estamos finalizando também o ajuste para os dados de convite para o banco completo, que foi um desafio grande também, que o banco tem mais de 120 mil observações, 40 e tantas variáveis, então foi um trabalho grande de limpeza, tratamento do banco de dados,

quando a gente pega ele a primeira vez, tem vários pepinos, então gastamos um tempão com isso.

Considerações Finais

Augusto Fadel: Eu acho que a gente falou bastante aqui sobre utilidade e risco, mas tem uma outra dimensão quando a gente vai escolher um método, que é a complexidade em dois aspectos, de implementação e para o usuário. Tem, por exemplo, esses métodos de Input que carregam uma complexidade muito grande para ambos, para a instituição detentora do dado implementar e para o usuário. Outros métodos nem tanto, então para você entender melhor como a coisa funciona. Então não existe um método que vai resolver tudo, provavelmente a gente vai ter que conviver com alguns tipos de divulgação diferente. Tem a questão do dado tabular, que a gente divulga o dado em tabela, se eu divulgo o meu dado, não preciso mais divulgar o dado em tabela. No momento particularmente eu não concordo com isso, eu acho que o dado em tabela atende uma parcela de usuário muito importante e vai continuar atendendo. Acho que a gente tem que trabalhar aí nesses microdados de uso público, que tem esse nome de public user files. Porque eu acho que ele atende uma parcela que já talvez já estivesse não tão bem atendida e vai ficar pior, porque com essa insegurança jurídica que o Felipe também colocou aí na apresentação dele, a gente tem que ter atenção a isso. Mas enfim, acho que é uma... por isso que o debate é interessante e longo. Acho que não tem solução única e a gente vai ter que realmente conviver com trabalhar e conviver com várias soluções, provavelmente. É isso, obrigado.

Erivelton: Tem duas coisinhas que eu esqueci de falar. Uma função que a gente usou legal também no Banco de Dados foram tabelas particionadas. O que é isso? Apesar de ter 36 anos, se eu quiser consultar um único ano, eu posso ir diretamente naquele ano fazendo um filtro simples por ano, e aí eu estou indo numa tabela única. Na verdade, eu tenho 36 tabelas. O comportamento delas é como se fosse uma só. Falta eu explicar esse detalhe. E eu esqueci de mencionar também uma coisa importante que pode interessar, principalmente a DIEESE, é que existe uma base de dados públicas chamada base dos dados que é um grupo que começou a organizar dados públicos abertos e dados que são difíceis manusear, são complexos, são grandes, mas eles estão colocando essa base na infraestrutura do Google. Então, só para terem uma ideia, uma consulta aqui, se eu quiser fazer uma consulta gigantesca aqui, eu levo 10 horas de processamento na minha base e levo 17 segundos na

base deles. E eles têm basicamente a mesma base que nós temos aqui, com exceção de que não tem nada identificado lá. Mas, por exemplo, se você quer fazer apenas grandes agregações, sem entrar em alguns detalhes, a base deles é fantástica, a consulta é super rápida e eles têm API para R, Python ou direto no Google. Só isso que eu queria completar por enquanto

Thais Paiva: Bom, vou complementar algumas coisas: sobre a questão de qual o limite da precisão. No caso da nossa metodologia, a gente gera uma coordenada X, Y, uma latitude e longitude sintética. Aí, na verdade, a gente vai divulgar várias cópias, então você não sabe, nenhuma delas é a localização real de uma pessoa. Mas, você vai ter várias coordenadas ali correspondentes que são sintéticas, então você não sabe onde que aquela pessoa mora de verdade, nenhuma delas é a verdadeira. Se isso for possível, você vai divulgar essas bases sintéticas. Mas a gente pode parar um nível antes. Eu não expliquei a fundo o método, mas pode ser uma grade regular artificial que a gente joga por cima do mapa. E essa grade é o que vai regular o meu nível de utilidade e de segurança, o meu risco. Então, se eu fizer uma grade mais fininha, eu vou estar muito próxima da realidade, então pode ser que as minhas coordenadas sintéticas fiquem praticamente todas no mesmo quarteirão, e eu não estou preservando muita coisa de segurança. A utilidade vai ficar super preservada, as análises que eu fizer nessa base vão ficar ótimas, mas eu estou quase que só jogando um ruídozinho bem pequeno. Ao contrário, se eu fizer uma grade muito esparsa, vamos supor que eu jogo só uma célula grandona. O que eu vou fazer no final das contas é jogar todo mundo aleatório no espaço. Eu perco toda a minha utilidade, porque aí eu não tenho nenhuma preservação espacial da distribuição espacial mais. Mas eu maximizo a minha segurança, porque agora está todo mundo aleatório no mapa inteiro. O equilíbrio está justamente nesse controle do tamanho da grade, no caso do nosso método. A instituição, quem tem os dados, o dono dos dados, é que vai controlar esse botão para saber se ele vai mais para um lado ou mais para o outro. Mas de qualquer forma a gente existe a possibilidade de divulgar. Então eu vou ter lá uma linha do microdado que é todas as características que eu posso estar querendo divulgar originais ou não, mas a parte não espacial, aquela parte do X lá que eu mostrei, a linha vai inteira completa e várias coluninhas de coordenadas X e Y. Latitude e longitude para aquelas pessoas todas do banco. Então eu tenho o nível máximo, posso dar o zoom que eu quiser, posso fazer dado agregado ou não. Uma outra possibilidade que a gente não explorou nos resultados, nem na simulação, mas pode ser aplicada, porque o modelo estatístico se adapta muito bem, ele é um modelo para dados de área. Então se eu tiver, por exemplo, bairro, setor censitário, ou alguma outra região que não seja a grade regular artificial, mas que eu tenha

estrutura de vizinhança, eu posso fazer meu método usando-a. Então, ao invés de fazer a minha grade artificial, eu posso aplicar usando, por exemplo, o setor censitário e vai funcionar também. Dentro do setor censitário, eu vou gerar, eu escolho o setor censitário sintético. Eu vou ter um lambda que me fala quais são as probabilidades de estar em cada um dos setores sensitários para mulheres, negras, com essas escolaridades e assim por diante. O lambda me fala isso, para cada setor censitário, para cada célula da grade. E aí lá dentro eu gero o uniforme. A gente assume que o lambda é constante dentro do quadradinho, ou do setor sensitário ou da célula da grade. Então, existe essa possibilidade de aplicar o método nesses dois sentidos. E aí vai depender da aplicação. Então, o detentor dos dados que vai decidir como fazer isso. E, mais importante, usar as medidas de risco e utilidade para decidir se vale a pena ou não. Porque, às vezes, um banco de dados de Covid, ele fala, não, vamos usar isso para fazer bastante análise, não tem tanto problema assim, mas agora, depois de dois anos, tanta gente já teve até a gente tirava as pessoas duplicadas para a gente só considerar um único indivíduo. Hoje a gente tem pessoas que já pegaram o Covid várias vezes, então a gente teria que mudar o banco de dados para inclusive considerar as pessoas, as diferentes ocorrências de um mesmo indivíduo, etc. Então perder um pouco essa questão de ser uma coisa muito sigilosa. Mas se for um outro banco de dados numa região, uma população muito menor, uma informação muito mais sensível, com certeza a gente vai ter outros critérios. Então, como regular esse tamanho, isso tudo vai depender muito da aplicação. E quem decide, quem vai controlar esse botão aí, vai ser a instituição. E aí o que o Augusto falou da questão do desafio das bases grandes, até então eu não tinha encontrado esse desafio ainda porque era muito difícil a gente conseguir ter uma base pública que eu possa aplicar, que seja legal e que eu possa aplicar e divulgar de qualquer jeito. Então sempre é uma base mais restrita, menos interessante, mais limpinha e muito menor. Agora com o banco de dados de Covid a gente já está tendo um pouco de desafio dessa questão do tamanho do banco, mas é um banco com 120 mil observações. Ele não é tão grande assim, poderia ser muito maior. Assim, o que vai acontecer é a gente, pelo que a gente está trabalhando agora nessa questão, como a gente faz uma contagem por células, o método não vai escalar tanto no número de observações. Ele escala mais com relação ao tamanho da grade e quantos atributos eu estou colocando. Isso que determina a maior complexidade do meu método no caso. Então, teoricamente, ele escalaria bem para um banco de dados muito maior, desde que eu não aumente tanto a minha região ou o meu número de atributos. Então, eu consigo aplicar isso a bases que aumentem de tamanho, porque não vai depender dessa ordem de grandeza. Mas, de qualquer forma, para que eu consiga estimar a variância que vem da imputação, eu

preciso de divulgar mais de uma cópia. Na literatura de imputação múltipla, a gente fala de quatro, cinco, dez bancos sintéticos, mas eu vejo muita gente falando de cinco e já é suficiente. Então eu vou ter que estimar o Lambda à parte cara computacionalmente. Eu faço isso só uma vez. Depois que eu tenho esse Lambda estimado, gerar quatro, cinco ou dez bancos sintéticos é quase que fácil. O difícil é estimar o Lambda. Então não seria tão problemático assim. Então, no caso, eu teria que ter quatro, cinco ou dez cópias, algo em torno disso, das coordenadas sintéticas que eu estou trocando. Eu vou guardar um banco original e vou substituir por quatro ou cinco sintéticos. Então, aumenta nessa ordem o meu volume de dados. Aí, o analista vai fazer a análise nesses cinco bancos sintéticos. A análise vai ser exatamente a mesma, estatística, se for uma análise simples, frequentista, que a gente fala assim, tem algumas fórmulas, eu posso passar a literatura depois, tem algumas fórmulas para eu poder fazer a combinação dos resultados. Quer fazer uma regressão? Aí eu faço nos cinco bancos e depois junto a esses cinco betas, vamos dizer assim, com as médias e uma formulinha para fazer os intervalos de confiança, que vai incorporar as duas partes de variabilidade. E, se for benziana, é ainda mais fácil, é só empilhar as coisas. Então, analisar o banco sintético também acredito que seja, se você tem o problema de analisar um banco sintético, um banco grande, você vai ter o problema de analisar quatro ou cinco bancos sintéticos. Aumenta nessa proporção, mas o número de bancos sintéticos não costuma passar muito de dez. Então Entre quatro e cinco, eu acho que fica até razoável ainda de se manipular. Uma outra coisa que eu queria comentar, acho que vendo bastante da discussão aqui, é que nem toda a aplicação vai funcionar para o banco sintético. Se eu preciso fazer uma estatística oficial, se eu preciso saber qual foi a verdadeira localização de um crime, alguma coisa assim, eu preciso do dado original. Não posso substituir isso para um dado sintético. Acontecer um crime em quatro coordenadas sintéticas não faz sentido. Para estatísticas oficiais, alguns órgãos ainda vão precisar das coordenadas originais, e aí tem a questão da restrição de acesso mesmo. Existe também toda uma literatura da estatística sobre como fazer um acesso controlado que de fato proteja, porque pode ser que ainda exista brecha, mesmo que eu não acesse o dado original, mesmo que eu só peça, por exemplo, algum resultado, então eu vou aplicar um banco, peço uma regressão ou um estimador para o servidor e o servidor só me retorna aquele número. Eu não tenho acesso ao banco completo, então eu só consigo saber o resultado das minhas pesquisas. Existem maneiras de fazer pesquisas que me devolvam informações muito identificável, digamos assim. Eu consigo fazer a combinação de alguns atributos ou alguma regressão, que eu vou ter praticamente a informação que eu gostaria de recuperar sem ter que colocar da mão no dado. Então existe

também essa literatura sobre essas maneiras de fazer isso. E outra coisa também que eu lembro muito de ver, meus colegas que trabalhavam como mesmo orientador lá, essa questão de como você consegue cruzar informações em bases diferentes. Eu tenho duas bases, uma tem um pedaço de uma informação, outra tem outro pedaço, mas se eu juntar aquilo ali eu identifico muita gente. E isso é bem problemático também.

Felipe: Só um comentário geral, como gestor de dados, a gente sempre sai, nesse debate de anonimização, cada dia com uma dúvida diferente. Interessante ver a abordagem que tivemos em uma auditoria recente do TCU sobre os dados do CAGED, nos questionando também, o motivo da retirada da divulgação dos dados do CAGED estabelecimentos, que entendemos que havia problema de facilidade de identificação e o técnico do TCU, que estava fazendo a auditoria, questionou por que a gente tirou, considerando, por exemplo, além de acesso à informação, a transparência do dado. E a abordagem hoje, em uma oficina com pesquisadores, a ponderação da divulgação, da segurança da informação e tal. Então, a gente está sempre atravessado por essas questões, e acho que a mensagem que eu queria deixar é que esse jeito que a gente faz hoje também é inseguro, né? A gente disponibilizar um TXT para algumas organizações que solicitam, que mesmo que passe algum critério de adequação formal, o mais criterioso que seja, ele também é vulnerável a diversos questionamentos jurídicos, não é? E hoje, enfim, a consultoria jurídica do Ministério ia adorar se a gente falasse, não vamos divulgar mais dados identificados, nem microdados para ninguém. A gente vive atormentando a vida da consultoria jurídica com essas questões para continuar viabilizando os trabalhos. Então construir saídas que evitem disponibilizar os TXTs também são saídas que a gente viabiliza como proporcionando maior segurança no médio prazo, porque a gente sabe que hoje a estrutura que a gente tem de disponibilização é frágil também. É por isso que a gente está pensando nesse projeto de construir bases mais integradas que diminuam a necessidade de disponibilização do TXT para fora do governo federal, porque até o momento a gente se sente mais seguro para disponibilizar dentro de órgãos do governo federal, mas para fora acho que tudo que a gente faz hoje, tanto a base da RAIS pública que a gente divulga hoje, quanto o acordo de cooperação que a gente celebra com organismos, é muito facilmente questionável nos termos da LGPD, e a gente, por outro lado, está fazendo o máximo para não gerar um apagão de dados, para não gerar questionamento do tipo que aconteceu com o MEC, com os dados da educação. Sempre estamos andando sobre uma linha fina, e por isso a queremos ouvir todo mundo e construir essa solução de forma coletiva.

Patrícia: Obrigada, a todos e todas. Estamos todos nós muito impactados com todos os elementos que foram trazidos, todas as possibilidades. Muitos obstáculos, muitas dificuldades, mas também muita luz. A gente não sabe exatamente o que vai conseguir construir, mas, tem muita coisa aí para a gente trabalhar a partir dessa contribuição tão generosa de vocês. Quero agradecer de novo, muitíssimo, a disponibilidade de todos e todas. O objetivo é, a partir daqui, com esses elementos, quebrar um pouco a cabeça, para construir uma proposta de caminho para esse desenvolvimento. Sabemos que isso aqui é só o início, então eu já queria contar com a participação de vocês nas próximas etapas, entraremos em contato individualmente com vocês, com o Erivelton, com a Thaís, com o Augusto e outros que não puderam estar aqui hoje para combinar um pouco como que a gente pode caminhar para frente.

Felipe: Também queria agradecer a todos, a Thaís, o Augusto, que sempre ajudou a gente, e o Erivelton pela disponibilidade, e vocês pela organização também, e é exatamente isso. Obrigado, Patrícia.

Diego: Só pra gente falar aqui um pouco pela parte do Ministério da Cidadania. Assim, a reunião foi muito proveitosa, assim, de novas informações, de novos conteúdos que a gente precisa aprimorar nossos processos internos, que a gente também está trabalhando com essa questão da anonimização das bases. E a gente tem muito interesse em continuar, em prosseguir, em participar dos próximos encontros e reuniões para a gente ter acesso a esse conhecimento compartilhado e participar também, ajudar no que for possível. A base do Cadastro Único também é um desafio enorme em termos de a gente conseguir ter dados sintéticos, conseguir também retirar algumas características que possibilitam identificar algumas pessoas. É um desafio enorme que a gente está tendo que encarar aqui também.

Augusto Fadel: Só um comentário bem rápido Patrícia, o Diogo comentou, eu acho que antes a Laís e a Celi tinham falado de atributos sensíveis, isso é uma discussão muito difícil e dinâmica, não é? Porque em algum momento uma nova base surge, uma nova condição surge, uma nova legislação surge, o que não era sensível passa a ser e, enfim, quando a gente vai olhando, daqui a pouco, se você vai olhando em detalhe, daqui a pouco você descobre que tudo é sensível, você fala, não, mas eu tenho que liberar alguma coisa e você começa a tentar baixar o padrão, enfim. Uma coisa que eu vi muito interessante é que tem uma discussão internacional, uma parte do grupo internacional que discute esse assunto está começando a chamar o que antes era chamado de Privacy Preserving Techniques, passando a chamar de Privacy Enhancing Techniques. Ou seja, a gente não está garantindo que a gente

vai preservar a privacidade, a gente está aprimorando a privacidade, porque essa garantia de que não vai haver revelação, a única forma de fazer isso, e talvez não seja o suficiente, é não divulgar o microdado. E mesmo assim você pode ter vazamento de dados, enfim. Então é realmente, eu acho que esse sentimento é esse mesmo, a gente tem que lutar com isso. E só para fechar, obrigado mais uma vez, Patrícia, foi sensacional o evento e conta comigo aí se vocês precisarem.

Erivelton: Eu quero só agradecer aqui a participação de todos e foi um prazer estar aqui e fico à disposição aí naquilo que cada um achar que eu possa contribuir, mas uma coisinha que eu esqueci de mencionar. Eu mencionei, mas não destaquei, a gente já referenciou os endereços do CNPJ, do cadastro CNPJ, não é? Foi um trabalho grande, são 10 milhões de registros. A gente tem uma ferramenta aqui interna que permite fazer isso em grande quantidade. Se a gente fosse fazer no Google, na parte gratuita, ia levar meses, porque só pode 5, 10 mil por dia, uma coisa assim. Então, tem umas falhas, alguns não foram encontrados, mas se alguém quiser os endereços compartilhados, georreferenciados, a gente está disponível.

Patrícia: Quero agradecer muitíssimo a disponibilidade, a generosidade de vocês compartilharem o trabalho e dizer que a gente está comprometido, então nós vamos colocar, vamos nos debruçar sobre o que vocês trouxeram, vamos propor aqui um caminho que a gente possa seguir e vamos contatá-los, então, para gente dar continuidade nesse trabalho. Felipe, muito obrigada aí também do esforço que vocês fizeram aí no Ministério. Obrigada a todos e uma ótima semana para vocês.

Anexo II - Plano de desenvolvimento do trabalho e relação de variáveis

Plano de desenvolvimento do trabalho

Estágios		Ações	Input	Output (resultado esperado)
Estágio 0	**Onde aplicar SDC**	Identificar o pedido de cruzamento de bases mais relevante. Seja do ponto de vista de frequência ou de complexidade.	Solicitações dos usuários.	Escolha do caso a ser estudado e qual ano.
Estágio 1	**Por que aplicar SDC?**	Interpretação da lei. Dado a legislação vigente, é necessário aplicar algum nível aplicar alguma proteção da confidencialidade?	Análise dos tipos de variáveis e unidades estatísticas nos microdados.	Decisão se é necessário aplicar proteção.
Estágio 2	**Quais são as principais características e uso dos dados**	Compreender as necessidades dos usuários finais e as requisições para apropriada anomização. De modo a conseguir mapear o que é preciso para disseminar um arquivo em conformidade com os requisitos de anonimização, mas que seja útil para os usuários	Análise do questionário e metodologia da pesquisa.	Identificação da estrutura de dados, lista de identificadores, quase-identificadores, variáveis sensíveis, variáveis confidenciais.
			Análise das necessidades dos usuários.	Lista de prioridade na importância das variáveis a serem incluídas, nível de detalhe nas classificações, tipos de análises estatísticas realizadas pelos usuários e tipos de disseminação dos microdados do ponto de vista dos usuários.

Estágios		Ações	Input	Output (resultado esperado)
			Política de disseminação da instituto produtora da pesquisa e necessidade dos usuários.	Decisão sobre os tipos de disseminação.
			Análise do plano de disseminação e tipos de disseminação.	Restrições provindas do produtor dos dados.
Estágio 3	**Risco de revelação: definição e avaliação**	Definir o que é revelação e descrever quais situações podem fazer que isto aconteça (cenário de revelação) e mensurar qual é o risco de revelação.	Decisão sobre o tipo de microdados a ser disseminado (por exemplo, PUF ou MFR).	Definição de cenários de revelação e risco de revelação.
			Definição de risco, cenários de revelação e análise de dados.	Identificação de métodos para estimar ou mensurar o risco de revelação.
Estágio 4	**Métodos de controle de revelação**	Aplicação de métodos de proteção e mensuração da perda de informação.	Necessidades dos usuários, política de disseminação e análise dos tipos de variáveis requeridas.	Identificações de métodos de limitação de revelação a serem aplicados e escolha de parâmetro e limiares.
			Análise dos tipos de variáveis envolvidas, métodos de controle de revelação usados e necessidades dos usuários.	Identificações dos métodos de mensuração a perda de informação.
Estágio 5	**Implementação**	Implementação de todo o processo para gerar o arquivo final a ser disseminado.	Análise da disponibilidade de software ou estimação do esforço requirido para reescrever rotinas ou criar novas.	Escolha dos instrumentos necessários para produzir os microdados protegidos.

Estágios		Ações	Input	Output (resultado esperado)
			Microdados originais, método de estimação do risco de revelação, softwares/routines.	Risco de revelação dos microdados.
			Microdados originais, método de limitação de revelação, softwares/routines.	Dados protegidos.
			Microdados originais, microdados protegidos e métodos para mensurar a perda de informação.	Análise da perda de informação, revisão do arquivo final para certificar que não haja mais registros em risco.
			Descrição da metodologia da pesquisa.	Documentação da metodologia da pesquisa.
			Descrição dos métodos de SDC utilizados.	Documentação dos métodos SDC utilizados para usuários.

Relação de Variáveis

Dimensão das Bases

Base	Número de Variávies	Tamanho (em GB)	Extensão	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
Rais Estabelecimentos	29	2	dta				x	x	x	x	x	x	x	x	x		
RAIS Vínculos	77	18	z				x	x	x	x	x	x	x	x	x		
CAGED	41	7,5	txt		x	x	x	x	x	x	x	x	x	x			
SD	153	6,5	dta	x	x	x	x	x	x	x	x	x					
BEM - Parcelas	24	5,6	dta														
BEM – Acordos	22	1,9	dta														
Total	**765**																

Variáveis Empilhadas

Ref.	Variáveis	Base	Variávies_comuns
11	Bairro Requerente	SD	bairro
8	cbo2002ocupação	CAGED	ocupação
37	cboocupação2002	RAIS	ocupação
12	CEP Requerente	SD	cep
61	cepestab	RAIS	cep
2	cepestab	RAIS_Estab	cep
15	Classe CNAE 2.0	SD	cnae
38	cnae20classe	RAIS	cnae
28	cnae20classe	RAIS_Estab	cnae
39	cnae20subclasse	RAIS	cnae
34	cnpjcei	RAIS	cnpj
14	cnpjcei	CAGED	cnpj
4	cnpjcei	RAIS_Estab	cnpj
35	cnpjraiz	RAIS	cnpj
13	cnpjraiz	CAGED	cnpj
5	cnpjraiz	RAIS_Estab	cnpj
28	Cód.Classe CNAE 2.0	SD	cnae
29	Cód.Divisão CNAE 2.0	SD	cnae
17	Cod.Gênero	SD	sexo
32	Cód.Grau Instrução	SD	escolaridade
33	Cód.Grupo CNAE 2.0	SD	cnae
39	Cód.Município Residência	SD	Munic_trabalhador
41	Cód.Ocupação CBO	SD	ocupação
44	Cód.Seção CNAE 2.0	SD	cnae
45	Cód.Subclasse CNAE 2.0	SD	cnae
32	cpf	RAIS	cpf
9	cpf	CAGED	cpf
13	CPF Requerente	SD	cpf
1	cpfrequerente	BEM – Acordos	cpf
51	Data Admissão Requerente	SD	data_admissao
54	Data Nascimento	SD	data_nascimento
32	dataadmissão	CAGED	data_admissao
20	dataadmissãodeclarada	RAIS	data_admissao
30	datadenascimento	RAIS	data_nascimento
30	datanascimento	CAGED	data_nascimento
19	datanascimento	BEM – Acordos	data_nascimento

Ref.	Variáveis	Base	Variávies_comuns
60	Divisão CNAE 2.0	SD	cnae
9	escolaridadeapós2005	RAIS	escolaridade
63	Faixa Salarial	SD	Fx_Salar/Renda
72	Gênero	SD	sexo
69	Grau Instrução	SD	escolaridade
11	graudeinstrução	CAGED	escolaridade
71	Grupo CNAE 2.0	SD	cnae
57	idade	RAIS	idade
12	idade	CAGED	idade
75	indtrabintermitente	RAIS	intermitente
23	indtrabintermitente	CAGED	intermitente
1	município	RAIS	Munic_trabalhador
4	município	CAGED	Munic_trabalhador
93	Município Residência	SD	Munic_trabalhador
62	muntrab	RAIS	Munic_trabalhador
29	nit	CAGED	nit
99	Num Insc Empregador (CEI/CNPJ)	SD	cnpj
104	Ocupação CBO	SD	ocupação
38	paísdenacionalidade	CAGED	origem
39	paísdeorigem	CAGED	origem
29	pis	RAIS	pis
105	PIS/PASEP/NIT	SD	nit
12	raçacor	RAIS	racacor
17	raçacor	CAGED	racacor
121	Razão Social Empregador	SD	razaosocial
63	razãosocial	RAIS	razaosocial
25	salário	CAGED	salar/rendimento
126	Seção CNAE 2.0	SD	cnae
18	sexo	CAGED	sexo
10	sexotrabalhador	RAIS	sexo
134	Subclasse CNAE 2.0	SD	cnae
25	tempoemprego	RAIS	tempoemprego
16	tempoemprego	CAGED	tempoemprego
77	uf	RAIS	uf
3	uf	CAGED	uf
153	Último Salário	SD	salar/rendimento
41	valorsalariofixo	CAGED	salar/rendimento

Anexo III – Planilha de Controle solicitação de bases identificadas e justificativas

Tipo	Bases Solicitadas	Finalidade	Grupo de Finalidade
Conselhos Profissionais	RAIS e CAGED	integração dos dados da RAIS e CAGED com os dados do sindicato	Fiscalização
Conselhos Profissionais	RAIS	fiscalizar os economistas no Brasil; e acompanhar o profissional no mercado de trabalho, verificando as mudanças que ocorrem no mundo do trabalho e como elas estão influenciando a inclusão do profissional	Fiscalização
Conselhos Profissionais		fiscalização do exercício ilegal da profissão de Administrador	Fiscalização
Conselhos Profissionais	RAIS	Fiscalização voltada especificamente aos profissionais que exercem atividades inerentes ao campo profissional da Administração e seus respectivos empregador.	Fiscalização
Conselhos Profissionais	RAIS e CAGED	Fiscalização do exercício profissional da Administração, em atendimento à manutenção, organização e execução da inspeção do trabalho, conforme art. 21, XXIV, Constituição Federal e artigo 8º, alínea “b”, da Lei 4.769/65.	Fiscalização
Conselhos Profissionais	RAIS e CAGED	Fiscalização do exercíccio das profissões de Engenheria e agronomia	Fiscalização
Conselhos Profissionais	RAIS	Fiscalização do exercício da Profissão de Administrador no estado	Fiscalização
Conselhos Profissionais	CAGED	relação de biólogos relacionados no CAGED nos estados de Pernambuco, Ceará, Maranhão, Paraíba, Piauí e Rio Grande do Norte.	Fiscalização
Conselhos Profissionais	RAIS	Fiscalização do exercício da Profissão de Administrador no estado	Fiscalização
Conselhos Profissionais	RAIS e CAGED	mapear o local de atuação de trabalho dos economistas com finalidade de regularizar e atualizar o cadastro dos profissionais do Conselho.	Fiscalização
Conselhos Profissionais	RAIS	Fiscalização do exercício das profissões contábil, economia e administração.	Fiscalização
Empresa Pública	RAIS e CAGED	aprimorar a qualidade e eficiência das atividades de controles internos e auditoria pelo SERPRO	Política Pública

Tipo	Bases Solicitadas	Finalidade	Grupo de Finalidade
Empresa Pública	RAIS e CAGED	Uso da RAIS para verificação de benefícios sobre atribuição da CAIXA	Política Pública
Entidade Privada	RAIS e CAGED	realizar uma análise econométrica do impacto dos incentivos fiscais dos estados brasileiros	Pesquisa
Entidade Privada	RAIS e CAGED	Realizar estudo sobre a dinâmica de formação e emprego dos indivíduos com titulações de mestrado e/ou doutorado	Pesquisa
Entidade Privada	RAIS	investigar os movimentos de médio e longo prazo no mercado de trabalho dos beneficiários de programas de transferência de renda e avaliar de maneira mais agregada a relação dos ciclos do mercado de trabalho com os ciclos dos programas de transferência.	Pesquisa
Entidade Privada	RAIS	analisar a relação entre a variação na quantidade de empregos e tipo de vínculo empregatício ao longo do tempo, por região demográfica.	Pesquisa
Entidade Privada		Aplicação nos modelos estatísticos de riscos de crédito dos clientes do Banco Santander.	Pesquisa
Entidade Privada	RAIS e CAGED	Desenvolver estudos e pesquisas para mapeamento da realidade de remuneração em âmbito Nacional.	Pesquisa
Entidade Privada	RAIS e CAGED	Realizar estudos estatísticos, em respeito à Lei nº 12.527, de 18 de novembro de 2011.	Pesquisa
Entidades Sindicais	RAIS e CAGED	Prevenir e combater a fraude contra o seguro, agindo assim a favor dos consumidores e beneficiando a sociedade em geral.	Fiscalização
Entidades Sindicais	RAIS	Manter controle das informações dos trabalhadores representados	Fiscalização
Entidades Sindicais	RAIS e CAGED	Defesa comum dos interesses da categoria profissional que representa.	Fiscalização
Entidades Sindicais	RAIS e CAGED	Realizar o mapeamento real do porte das empresas que compõem a base de representação e assim formular ações de assistências a elas dirigidas..	Fiscalização
Entidades Sindicais	RAIS e CAGED	Realizar os cálculos de liquidação na ação de cumprimento nº 1000269-35.2018.5.02.0317, no qual tramita perante a 7ª Vara do Trabalho de Guarulhos/SP	Fiscalização
Entidades Sindicais	RAIS e CAGED	Manter controle das informações dos trabalhadores representados	Fiscalização

Tipo	Bases Solicitadas	Finalidade	Grupo de Finalidade
Entidades Sindicais	RAIS e CAGED	Acompanhar a evolução da categoria na área de representação do sindicato nos municípos, através de estudos e pesquisas.	Fiscalização
Entidades Sindicais	RAIS e CAGED	Manter controle das informações dos trabalhadores representados	Fiscalização
Entidades Sindicais	RAIS e CAGED	Fiscalização do cumprimento das normas coletivas estampadas na Consolidação das leis Trabalhistas, bem como nas Convenções Coletivas de Trabalho.	Fiscalização
Entidades Sindicais	RAIS e CAGED	requer informações sobre os trabalhadores que fazem parte do sindicato.	Fiscalização
Entidades Sindicais	RAIS e CAGED	Verificar se os empregadores rurais descontaram e recolheram a devida contribuição sindical rural dos empregados na base territorial da entidade sindical supramencionada.	Fiscalização
Entidades Sindicais	CAGED	Conferir maior segurança jurídica e transparência às contratações de planos de saúde coletivos empresariais, impedindo que pessoas sem qualquer vínculo com as entidades contratantes sejam beneficiadas.	Fiscalização
Entidades Sindicais	RAIS e CAGED	estudo, coordenação, proteção, orientação e representação legal, em nível nacional, dos trabalhadores do Grupo “Turismo e Hospitalidade”.	Fiscalização
Entidades Sindicais	RAIS e CAGED	Substidiar trabalho de assistência à categoria representada	Fiscalização
Entidades Sindicais	RAIS e CAGED	Manter controle das informações dos trabalhadores representados	Fiscalização
Entidades Sindicais	RAIS e CAGED	Realizar o cruzamento de dados quantitativos das empresas representadas pelo Sindmóveis visando acompanhar as informações prestadas pelas mesmas para verificar a aplicabilidade das convenções coletivas de trabalho	Fiscalização
Entidades Sindicais	RAIS e CAGED	utilizar as informações para dar eficácia ao direito dos trabalhadores representados à percepção do adicional de insalubridade, pelo grau máximo, conforme pleiteado nas ações civis pública relacionadas na solicitação.	Fiscalização

Tipo	Bases Solicitadas	Finalidade	Grupo de Finalidade
Entidades Sindicais	RAIS e CAGED	utilização em suas atividades institucionais.	Fiscalização
Entidades Sindicais	CAGED	análise e possível atendimento á solicitação objeto do mencionado ofício.	Fiscalização
Entidades Sindicais	RAIS e CAGED	acompanhamento dos seus representados em sua base territorial, mapeamento do número de empregados, sendo eles o que estão empregados, os admitidos e demitidos, a fim de desenvolver programa para manutenção de emprego e para contratação.	Fiscalização
Entidades Sindicais	RAIS e CAGED	realizar o mapeamento estatístico, traçar o perfil e traçar o perfil dos salários de todos os colaboradores da Rede Apae no Brasil.	Pesquisa
Entidades Sindicais	RAIS e CAGED	Produzir estatísticas inerentes ao mercado de trabalho do setor de prestação de serviços de limpeza e conservação na construção de projeto de business intelligence para acompanhar a evolução do mercado de trabalho do setor.	Pesquisa
Entidades Sindicais	RAIS e CAGED	Estudos e pesquisas de interesse dos entes e instituições vinculadas ao Sistema FIEC.	Pesquisa
Entidades Sindicais	RAIS e CAGED	obter os dados do trabalhadores integrantes da categoria, viabilizando a atuação da entidade sindical na fiscalização e cumprimento das normas trabalhistas, acordos coletivos e convenção coletiva de trabalho.	Pesquisa
Entidades Sindicais	RAIS e CAGED	conhecer o universo de trabalhadores atingidos pela pandemia de alguma forma, com desemprego ou redução salarial, para propiciar um estudo para viabilizar o incentivo a recolocação profissional e adoção de medidas que possam dar assistência aos trabalhadores e as empresas, bem como prestar orientação na prática de políticas públicas	Pesquisa
Organismos Internacionais		Apoiar questões de implementação e monitorar os efeitos das medidas em nível de emergência; (2) avaliar os impactos dessas políticas e extrair recomendações de políticas públicas para estratégias de médio prazo no Brasil, ou mesmo em outros países da	Pesquisa

Tipo	Bases Solicitadas	Finalidade	Grupo de Finalidade
		região que tenham cenários semelhantes.	
Órgãos Públicos Municipais e Estaduais	RAIS e CAGED	auxiliar nas medidas judiciais e administrativas para a recuperação de ativos de titularidade do Estado do Rio Grande do Sul e combate às fraudes fiscais	Fiscalização
Órgãos Públicos Municipais e Estaduais	CAGED	facilitar e dar maior eficiência às ações do Ministério Público de Santa Catarina nas mais variadas áreas de atuação, além de auxiliar na tomada de decisões dos órgãos de execução.	Fiscalização
Órgãos Públicos Municipais e Estaduais	RAIS e CAGED	incrementar as bases disponíveis ao órgão para o exercício de sua atribuição quanto ao controle externo da Administração Pública	Fiscalização
Órgãos Públicos Municipais e Estaduais	RAIS e CAGED	Auxiliar na atividade ministerial, visando a consecução de seus objetivos institucionais.	Fiscalização
Órgãos Públicos Municipais e Estaduais	RAIS e CAGED	auxiliar na instrução procedimental e processual dos feitos afetos ao MPPE	Fiscalização
Órgãos Públicos Municipais e Estaduais	RAIS e CAGED	Verificação da regularidade na concessão de benefícios assistenciais e previdenciários, na prestação de serviços de saúde e no custeio da folha de pagamento.	Fiscalização
Órgãos Públicos Municipais e Estaduais	RAIS e CAGED	cumprimento dos compromissos firmados pelos contribuintes em protocolos de intenções e regimes especiais de tributação; dar maior segurança ao crédito tributário tramitando na esfera administrativa; e utilizar a evolução do emprego e renda como indicadores para o planejamento e política tributária e fiscal	Fiscalização
Órgãos Públicos Municipais e Estaduais	RAIS e CAGED	subsidiar a atuação investigativa e produção de conhecimento de inteligência do Ministério Público Estadual, inclusive com a criação de sistemas big data de cruzamento de dados.	Fiscalização
Órgãos Públicos Municipais e Estaduais	RAIS e CAGED	utilizar no exercício das atividades de controle externo que envolvem a avaliação de políticas e a correta aplicação dos recursos públicos	Fiscalização

Tipo	Bases Solicitadas	Finalidade	Grupo de Finalidade
Órgãos Públicos Municipais e Estaduais	RAIS e CAGED	Incremento de bases de dados para combate à sonegação fiscal e às fraudes	Fiscalização
Órgãos Públicos Municipais e Estaduais	RAIS e CAGED	Implantação de dados atualizados no Sistema Busca Integrada de Dados-BID para acesso a todos os membros do Ministério Público do Estado do Piauí.	Fiscalização
Órgãos Públicos Municipais e Estaduais	RAIS e CAGED	Otimizar a atuação do MPCE na defesa dos interesses sociais e individuais	Fiscalização
Órgãos Públicos Municipais e Estaduais	RAIS e CAGED	Apoiar suas atividades institucionais de supervisão e coordenação do sistema de controle interno na apuração de indícios de irregularidades administrativas; de correição administrativa, nas apurações de responsabilidades no âmbito do Poder Executivo do Distrito Federal, por meio de Procedimento Investigatório e Disciplinares; além da defesa do patrimônio público, da transparência e do Combate à Corrupção.	Fiscalização
Órgãos Públicos Municipais e Estaduais	RAIS e CAGED	Atender ao interesse público e sua persecução, com o objetivo de executar as competências legais atribuídas aos Tribunais de Contas em seu papel precípuo de órgão controlador externo.	Fiscalização
Órgãos Públicos Municipais e Estaduais	RAIS e CAGED	Facilitar a coleta de dados necessários para subsidiar atividades relacionadas a processos judiciais e extrajudiciais no âmbito da Procuradoria-Geral do Estado de Goiás.	Fiscalização
Órgãos Públicos Municipais e Estaduais	RAIS e CAGED	Atividades do sistema de controle interno	Fiscalização
Órgãos Públicos Municipais e Estaduais	RAIS e CAGED	aprimorar sua atividade fiscalizatória.	Fiscalização
Órgãos Públicos Municipais e Estaduais	RAIS e CAGED	Realizar atividades de auditoria e fiscalização nos sistemas contábil, financeiro, orçamentário, patrimonial, de pessoal e de recursos externos	Fiscalização

Tipo	Bases Solicitadas	Finalidade	Grupo de Finalidade
Órgãos Públicos Municipais e Estaduais	RAIS e CAGED	Realizar o cruzamento das informações com o sistema de folha de pagamento do Estado de Pernambuco a fim de implantar controles internos preventivos e mitigadores de erros e fraudes.	Fiscalização
Órgãos Públicos Municipais e Estaduais	RAIS e CAGED	realizar estudos e avaliações dos programas de crédito do BNB.	Pesquisa
Órgãos Públicos Municipais e Estaduais	RAIS e CAGED	realização de pesquisa e certificação/validação de dados e informações cadastrais de funcionários que mantem ou mantiveram vínculo com os Entes desse Governo	Pesquisa
Órgãos Públicos Municipais e Estaduais	RAIS e CAGED	Elaborar estudos de caráter preditivo e analítico, a fim de mitigar riscos de fraudes e conflitos de interesse nas contratações públicas	Pesquisa
Órgãos Públicos Municipais e Estaduais	RAIS e CAGED	Responder questões relativas ao emprego na cidade de São José dos Campos	Pesquisa
Órgãos Públicos Municipais e Estaduais	RAIS e CAGED	Produzir conhecimento e subsidiar políticas públicas através da elaboração e implementação de estudos e pesquisas econômicas, voltadas ao desenvolvimento socioeconômico do Espirito Santo.	Pesquisa
Órgãos Públicos Municipais e Estaduais		realizar análise quanto ao cumprimento de requisitos para a fruição de benecios fiscais.	Pesquisa
Órgãos Públicos Municipais e Estaduais		Estudos e pesquisas, acompanhamento e avaliação de planos, programas e projetos, para manter o sistema de informações e carttografia.	Pesquisa
Órgãos Públicos Municipais e Estaduais	RAIS	elaborar estudos e projetos, bem como prover a base e difundir as informações estatísticas e geográficas, visando à formulação e avaliação de políticas públicas, planos e programas de desenvolvimento do estado	Pesquisa
Órgãos Públicos Municipais e Estaduais	RAIS e CAGED	analisar a concentração espacial e a especialização econômica pelo território de Diadema.	Pesquisa

Tipo	Bases Solicitadas	Finalidade	Grupo de Finalidade
Órgãos Públicos Municipais e Estaduais	RAIS	ampliar a qualidade da informação ofertada oa governo do estado para a tomada de decisão a respeito das políticas públicas.	Política pública
Órgãos Públicos Municipais e Estaduais	RAIS e CAGED	Avançar da disponibilidade de indicadores contextualizados com informações extraídas diretamente dos microdados disponíveis no portal do Ministério da Economia.	Política pública
Órgãos Públicos Municipais e Estaduais	RAIS	Avaliar o impacto dos programas de renúncia fiscal na geração de emprego, renda e arrecadação tributária local no estado do Goiás	Política pública
Órgãos Públicos Municipais e Estaduais	CAGED e RAIS	Atuação judicial e extrajudicial visando à recuperação de ativos inscritos na dívida ativa do Estado de São Paulo.	Política pública
Órgãos Públicos Municipais e Estaduais	RAIS e CAGED	Servir de subsídio informacional para elaboração de políticas públicas de trabalho, emprego e geração de renda.	Política pública
Órgãos Públicos Municipais e Estaduais	CAGED e RAIS	para realizar política pública de acompanhamento e desenvolvimento econômico no Estado do RS	Política pública
Órgãos Públicos Municipais e Estaduais	RAIS e CAGED	Atividade de investigação de polícia	Política pública
Órgãos Públicos Municipais e Estaduais	RAIS	compor a base de cálculo do PIB turístico de Santa Catarina.	Política pública
Órgãos Públicos Municipais e Estaduais	RAIS e CAGED	auxiliar e contribuir na aplicação de Políticas Públicas do Estado de São Paulo, preliminarmente do Programa Bolsa do Povo	Política Pública
Órgãos Públicos Municipais e Estaduais	RAIS	aprimorar a base de informações e indicadores de ciência e tecnologia do estado de São Paulo.	Política pública
Órgãos Públicos Municipais e Estaduais	RAIS e CAGED	realizar o cruzamento de informações dos contribuintes com domicílio na cidade de Goiânia a fim de verificar se cumprem os requisitos para obtenção de auxílios instituídos pelo Renda Família e IPTU Social.	Política pública
Órgãos Públicos		desenvolver uma análise espacial da concentração de empregos formais dos municípios do Amazonas, a fim de	Política pública

Tipo	Bases Solicitadas	Finalidade	Grupo de Finalidade
Municipais e Estaduais		melhor projetar a logística de locomoção da população economicamente ativa.	
Órgãos Públicos Municipais e Estaduais	RAIS	viabilizar a avaliação de impacto de políticas públicas do Estado do Rio Grande do Sul.	Política pública
Órgãos Públicos Municipais e Estaduais	RAIS e CAGED	formular políticas públicas baseadas na observação empírica das variáveis relacionadas a estrutura produtiva do município de Campinas e ao seu mercado de trabalho.	Política pública
Órgãos Públicos Municipais e Estaduais	RAIS e CAGED	Fiscalização com objetivo de aumentar a arrecadação do Imposto sobre Serviços de Qualquer Natureza (ISSQN), identificando a falta de seu recolhimento, principalmente, relacionada à fraude de domicílio tributário.	Política pública
Órgãos Públicos Municipais e Estaduais	RAIS e CAGED	Aprimoramento da atuação institucional desta Agência.	Política Pública
Sistema S	RAIS e CAGED	ampliar as análises agregadas circunscritas à missão institucional do SENAI/PR, a partir da combinação das bases RAIS e CAGED com outras bases de dados mantidas pelo SENAI/PR	Pesquisa
Sistema S	RAIS	produzir estatísticas e estudos sobre o mercado de trabalho que contribuam para o aprimoramento de políticas voltadas para a promoção de um ambiente favorável aos negócios, à competitividade e ao desenvolvimento do Brasil.	Pesquisa
Sistema S	RAIS	identificar os egressos concluintes de cursos do Senac, avaliando os perfis de suas ocupações, atividades econômicas, regiões geográficas, assim como das características dos cursos que realizaram.	Pesquisa
Sistema S	RAIS e CAGED	Realizar estudos sobre a dinâmica do emprego nas micro e pequenas empresas do Brasil, compreendendo as movimentações dos indivíduos presentes no mercado formal que se tornam empresários, bem como, empresários que fecham suas empresas e retornam ao mercado de trabalho formal.	Pesquisa

Tipo	Bases Solicitadas	Finalidade	Grupo de Finalidade
Sistema S		Realizar estudo sobre migração interestadual de mão de obra qualificada no emprego formal; acompanhar a inserção no mercado de trabalho formal de estudantes egressos do Senai do Espírito Santo, com objetivo de compreender a participação deles no mercado de trabalho capixaba; acessar o universo das empresas do Espírito Santo a fim de estabelecer amostras estatísticas para as pesquisas estaduais.	Pesquisa
Universidades e Instituições de Ensino e Pesquisa (Exceto Federais)	RAIS	Projetos de pesquisa descritos de docentes, discentes e pesquisadores	Pesquisa
Universidades e Instituições de Ensino e Pesquisa (Exceto Federais)	RAIS e CAGED	desenvolver pesquisas e relatórios técnicos-científicos desidentificados sobre mercado de trabalho e a migração interna e internacional no brasil a partir de informações sobre a inserção de trabalhadores – migrantes ou não - no setor formal brasileiro	Pesquisa
Universidades e Instituições de Ensino e Pesquisa (Exceto Federais)	RAIS e CAGED	aprimorar as análises dos docentes e discentes, buscando cumprir a missão institucional desta Universidade de produção de conhecimento científico.	Pesquisa
Universidades e Instituições de Ensino e Pesquisa (Exceto Federais)	RAIS e CAGED	realizar pesquisa acadêmica na forma de artigos, monografias de graduação, dissertações de mestrado e teses de doutorado sobre vários aspectos do mercado de trabalho no Brasil	Pesquisa
Universidades e Instituições de Ensino e Pesquisa (Exceto Federais)	RAIS	acessar as informações de empregabilidade dos egressos constantes na base da RAIS	Pesquisa
Universidades e Instituições de Ensino e Pesquisa (Exceto Federais)	RAIS	Realização de pesquisas, estudos e projetos das atividades comerciais e de outros setores	Pesquisa
Universidades e Instituições de Ensino e Pesquisa (Exceto Federais)	RAIS e CAGED	Realizar pesquisas acadêmicas sobre o mercado de trabalho	Pesquisa

Tipo	Bases Solicitadas	Finalidade	Grupo de Finalidade
Universidades e Instituições de Ensino e Pesquisa (Exceto Federais)	RAIS e CAGED	Realizar pesquisa acadêmica sobre práticas nepotistas	Pesquisa
Universidades e Instituições de Ensino e Pesquisa (Exceto Federais)	RAIS	avaliar o impacto econômico das Universidades Públicas do Paraná no mercado de trabalho	Pesquisa
Universidades e Instituições de Ensino e Pesquisa (Exceto Federais)	RAIS	Realizar pesquisa sobre micro e pequenas empresas	Pesquisa
Universidades e Instituições de Ensino e Pesquisa (Exceto Federais)	RAIS e CAGED	conhecer o número de egressos desta Instituição de Ensino Superior (IES) que estão inseridos no mundo do trabalho.	Pesquisa
Universidades e Instituições de Ensino e Pesquisa (Exceto Federais)	RAIS	Pesquisas dos mestrandos e doutorando, bem como nos projetos de pesquisa desenvolvidos pelos professores desses Programa.	Pesquisa

Anexo IV – Planilha de dúvidas sobre as bases e variáveis

RAIS

variavel, observacao,duvida,inconsistencia,resposta
municÃpio,,nÃ£o,nÃ£o,
vÃnculoativo3112,,nÃ£o,nÃ£o,
tipovÃnculo,,nÃ£o,nÃ£o,
motivodesligamento,,nÃ£o,nÃ£o,
mÃªsdesligamento,,nÃ£o,nÃ£o,
indvÃnculoalvarÃ¡,,nÃ£o,nÃ£o,
tipoadmissÃ£o,,nÃ£o,nÃ£o,
tiposalÃ¡rio,,nÃ£o,nÃ£o,
escolaridadeapÃ³s2005,,nÃ£o,nÃ£o,
sexotrabalhador,,nÃ£o,nÃ£o,
nacionalidade,,nÃ£o,nÃ£o,
raÃ§acor,,nÃ£o,nÃ£o,
indportadordefic,,nÃ£o,nÃ£o,
tamanhoestabelecimento,,nÃ£o,nÃ£o,
naturezajurÃdica,"Checar o que ocorre qdo natureza jurÃdica for 2135 (firma mercantil individual), 4014 (Empresa Indiv
RAIS2007)",nÃ£o,nÃ£o,
indceivinculado,,nÃ£o,nÃ£o,
tipoestab,"Esclarecer: CNPJ ou CEI? | NÃ£o bate com o dicionÃ¡rio (1, 3, 9 X 1, 3, 5 na base)",sim,nÃ£o,
indestabparticipapat,,nÃ£o,nÃ£o,
indsimples,,nÃ£o,nÃ£o,
dataadmissÃ£odeclarada,NÃ£o pode ser sÃ³ mÃªs/ano?,sim,nÃ£o,
vlremunmÃ©dianom,,nÃ£o,nÃ£o,
vlremunmÃ©diasm,,nÃ£o,nÃ£o,
vlremundezembronom,,nÃ£o,nÃ£o,
vlremundezembrosm,,nÃ£o,nÃ£o,
tempoemprego,Tempo em meses?,sim,nÃ£o,
qtdhoracontr,,nÃ£o,nÃ£o,
vlÃºltimaremuneraÃ§Ã£oano,,nÃ£o,nÃ£o,
vlsalÃ¡riocontratual,,nÃ£o,nÃ£o,
pis,,nÃ£o,nÃ£o,
datadenascimento,,nÃ£o,nÃ£o,
nÃºmeroctps,,nÃ£o,nÃ£o,
cpf,,nÃ£o,nÃ£o,
ceivinculado,Esclarecer: CNPJ ou CEI?,sim,nÃ£o,
cnpjcei,,nÃ£o,nÃ£o,
cnpjraiz,,nÃ£o,nÃ£o,
nometrabalhador,,nÃ£o,nÃ£o,
cboocupaÃ§Ã£o2002,,nÃ£o,nÃ£o,
cnae20classe,,nÃ£o,nÃ£o,
cnae20subclasse,"SÃ³ serÃ¡ um problema, se for com as MEIs*",nÃ£o,nÃ£o,
tipodefic,,nÃ£o,nÃ£o,
causaafastamento1,"Esse dados sobre afastamento precisam estar desagregados desssa forma, com essa granularida
diainiaf1,Idem acima,sim,nÃ£o,

mÃªsiniaf1,Idem acima,sim,nÃ£o,
diafimaf1,Idem acima,sim,nÃ£o,
mÃªsfimaf1,Idem acima,sim,nÃ£o,
causaafastamento2,Idem acima,sim,nÃ£o,
diainiaf2,Idem acima,sim,nÃ£o,
mÃªsiniaf2,Idem acima,sim,nÃ£o,
diafimaf2,Idem acima,sim,nÃ£o,
mÃªsfimaf2,Idem acima,sim,nÃ£o,
causaafastamento3,Idem acima,sim,nÃ£o,
diainiaf3,Idem acima,sim,nÃ£o,
mÃªsiniaf3,Idem acima,sim,nÃ£o,
diafimaf3,Idem acima,sim,nÃ£o,
mÃªsfimaf3,Idem acima,sim,nÃ£o,
qtddiasafastamento,Idem acima,sim,nÃ£o,
idade,,nÃ£o,nÃ£o,
diadedesligamento,,nÃ£o,nÃ£o,
ibgesubsetor,,nÃ£o,nÃ£o,
anochegadabrasil,,nÃ£o,nÃ£o,
cepestab,,nÃ£o,nÃ£o,
muntrab,MunicÃpio onde pessoa de fato trabalha?,sim,nÃ£o,
razÃ£osocial,Como tratar razÃ£o social de MEI na regra antiga? Como achar MEIs na base?,sim,nÃ£o,
vlremjaneirosc,O que significa sc do fim da variÃ¡vel? | Ã‰ necessÃ¡ria esta granularidade do dado?,sim,nÃ£o,
vlremfevereirosc,Idem acima,sim,nÃ£o,
vlremmarÃ§osc,Idem acima,sim,nÃ£o,
vlremabrilsc,Idem acima,sim,nÃ£o,
vlremmaiosc,Idem acima,sim,nÃ£o,
vlremjunhosc,Idem acima,sim,nÃ£o,
vlremjulhosc,Idem acima,sim,nÃ£o,
vlremagostosc,Idem acima,sim,nÃ£o,
vlremsetembrosc,Idem acima,sim,nÃ£o,
vlremoutubrosc,Idem acima,sim,nÃ£o,
vlremnovembrosc,Idem acima,sim,nÃ£o,
indtrabintermitente,,nÃ£o,nÃ£o,
indtrabparcial,,nÃ£o,nÃ£o,
uf,Foi um marcador da seleÃ§Ã£o desta amostra?,sim,nÃ£o,
munic_selec,Foi um marcador da seleÃ§Ã£o desta amostra?,sim,nÃ£o,

CAGED

variavel,observacao,duvida,inconsistencia,resposta
V1,id da linha,nÃ£o,nÃ£o,
competÃªnciamov,,nÃ£o,nÃ£o,
regiÃ£o,,nÃ£o,nÃ£o,
uf,,nÃ£o,nÃ£o,
municÃpio,HÃ¡ ocorrÃªncias Ãºnicas de municÃpios combinadas com subclasse,nÃ£o,nÃ£o,
seÃ§Ã£o,,nÃ£o,nÃ£o,
subclasse,,nÃ£o,nÃ£o,

saldomovimentaÃ§Ã£o,O resumo entre entradas (+1) e saÃdas (-1),nÃ£o,nÃ£o,
cbo2002ocupaÃ§Ã£o,,nÃ£o,nÃ£o,
cpf,,nÃ£o,nÃ£o,
categoria,,nÃ£o,nÃ£o,
graudeinstruÃ§Ã£o,,nÃ£o,nÃ£o,
idade,,nÃ£o,nÃ£o,
cnpjraiz,,nÃ£o,nÃ£o,
cnpjcei,"Significado e composiÃ§Ã£o da variÃ¡vel?
CNPJ Ã© para pessoa jurÃdica, CEI para pessoa fÃsica, cnpjcei Ã© para construÃ§Ã£o civil
Nesta base parece ser um ""CNPJ completo""",sim,nÃ£o,
horascontratuais,,nÃ£o,nÃ£o,
tempoemprego,Tempo em meses. O range da base indica que hÃ¡ inconsistÃªncias.,nÃ£o,sim,
raÃ§acor,,nÃ£o,nÃ£o,
sexo,,nÃ£o,nÃ£o,
tipoempregador,Onde estariam os CFPs caso o tipoempregador seja == 2?,sim,nÃ£o,
tipoestabelecimento,Onde estariam os CFPs caso o tipoestabelecimento seja == 3 ou ==5?,sim,nÃ£o,
tipomovimentaÃ§Ã£o,,nÃ£o,nÃ£o,
tipodedeficiÃªncia,,nÃ£o,nÃ£o,
indtrabintermitente,,nÃ£o,nÃ£o,
indtrabparcial,,nÃ£o,nÃ£o,
salÃ¡rio,,nÃ£o,nÃ£o,
tamestabjan,,nÃ£o,nÃ£o,
indicadoraprendiz,,nÃ£o,nÃ£o,
origemdainformaÃ§Ã£o,,nÃ£o,nÃ£o,
nit,,nÃ£o,nÃ£o,
datanascimento,,nÃ£o,nÃ£o,
diamovimentaÃ§Ã£o,,nÃ£o,nÃ£o,
dataadmissÃ£o,,nÃ£o,nÃ£o,
competÃªnciadec,,nÃ£o,nÃ£o,
indicadordeforadoprazo,,nÃ£o,nÃ£o,
paÃsdenacionalidade,,nÃ£o,nÃ£o,
paÃsdeorigem,,nÃ£o,nÃ£o,
unidadesalÃ¡riocÃ³digo,,nÃ£o,nÃ£o,
valorsalÃ¡riofixo,,nÃ£o,nÃ£o,
regiao,,nÃ£o,nÃ£o,

BEM

variavel,observacao,duvida,inconsistencia,resposta
AgÃªncia BancÃ¡ria,,nÃ£o,nÃ£o,
AgÃªncia BancÃ¡ria DV,Essa informaÃ§Ã£o nÃ£o precisa ser divulgada?,nÃ£o,nÃ£o,
Banco,,nÃ£o,nÃ£o,
CPF Requerente,,nÃ£o,nÃ£o,
CPF UsuÃ¡rio Portal,,nÃ£o,nÃ£o,
Classe CNAE 2.0,deve ser igual Ã cnae20classe da base da RAIS,nÃ£o,nÃ£o,
Compet. Acordo,Referente ao presente ou ao passado? Qual passado?,sim,nÃ£o,

Compet. AntepenÃºltimo SalÃ¡rio,,nÃ£o,nÃ£o,
Compet. AtualizaÃ§Ã£o Status Req.,,nÃ£o,nÃ£o,
Compet. Cancelamento Acordo,,nÃ£o,nÃ£o,
Compet. DigitaÃ§Ã£o Acordo,,nÃ£o,nÃ£o,
Compet. FinalizaÃ§Ã£o Acordo,,nÃ£o,nÃ£o,
Compet. InclusÃ£o Acordo,,nÃ£o,nÃ£o,
Compet. PenÃºltimo SalÃ¡rio,,nÃ£o,nÃ£o,
Compet. Recebimento Acordo,,nÃ£o,nÃ£o,
Compet. Requerente BeneficiÃ¡rio,,nÃ£o,nÃ£o,
Compet. Requerente Segurado,,nÃ£o,nÃ£o,
Compet. Requerimento,,nÃ£o,nÃ£o,
Compet. Ãšltimo SalÃ¡rio,,nÃ£o,nÃ£o,
CÃ³d. Classe Cnae 2.0,deve ser igual Ã cnae20classe da base da RAIS,nÃ£o,nÃ£o,
CÃ³d. Cnae 2.0,,nÃ£o,nÃ£o,
CÃ³d. Faixa EtÃ¡ria,8 nÃveis,nÃ£o,nÃ£o,
CÃ³d. Faixa Salarial BEm,5 nÃveis,nÃ£o,nÃ£o,
CÃ³d. Faixa Salarial MTE,12 nÃveis,nÃ£o,nÃ£o,
CÃ³d. Faixa Salarial SD,8 nÃveis,nÃ£o,nÃ£o,
CÃ³d. Faixa Tempo Acordo,5 nÃveis,nÃ£o,nÃ£o,
CÃ³d. MunicÃpio Empresa,,nÃ£o,nÃ£o,
CÃ³d. Programa,,nÃ£o,nÃ£o,
CÃ³d. SeÃ§Ã£o Cnae 2.0,,nÃ£o,nÃ£o,
CÃ³d. Tipo Empregador,Qual Ã© o significado?,sim,nÃ£o,
Data Acordo,,nÃ£o,nÃ£o,
Data AdmissÃ£o,datas estranhas: 01/11/1199,nÃ£o,sim,
Data AtualizaÃ§Ã£o Status Req.,,nÃ£o,nÃ£o,
Data Cancelamento Acordo,,nÃ£o,nÃ£o,
Data DigitaÃ§Ã£o Acordo,,nÃ£o,nÃ£o,
Data Final Informada,,nÃ£o,nÃ£o,
Data FinalizaÃ§Ã£o Acordo,,nÃ£o,nÃ£o,
Data InclusÃ£o Acordo,,nÃ£o,nÃ£o,
Data Nascimento,,nÃ£o,nÃ£o,
Data Nascimento RFB,,nÃ£o,nÃ£o,
Data Processamento,,nÃ£o,nÃ£o,
Data Recebimento Acordo,,nÃ£o,nÃ£o,
Data Requerente BeneficiÃ¡rio,,nÃ£o,nÃ£o,
Data Requerente Segurado,,nÃ£o,nÃ£o,
Data Requerimento,,nÃ£o,nÃ£o,
Data Ãšltimo Batimento,,nÃ£o,nÃ£o,
Dias DuraÃ§Ã£o Informado,,nÃ£o,nÃ£o,
DivisÃ£o CNAE 2.0,mais genÃ©rico que o classe CNAE,nÃ£o,nÃ£o,
DuraÃ§Ã£o Dias SuspensÃ£o,Ã‰ necessÃ¡rio disponibilizar essa informaÃ§Ã£o,nÃ£o,nÃ£o,
DuraÃ§Ã£o Meses SuspensÃ£o,,nÃ£o,nÃ£o,
Faixa EtÃ¡ria,,nÃ£o,nÃ£o,
Faixa Salarial BEm,,nÃ£o,nÃ£o,
Faixa Salarial MTE,,nÃ£o,nÃ£o,
Faixa Salarial SD,,nÃ£o,nÃ£o,

Faixa Tempo Acordo,,nÃ£o,nÃ£o,
Faixa Tempo Trabalhado,,nÃ£o,nÃ£o,
Grande Agrupamento Cnae 2.0,mais genÃ©rico que o classe CNAE,nÃ£o,nÃ£o,
Grupo CNAE 2.0,mais genÃ©rico que o classe CNAE,nÃ£o,nÃ£o,
Indicador Acordo Cancelado,,nÃ£o,nÃ£o,
Indicador Acordo Finalizado,,nÃ£o,nÃ£o,
Indicador Anistiado,nÃ£o sabemos o significado,sim,nÃ£o,
Indicador Batimento DiÃ¡rio,nÃ£o sabemos o significado,sin,nÃ£o,
Indicador BeneficiÃ¡rio,Trabalhador elegÃvel (aprovado) para o benifÃcio,nÃ£o,nÃ£o,
Indicador Dividiu SalÃ¡rio,nÃ£o sabemos o significado (tem sim e nulo na base),sim,nÃ£o,
Indicador Faturamento Sup. 4,8M (2019),Ã‰ um indicador ligado Ã empresa,nÃ£o,nÃ£o,
Indicador Requerente,"Trabalhador que requereu o benifÃcio, cadastrado pela empresa",nÃ£o,nÃ£o,
Indicador Requerimento Processado,,nÃ£o,nÃ£o,
Indicador Segurado,Trabalhador elegÃvel passa a receber o benifÃcio,nÃ£o,nÃ£o,
Lei,,nÃ£o,nÃ£o,
MatrÃcula eSocial,estranho estar tudo vazio,nÃ£o,sim?,
MunicÃpio Empresa,,nÃ£o,nÃ£o,
NIT Requerente,,nÃ£o,nÃ£o,
Nome MÃ£e Requerente,,nÃ£o,nÃ£o,
Nome Requerente,,nÃ£o,nÃ£o,
NÃºm InscriÃ§Ã£o Empregador,"CNPJ, certo?",sim,nÃ£o,
NÃºm. AgÃªncia BancÃ¡ria,,nÃ£o,nÃ£o,
NÃºmero Conta DV,,nÃ£o,nÃ£o,
NÃºmero Requerimento,,nÃ£o,nÃ£o,
NÃºmero da Conta,,nÃ£o,nÃ£o,
Percentual ReduÃ§Ã£o Carga HorÃ¡ria,"varia estranhamente, muitos em branco e alguns 0",nÃ£o,sim?,
Programa,"1 - 2020, 2 - inicia em abril/2021",nÃ£o,nÃ£o,
Qtd Meses Trabalhados,,nÃ£o,nÃ£o,
Qtd NotificaÃ§Ãµes,,nÃ£o,nÃ£o,
Qtd Parcelas Bloqueadas,,nÃ£o,nÃ£o,
Qtd Parcelas Emitidas,,nÃ£o,nÃ£o,
Qtd Parcelas Pagas,,nÃ£o,nÃ£o,
Qtd Parcelas Previstas,,nÃ£o,nÃ£o,
Qtd Recursos Deferidos,,nÃ£o,nÃ£o,
Qtd Recursos Indeferidos,,nÃ£o,nÃ£o,
RegiÃ£o GeogrÃ¡fica Empresa,,nÃ£o,nÃ£o,
Setor EconÃ´mico CNAE 2.0,,nÃ£o,nÃ£o,
SeÃ§Ã£o CNAE 2.0,,nÃ£o,nÃ£o,
Sigla UF Empresa,,nÃ£o,nÃ£o,
SituaÃ§Ã£o Acordo,,nÃ£o,nÃ£o,
SituaÃ§Ã£o Requerimento,,nÃ£o,nÃ£o,
Status Requerimento,,nÃ£o,nÃ£o,
Subclasse CNAE 2.0,,nÃ£o,nÃ£o,
Tipo AdesÃ£o,,nÃ£o,nÃ£o,
Tipo Empregador,Qual Ã© o significado?,sim,nÃ£o,
UF Empresa,,nÃ£o,nÃ£o,
Valor AntepenÃºltimo SalÃ¡rio,,nÃ£o,nÃ£o,

Valor AntepenÃºltimo SalÃ¡rio CNIS,Qual diferenÃ§a entre salÃ¡rio e salÃ¡rio CNIS?,sim,nÃ£o,
Valor MÃ©dia SalÃ¡rios,,nÃ£o,nÃ£o,
Valor Parcelas Bloqueadas,HÃ¡ nÃºmeros negativos,nÃ£o,sim?,
Valor Parcelas Emitidas,,nÃ£o,nÃ£o,
Valor Parcelas Pagas,,nÃ£o,nÃ£o,
Valor Parcelas Previstas,,nÃ£o,nÃ£o,
Valor PenÃºltimo SalÃ¡rio,,nÃ£o,nÃ£o,
Valor PenÃºltimo SalÃ¡rio CNIS,,nÃ£o,nÃ£o,
Valor Soma SalÃ¡rio,,nÃ£o,nÃ£o,
Valor Ãšltimo SalÃ¡rio,,nÃ£o,nÃ£o,
Valor Ãšltimo SalÃ¡rio CNIS,,nÃ£o,nÃ£o,

RAIS ESTABELECIMENTOS

variavel,observacao,duvida,inconsistencia,resposta
ceivinculado,CNPJ completo ou CEI do INSS?,sim,nÃ£o,
cepestab,,nÃ£o,nÃ£o,
cnae95classe,,nÃ£o,nÃ£o,
cnpjcei,CNPJ completo ou CEI do INSS?,sim,nÃ£o,
cnpjraiz,,nÃ£o,nÃ£o,
dataabertura,,nÃ£o,nÃ£o,
databaixa,,nÃ£o,nÃ£o,
dataencerramento,,nÃ£o,nÃ£o,
emailestabelecimento,"Pode identificar pessoas, principalmente MEIs (e nÃ£o Ã© relevante para anÃ¡lises de bases pÃ
indceivinculado,,nÃ£o,nÃ£o,
indestabparticipapat,,nÃ£o,nÃ£o,
indraisnegativa,NÃ£o tem no dicionÃ¡rio â€“ o que significa?,sim,nÃ£o,
indsimples,,nÃ£o,nÃ£o,
municÃpio,,nÃ£o,nÃ£o,
naturezajurÃdica,"Checar o que ocorre qdo natureza jurÃdica for 2135 (firma mercantil individual), 4014 (Empresa Indiv
RAIS2007)",nÃ£o,nÃ£o,
nomelogradouro,,nÃ£o,nÃ£o,
nÃºmerologradouro,,nÃ£o,nÃ£o,
nomebairro,,nÃ£o,nÃ£o,
nÃºmerotelefoneempresa,"Pode identificar pessoas, principalmente MEIs (e nÃ£o Ã© relevante para anÃ¡lises de bases
qtdvÃnculosativos,,nÃ£o,nÃ£o,
qtdvÃnculosclt,,nÃ£o,nÃ£o,
qtdvÃnculosestatutÃ¡rios,,nÃ£o,nÃ£o,
razÃ£osocial,Como tratar razÃ£o social de MEI na regra antiga? Como achar MEIs na base?,sim,nÃ£o,
tamanhoestabelecimento,,nÃ£o,nÃ£o,
tipoestab,Esclarecer: CNPJ ou CEI?,sim,nÃ£o,
ibgesubsetor,,nÃ£o,nÃ£o,
indatividadeano,NÃ£o sabemos o significado - esclarecer,sim,nÃ£o,
cnae20classe,,nÃ£o,nÃ£o,
cnae20subclasse,"SÃ³ serÃ¡ um problema, se for com as MEIs*",nÃ£o,nÃ£o,
uf,Foi um marcador da seleÃ§Ã£o desta amostra?,sim,nÃ£o,

munic_selec,Foi um marcador da seleÃ§Ã£o desta amostra?,sim,nÃ£o,

SEGURO DESEMPREGO – SD

variavel,observacao,duvida,inconsistencia,resposta
agentedigitaÃ§Ã£o,,nÃ£o,nÃ£o,
agenterecepÃ§Ã£o,,nÃ£o,nÃ£o,
anoacessodigitaÃ§Ã£o,,nÃ£o,nÃ£o,
anobeneficiario,,nÃ£o,nÃ£o,
anodemissÃ£orequerente,,nÃ£o,nÃ£o,
anodigitaÃ§Ã£o,,nÃ£o,nÃ£o,
anorequerente,,nÃ£o,nÃ£o,
anosegurado,,nÃ£o,nÃ£o,
anostatusrequerimento,,nÃ£o,nÃ£o,
bairrorequerente,,nÃ£o,nÃ£o,
ceprequerente,,nÃ£o,nÃ£o,
cpfrequerente,,nÃ£o,nÃ£o,
canaldeacesso,,nÃ£o,nÃ£o,
competacessodigitaÃ§Ã£o,,nÃ£o,nÃ£o,
competÃªnciabeneficiÃ¡rio,,nÃ£o,nÃ£o,
competÃªnciademissÃ£orequerente,,nÃ£o,nÃ£o,
competÃªnciadigitaÃ§Ã£o,,nÃ£o,nÃ£o,
competÃªnciarequerente,,nÃ£o,nÃ£o,
competÃªnciasegurado,,nÃ£o,nÃ£o,
competÃªnciastatusrequerimento,,nÃ£o,nÃ£o,
computadordigitaÃ§Ã£o,identificador da mÃ¡quina - confirmar se informaÃ§Ã£o de patrimÃ´nio e quem usa nÃ£o Ã© pÃ
cÃ³dagenterecepÃ§Ã£o,verificar se esse nÃºmero funcional Ã© de uso interno,sim,nÃ£o,
cÃ³dconvÃªnio,o que significa?,sim,nÃ£o,
codgrauinstrucao,,nÃ£o,nÃ£o,
codgenero,,nÃ£o,nÃ£o,
cÃ³dindicadoravisoprÃ©vio,,nÃ£o,nÃ£o,
cÃ³dindicadordireitosaldo,,nÃ£o,nÃ£o,
cÃ³dindicadorsentenÃ§ajudicial,,nÃ£o,nÃ£o,
cÃ³dmunicÃpiopostodigitaÃ§Ã£o,,nÃ£o,nÃ£o,
cÃ³dmunicÃpiopostorecepÃ§Ã£o,,nÃ£o,nÃ£o,
codmunicipioresidencia,,nÃ£o,nÃ£o,
cÃ³docupaÃ§Ã£ocbo,,nÃ£o,nÃ£o,
cÃ³dpostodigitaÃ§Ã£o,,nÃ£o,nÃ£o,
codpostorecepcao,,nÃ£o,nÃ£o,
cÃ³dsituaÃ§Ã£orequerente,,nÃ£o,nÃ£o,
cÃ³dstatusrequerimento,,nÃ£o,nÃ£o,
dataacessodigitaÃ§Ã£o,Ã© preciso abrir este tipo de informaÃ§Ã£o?,sim,nÃ£o,
dataadmissÃ£o,,nÃ£o,nÃ£o,
databeneficiario,Ã© preciso abrir este tipo de informaÃ§Ã£o?,sim,nÃ£o,
datademissÃ£orequerente,,nÃ£o,nÃ£o,
datadigitaÃ§Ã£orequerimento,Ã© preciso abrir este tipo de informaÃ§Ã£o?,sim,nÃ£o,

datanascimento,,nÃ£o,nÃ£o,
datarequerente,Ã© preciso abrir este tipo de informaÃ§Ã£o?,sim,nÃ£o,
datasatusrequerimento,Ã© preciso abrir este tipo de informaÃ§Ã£o?,sim,nÃ£o,
datasegurado,Ã© preciso abrir este tipo de informaÃ§Ã£o?,sim,nÃ£o,
datasentenÃ§ajudicial,Ã© preciso abrir este tipo de informaÃ§Ã£o?,sim,nÃ£o,
endereÃ§orequerente,,nÃ£o,nÃ£o,
faixaetaria,,nÃ£o,nÃ£o,
formadepagamento,,nÃ£o,nÃ£o,
grauinstrucao,,nÃ£o,nÃ£o,
genero,,nÃ£o,nÃ£o,
horaacessodigitaÃ§Ã£o,Ã© preciso abrir este tipo de informaÃ§Ã£o?,sim,nÃ£o,
ipinternetdigitaÃ§Ã£o,Ã© preciso abrir este tipo de informaÃ§Ã£o?,sim,nÃ£o,
ipintranetdigitaÃ§Ã£o,Ã© preciso abrir este tipo de informaÃ§Ã£o?,sim,nÃ£o,
indicadoravisoprÃ©vio,,nÃ£o,nÃ£o,
indicadordigital,,nÃ£o,nÃ£o,
indicadordireitosaldo,,nÃ£o,nÃ£o,
indicadorsentenÃ§ajudicial,,nÃ£o,nÃ£o,
indicadorsistema,Indica o sistema de pagamento ou de registro?,sim,nÃ£o,
intervalseguradobeneficiÃ¡rio,o que significa?,sim,nÃ£o,
intervalodemissÃ£orecepÃ§Ã£o,o que significa?,sim,nÃ£o,
intervalodigitaÃ§Ã£osegurado,o que significa?,sim,nÃ£o,
intervalorecepÃ§Ã£odigitaÃ§Ã£o,o que significa?,sim,nÃ£o,
leivigente,,nÃ£o,nÃ£o,
loterequerimento,,nÃ£o,nÃ£o,
matragentedigitaÃ§Ã£o,,nÃ£o,nÃ£o,
motivocancelamento,,nÃ£o,nÃ£o,
municÃpiopostodigitaÃ§Ã£o,,nÃ£o,nÃ£o,
municÃpiopostorecepcao,,nÃ£o,nÃ£o,
municipioresidencia,,nÃ£o,nÃ£o,
mÃªsÃºltimosalÃ¡rio,,nÃ£o,nÃ£o,
nitpisseguradoespecial,,nÃ£o,nÃ£o,
nomeconvÃªnio,,nÃ£o,nÃ£o,
nomemÃ£erequerente,,nÃ£o,nÃ£o,
nomerequerente,,nÃ£o,nÃ£o,
nÃºmeroctps,,nÃ£o,nÃ£o,
nÃºmeroinscriÃ§Ã£oempregador,"NÃ£o Ã© CNPJ, o que Ã©?",sim,nÃ£o,
nÃºmerorequerimento,"individualiza observaÃ§Ãµes na tabela, Ã© sequencial",nÃ£o,nÃ£o,
nÃºmerosentenÃ§ajudicial,Ã© preciso abrir este tipo de informaÃ§Ã£o?,sim,nÃ£o,
nÃºmerosÃ©riectps,,nÃ£o,nÃ£o,
ocupaÃ§Ã£ocbo,,nÃ£o,nÃ£o,
pispasepnit,,nÃ£o,nÃ£o,
placaderededigitaÃ§Ã£o,,nÃ£o,nÃ£o,
postodigitaÃ§Ã£o,,nÃ£o,nÃ£o,
postorecepcao,,nÃ£o,nÃ£o,
qtdbeneficiÃ¡rios,O que seriam os 0?,sim,nÃ£o,
qtdcontribuiÃ§Ãµesfgts,,nÃ£o,nÃ£o,
qtdcontribuiÃ§Ãµesinss,,nÃ£o,nÃ£o,

qtdnotificaÃ§Ãµes,,nÃ£o,nÃ£o,
qtdparcelaspagas,,nÃ£o,nÃ£o,
qtdparcelasprevistas,,nÃ£o,nÃ£o,
qtdrecursosdeferidos,,nÃ£o,nÃ£o,
qtdrecursosindeferidos,,nÃ£o,nÃ£o,
qtdrequerentes,,nÃ£o,nÃ£o,
qtdrequerimentos,,nÃ£o,nÃ£o,
qtdsegurados,,nÃ£o,nÃ£o,
regiÃ£opostodigitaÃ§Ã£o,,nÃ£o,nÃ£o,
regiÃ£orecepÃ§Ã£o,,nÃ£o,nÃ£o,
regiÃ£oresidÃªncia,,nÃ£o,nÃ£o,
siglaufdigitaÃ§Ã£o,,nÃ£o,nÃ£o,
siglaufemissÃ£octps,,nÃ£o,nÃ£o,
siglaufrecepÃ§Ã£o,,nÃ£o,nÃ£o,
siglaufresidÃªncia,,nÃ£o,nÃ£o,
situacaorequerimento,,nÃ£o,nÃ£o,
statusrequerimento,,nÃ£o,nÃ£o,
telefonerequerente,,nÃ£o,nÃ£o,
tempoempregorequerente,Qual Ã© a unidade de tempo? P/ enteder a granularidade do dado,sim,nÃ£o,
tipodepostodigitaÃ§Ã£o,,nÃ£o,nÃ£o,
tipodepostorecepÃ§Ã£o,,nÃ£o,nÃ£o,
ufpostodigitaÃ§Ã£o,,nÃ£o,nÃ£o,
ufpostorecepÃ§Ã£o,,nÃ£o,nÃ£o,
ufresidÃªncia,,nÃ£o,nÃ£o,
valorarredondparcprevistas,,nÃ£o,nÃ£o,
valorarredondparcemitidas,,nÃ£o,nÃ£o,
valorarredondparcelaspagas,,nÃ£o,nÃ£o,
valorcompensado,,nÃ£o,nÃ£o,
valorparcelasemitidas,,nÃ£o,nÃ£o,
valorparcelasprevistas,,nÃ£o,nÃ£o,
valortotalparcelaspagas,,nÃ£o,nÃ£o,
ÃºltimosalÃ¡rio,,nÃ£o,nÃ£o,
modalidade,,nÃ£o,nÃ£o,
agenteliberaÃ§Ã£oÃºltimanotiffase2,O que significa? NÃ£o tem nenhum preenchdo na amostra...,sim,nÃ£o,
anoanÃ¡lisefase1,,nÃ£o,nÃ£o,
anodemissÃ£o,,nÃ£o,nÃ£o,
anoliberaÃ§Ã£onotiffase1,,nÃ£o,nÃ£o,
anoliberaÃ§Ã£oÃºltimanotiffase2,,nÃ£o,nÃ£o,
anorequerimento,,nÃ£o,nÃ£o,
anoÃºltimanotiffase2,,nÃ£o,nÃ£o,
cpfgestorempweb,,nÃ£o,nÃ£o,
cpfprocuradorempweb,,nÃ£o,nÃ£o,
classecnae20,,nÃ£o,nÃ£o,
codindicadoravisoprÃ©vio,veio vazio - foi limpo antes do envio?,sim,nÃ£o,
codindicadordireitosaldo,veio vazio - foi limpo antes do envio?,sim,nÃ£o,
codindicadormesmoempregador,veio vazio - foi limpo antes do envio?,sim,nÃ£o,
codindicadorrecebeusalÃ¡rio,veio vazio - foi limpo antes do envio?,sim,nÃ£o,

competantepenÃºltimosalÃ¡rio,,nÃ£o,nÃ£o,
competanÃ¡lisefase1,,nÃ£o,nÃ£o,
competbeneficiÃ¡rio,,nÃ£o,nÃ£o,
competdemissÃ£o,,nÃ£o,nÃ£o,
competdigitaÃ§Ã£o,,nÃ£o,nÃ£o,
competliberaÃ§Ã£onotiffase1,,nÃ£o,nÃ£o,
competliberaÃ§Ã£oÃºltimanotiffase2,,nÃ£o,nÃ£o,
competpenÃºltimosalÃ¡rio,,nÃ£o,nÃ£o,
competrequerimento,,nÃ£o,nÃ£o,
competsegurado,,nÃ£o,nÃ£o,
competstatusrequerimento,,nÃ£o,nÃ£o,
competÃºltimanotiffase2,,nÃ£o,nÃ£o,
competÃºltimosalÃ¡rio,,nÃ£o,nÃ£o,
cÃ³dagenteliberaÃ§Ã£oÃºltimanotiffas,O que significa? NÃºmero funcional?,sim,nÃ£o,
cÃ³dmotivoliberaÃ§Ã£oÃºltimanotiffas,O que significa?,sim,nÃ£o,
cÃ³dagentedigitaÃ§Ã£o,O que significa? NÃºmero funcional?,sim,nÃ£o,
cÃ³dclassecnae20,,nÃ£o,nÃ£o,
cÃ³dfaixatempotrab,,nÃ£o,nÃ£o,
cÃ³dfaixatempotrabmp665,,nÃ£o,nÃ£o,
cÃ³dgrandesetoribge,5 nÃveis,nÃ£o,nÃ£o,
cÃ³dhabilitaÃ§Ã£o,4 nÃveis - nÃ£o sabemos o que Ã©,sim,nÃ£o,
cÃ³dinscriÃ§Ã£oempregador,,nÃ£o,nÃ£o,
cÃ³dmotivobloqueio,,nÃ£o,nÃ£o,
cÃ³dmotivocancelamento,,nÃ£o,nÃ£o,
cÃ³dmotivodispensa,,nÃ£o,nÃ£o,
cÃ³dmotivorequerimento,,nÃ£o,nÃ£o,
cÃ³dmunicÃpiodemissÃ£o,,nÃ£o,nÃ£o,
cÃ³dmunicÃpiodigitaÃ§Ã£o,,nÃ£o,nÃ£o,
cÃ³dmunicÃpiorecepÃ§Ã£o,,nÃ£o,nÃ£o,
cÃ³dpostorecepÃ§Ã£o,,nÃ£o,nÃ£o,
cÃ³dsituaÃ§Ã£orequerimento,,nÃ£o,nÃ£o,
cÃ³dsubclassecnae20,,nÃ£o,nÃ£o,
cÃ³dtipopostodigitaÃ§Ã£o,,nÃ£o,nÃ£o,
cÃ³dtipopostorecepÃ§Ã£o,,nÃ£o,nÃ£o,
dddgestorempweb,,nÃ£o,nÃ£o,
dddprocuradorempweb,,nÃ£o,nÃ£o,
dddtelefonerequerente,,nÃ£o,nÃ£o,
dataadmissÃ£oprocuradorempweb,,nÃ£o,nÃ£o,
dataadmissÃ£oreq,,nÃ£o,nÃ£o,
dataanÃ¡lisefase1,,nÃ£o,nÃ£o,
databeneficiÃ¡rio,,nÃ£o,nÃ£o,
dataliberaÃ§Ã£onotiffase1,,nÃ£o,nÃ£o,
dataliberaÃ§Ã£oÃºltimanotiffase2,,nÃ£o,nÃ£o,
datastatusrequerimento,,nÃ£o,nÃ£o,
dataÃºltimanotiffase2,,nÃ£o,nÃ£o,
descriÃ§Ã£oqtdparcprevistas,,nÃ£o,nÃ£o,
diabeneficiÃ¡rio,,nÃ£o,nÃ£o,

diademissÃ£o,,nÃ£o,nÃ£o,
diadigitaÃ§Ã£o,,nÃ£o,nÃ£o,
diarequerimento,,nÃ£o,nÃ£o,
diasegurado,,nÃ£o,nÃ£o,
diastatus,,nÃ£o,nÃ£o,
emalgestorempweb,,nÃ£o,nÃ£o,
faixareincidÃªncia,,nÃ£o,nÃ£o,
faixasalarial,,nÃ£o,nÃ£o,
faixasalarialsd,,nÃ£o,nÃ£o,
faixatempotrabmp665,,nÃ£o,nÃ£o,
faixatempotrabalhado,,nÃ£o,nÃ£o,
formapagamento,,nÃ£o,nÃ£o,
grandesetoribge,,nÃ£o,nÃ£o,
habilitaÃ§Ã£o,O que significa?,sim,nÃ£o,
indicadoranalisadofase1,,nÃ£o,nÃ£o,
indicadordiferenÃ§asalÃ¡riocnis,,nÃ£o,nÃ£o,
indicadorestÃ¡notificadofase2,,nÃ£o,nÃ£o,
indicadorfoinotificadofase2,,nÃ£o,nÃ£o,
indicadorliberadonotiffase1,,nÃ£o,nÃ£o,
indicadormesmoempregador,,nÃ£o,nÃ£o,
indicadorpronatec,,nÃ£o,nÃ£o,
indicadorpublicoprioritÃ¡rio,,nÃ£o,nÃ£o,
indicadorrecebeusalÃ¡rio,,nÃ£o,nÃ£o,
indicadorsegurocompleto,,nÃ£o,nÃ£o,
inscriÃ§Ã£oempregadorceicnpj,,nÃ£o,nÃ£o,
intervaloseguradobeneficiÃ¡rio,O que significa?,sim,nÃ£o,
logradourorequerente,,nÃ£o,nÃ£o,
lote,,nÃ£o,nÃ£o,
motivobloqueio,,nÃ£o,nÃ£o,
motivoliberaÃ§Ã£oÃºltimanotiffase2,,nÃ£o,nÃ£o,
municÃpiodemissÃ£o,,nÃ£o,nÃ£o,
municÃpiodigitaÃ§Ã£o,,nÃ£o,nÃ£o,
municÃpiorecepÃ§Ã£o,,nÃ£o,nÃ£o,
nomefantasiaempregador,,nÃ£o,nÃ£o,
nomegestorempweb,,nÃ£o,nÃ£o,
nomeprocuradorempweb,,nÃ£o,nÃ£o,
numqtdparcprevistas,,nÃ£o,nÃ£o,
numeroctps,,nÃ£o,nÃ£o,
numerocomunicadodispensa,"parece individualizar o registro, resta saber se Ã© acessÃvel externamente",sim,nÃ£o,
numeroprotocolo,"parece individualizar o registro, resta saber se Ã© acessÃvel externamente",sim,nÃ£o,
numerorequerimento,"parece individualizar o registro, resta saber se Ã© acessÃvel externamente",sim,nÃ£o,
numerosentenÃ§ajudicial,,nÃ£o,nÃ£o,
numerosÃ©riectps,,nÃ£o,nÃ£o,
origemreqmigraÃ§Ã£o,O que significa?,sim,nÃ£o,
pispasepnitbase,,nÃ£o,nÃ£o,
pispasepprocuradorempweb,,nÃ£o,nÃ£o,
qtdmesesempregorequerente,,nÃ£o,nÃ£o,

qtdparcelasbloqueadas,,nÃ£o,nÃ£o,
qtdparcelasbloqueadaspronatec,,nÃ£o,nÃ£o,
qtdpublicoprioritÃ¡rio,,nÃ£o,nÃ£o,
qtdreincidÃªncias,,nÃ£o,nÃ£o,
razÃ£osocialempregador,"Como tratar razÃ£o social de MEI na regra antiga? Como achar MEIs na base? Aplicar filtro com mÃ¡scara?",sim,nÃ£o,
regiÃ£odemissÃ£o,,nÃ£o,nÃ£o,
regiÃ£odigitaÃ§Ã£o,,nÃ£o,nÃ£o,
siglaufctps,,nÃ£o,nÃ£o,
siglaufdemissÃ£o,,nÃ£o,nÃ£o,
subclassecnae20,,nÃ£o,nÃ£o,
telefonegestorempweb,,nÃ£o,nÃ£o,
telefoneprocuradorempweb,,nÃ£o,nÃ£o,
tipoinscriÃ§Ã£oempregador,,nÃ£o,nÃ£o,
tipopostodigitaÃ§Ã£o,,nÃ£o,nÃ£o,
tipopostorecepÃ§Ã£o,,nÃ£o,nÃ£o,
ufdemissÃ£o,,nÃ£o,nÃ£o,
ufdigitaÃ§Ã£o,,nÃ£o,nÃ£o,
ufrecepÃ§Ã£o,,nÃ£o,nÃ£o,
ufresidencia,,nÃ£o,nÃ£o,
usuÃ¡riogestorempweb,O que Ã©? Valor numÃ©rico que nÃ£o parece bater com nenhum formato conhecido.,sim,nÃ£o
valorantepenÃºltimosalÃ¡rio,,nÃ£o,nÃ£o,
valorantepenÃºltimosalÃ¡riocnis,,nÃ£o,nÃ£o,
valorarredondparcpagas,,nÃ£o,nÃ£o,
valormÃ©diasalÃ¡rios,,nÃ£o,nÃ£o,
valorparcelasbloqueadas,,nÃ£o,nÃ£o,
valorparcelasbloqueadaspronatec,,nÃ£o,nÃ£o,
valorparcelascompensadas,,nÃ£o,nÃ£o,
valorparcelaspagas,,nÃ£o,nÃ£o,
valorpenÃºltimosalÃ¡rio,,nÃ£o,nÃ£o,
valorpenÃºltimosalÃ¡riocnis,,nÃ£o,nÃ£o,
valorsomasalÃ¡rio,,nÃ£o,nÃ£o,
valorÃºltimosalÃ¡rio,,nÃ£o,nÃ£o,
valorÃºltimosalÃ¡riocnis,,nÃ£o,nÃ£o,
Ãºltimaregrafase2,O que significa?,sim,nÃ£o,
codtipopostorecepcao,,nÃ£o,nÃ£o,
anodigitaÃ§Ã£orequerimento,,nÃ£o,nÃ£o,
anosuspensÃ£o,,nÃ£o,nÃ£o,
antepenÃºltimosalÃ¡rio,,nÃ£o,nÃ£o,
codmotivocancelamento,,nÃ£o,nÃ£o,
codsituacaorequerimento,,nÃ£o,nÃ£o,
codstatusrequerimento,,nÃ£o,nÃ£o,
competdigitaÃ§Ã£orequerimento,,nÃ£o,nÃ£o,
competÃªnciaacessodigitaÃ§Ã£o,,nÃ£o,nÃ£o,
competÃªnciasuspensÃ£o,,nÃ£o,nÃ£o,
cÃ³ddivisÃ£ocnae20,,nÃ£o,nÃ£o,
cÃ³dentidadeconveniada,O que significa?,sim,nÃ£o,

codfaixatempotrabalhado,,nÃ£o,nÃ£o,
cÃ³dgrupocnae20,,nÃ£o,nÃ£o,
cÃ³dindicadorrecebeusalÃ¡rio,,nÃ£o,nÃ£o,
codmunicipiorecepcao,,nÃ£o,nÃ£o,
cÃ³dmunicÃpiosuspensÃ£o,,nÃ£o,nÃ£o,
cÃ³dseÃ§Ã£ocnae20,,nÃ£o,nÃ£o,
codsubclassecnae20,,nÃ£o,nÃ£o,
cÃ³dtipoinscriÃ§Ã£oempregador,,nÃ£o,nÃ£o,
ddd,,nÃ£o,nÃ£o,
dataadmissÃ£orequerente,,nÃ£o,nÃ£o,
datasuspensÃ£o,,nÃ£o,nÃ£o,
divisÃ£ocnae20,,nÃ£o,nÃ£o,
famÃliacbo,,nÃ£o,nÃ£o,
grupocbo,,nÃ£o,nÃ£o,
grupocnae20,,nÃ£o,nÃ£o,
intervalosuspensÃ£orecepÃ§Ã£o,,nÃ£o,nÃ£o,
matragenterecepÃ§Ã£o,verificar se esse nÃºmero funcional Ã© de uso interno,sim,nÃ£o,
municipiorecepcao,,nÃ£o,nÃ£o,
municÃpiosuspensÃ£o,,nÃ£o,nÃ£o,
numinscempregadorceicnpj,,nÃ£o,nÃ£o,
penÃºltimosalÃ¡rio,,nÃ£o,nÃ£o,
percentualaulasprÃ¡ticas,,nÃ£o,nÃ£o,
qtdhorascurso,,nÃ£o,nÃ£o,
qtdmesesprorrogados,,nÃ£o,nÃ£o,
regiÃ£opostorecepÃ§Ã£o,,nÃ£o,nÃ£o,
regiÃ£osuspensÃ£o,,nÃ£o,nÃ£o,
seÃ§Ã£ocnae20,,nÃ£o,nÃ£o,
siglaufpostodigitaÃ§Ã£o,,nÃ£o,nÃ£o,
siglaufpostorecepÃ§Ã£o,,nÃ£o,nÃ£o,
siglaufsuspensÃ£o,,nÃ£o,nÃ£o,
subgrupocbo,,nÃ£o,nÃ£o,
subgrupoprincipalcbo,,nÃ£o,nÃ£o,
tipodepostorecepcao,,nÃ£o,nÃ£o,
ufsuspensÃ£o,,nÃ£o,nÃ£o,
valorarredondparcemitida,,nÃ£o,nÃ£o,
codgrandesetoribge,,nÃ£o,nÃ£o,
anocadastrodefeso,,nÃ£o,nÃ£o,
anofimdefeso,,nÃ£o,nÃ£o,
anoinÃciodefeso,,nÃ£o,nÃ£o,
apelidorequerente,,nÃ£o,nÃ£o,
atividadepesqueira,,nÃ£o,nÃ£o,
ceirequerente,,nÃ£o,nÃ£o,
cnpjcolÃ´nia,o que seria colÃ´nia?,sim,nÃ£o,
competÃªnciacadastrodefeso,,nÃ£o,nÃ£o,
competÃªnciafimdefeso,,nÃ£o,nÃ£o,
competÃªnciainÃciodefeso,,nÃ£o,nÃ£o,
competÃªnciarequerimento,,nÃ£o,nÃ£o,

v27,o que significa? AA/MMMM,sim,nÃ£o,
cÃ³dmunicÃpiocolÃ´nia,o que seria colÃ´nia?,sim,nÃ£o,
cÃ³dtipomotivocancelamento,,nÃ£o,nÃ£o,
datacadastrodefeso,,nÃ£o,nÃ£o,
datacredenciamento,,nÃ£o,nÃ£o,
datadigitaÃ§Ã£o,,nÃ£o,nÃ£o,
datafimdefeso,,nÃ£o,nÃ£o,
datahoraacessodigitaÃ§Ã£o,,nÃ£o,nÃ£o,
datainÃciodefeso,,nÃ£o,nÃ£o,
dataportariadefeso,,nÃ£o,nÃ£o,
dataprimeirorgp,,nÃ£o,nÃ£o,
datarequerimento,,nÃ£o,nÃ£o,
dataÃºltimaalteraÃ§Ã£o,,nÃ£o,nÃ£o,
descriÃ§Ã£odefeso,O que significa DEFESO?,sim,nÃ£o,
faixaintervalodefeso,,nÃ£o,nÃ£o,
identidaderequerente,,nÃ£o,nÃ£o,
logradouro,,nÃ£o,nÃ£o,
municÃpiocolÃ´nia,,nÃ£o,nÃ£o,
nitpisrequerente,,nÃ£o,nÃ£o,
nomecolÃ´nia,o que seria colÃ´nia?,sim,nÃ£o,
nÃºmerodefeso,,nÃ£o,nÃ£o,
nÃºmeroportariadefeso,,nÃ£o,nÃ£o,
nÃºmerorgp,o que significa?,sim,nÃ£o,
siglaufcolÃ´nia,o que seria colÃ´nia?,sim,nÃ£o,
tipomotivocancelamento,,nÃ£o,nÃ£o,
anoresgate,,nÃ£o,nÃ£o,
competÃªnciaresgate,,nÃ£o,nÃ£o,
cÃ³dmunicÃpionaturalidade,,nÃ£o,nÃ£o,
dataresgate,,nÃ£o,nÃ£o,
dataresgaterequerente,,nÃ£o,nÃ£o,
v44,o que significa? data DD/MM/AAAA,sim,nÃ£o,
formaÃ§Ã£opretendidacbo,,nÃ£o,nÃ£o,
formaÃ§Ã£opretendidagrupo,,nÃ£o,nÃ£o,
grupocboatual,,nÃ£o,nÃ£o,
grupocbopretendido,,nÃ£o,nÃ£o,
v59,o que significa? booleana sim/nÃ£o,sim,nÃ£o,
municÃpionaturalidade,,nÃ£o,nÃ£o,
ocupaÃ§Ã£ocboatual,,nÃ£o,nÃ£o,
ocupaÃ§Ã£ocbopretendida,,nÃ£o,nÃ£o,
raÃ§arequerente,,nÃ£o,nÃ£o,
regiÃ£onaturalidade,,nÃ£o,nÃ£o,
siglaufnaturalidade,,nÃ£o,nÃ£o,
ufemissÃ£octps,,nÃ£o,nÃ£o,
ufnaturalidade,,nÃ£o,nÃ£o,