Integração do **Cadastro Único com Cadastro Nacional de Endereços** para Fins Estatístico através da modelagem de um banco de dados espacial

MINISTÉRIO DO
DESENVOLVIMENTO E ASSISTÊNCIA SOCIAL, FAMÍLIA E COMBATE À FOME

GOVERNO FEDERAL
BRASIL
UNIÃO E RECONSTRUÇÃO

Ministério do Desenvolvimento e Assistência
Social, Família e Combate à Fome

Integração do Cadastro Único com Cadastro Nacional de Endereços para Fins Estatístico através da modelagem de um banco de dados espacial

2023

Sumário

Lista de Figuras

Lista de Tabelas

Prefácio

A Secretaria de Avaliação, Gestão da Informação e Cadastro Único (Sagicad) produz pesquisas de avaliação e estudos técnicos sobre programas e políticas do Ministério do Desenvolvimento e Assistência Social, Família e Combate à Fome. Esses instrumentos permitem a elaboração de diagnósticos aprofundados acerca dos públicos-alvo das políticas, insumos para desenho e redesenho de programas, e conhecimento geral acerca das ações governamentais. Por meio de cooperação técnica com entidades públicas e privadas, espera-se contribuir na reflexão e apontamento de caminhos que conduzam ao amadurecimento institucional das políticas sociais do país.

Em setembro de 2019 a então denominada Sagi, por meio de seu Departamento de Avaliação, firmou um Termo de Execução Descentralizada com o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), com objetivo de lançar Chamada Pública que selecionasse projetos de pesquisa previamente demandados pelas áreas finalísticas do, à época, Ministério da Cidadania, resultando na Chamada Pública CNPq/ Ministério da Cidadania nº 30/2019.

A Chamada trabalhou 26 temas de pesquisa referentes a diversas políticas ou programas do Ministério, dentre os quais estavam as demandas para qualificação de base de dados de Cadastro Único para Programas Sociais do Governo Federal (Cadastro Único), voltadas para o atendimento das necessidades da área gestora do Cadastro e demais órgãos do Ministério que operam com esse sistema de informações.

Uma das linhas de pesquisa da chamada pública tratou da proposta de aperfeiçoamento da metodologia informacional de pareamento dos endereços do Cadastro Único com o Cadastro Nacional de Endereços (CNEFE) do Instituto Brasileiro de Geografia e Estatística (IBGE). O pa-

reamento dessas informações possui como finalidade principal qualificar e padronizar as informações de endereço disponíveis no Cadastro Único. Além disso, tal produto irá permitir uma atualização mais frequente dos endereços, bem como a geolocalização por coordenadas geográficas dos dados de famílias e pessoas, para fins de análise e formulação de políticas públicas. Para tanto, foram utilizados como referência todos os registros anuais do Cadastro Único contidos entre 2012 e 2019 e uma extração única mais recente para a base do CNEFE.

Os bancos de dados geográficos representam as entidades reais e suas diferentes feições de diferentes formas, sendo que nem sempre essas representações são compatíveis ou equiparáveis entre os diferentes mapeamentos. Dessa forma, é muito comum que existam erros ou divergências nas representações semânticas de um atributo espacial que gerem ambiguidades ou imprecisão na geolocalização.

Esses equívocos se devem a erro de digitação, duplicidade nas representações (nomes diferentes para a mesma rua), abreviações e desconhecimento da grafia por parte dos registradores. Por conta dessas diferenças representativas e semânticas, ao se parear dois bancos de dados, faz-se necessário encontrar métricas capazes de avaliar e quantificar o grau de similaridade entre duas representações.

Frente ao desafio apresentado, o produto propõe a construção de algoritmo capaz de comparar os endereços oriundos dos dois bancos, de acordo com seu grau de similaridade. Para cada instância do Cadastro Único é escolhida uma melhor instância do CNEFE que possua o nome do bairro, rua e número o mais similar possível. Desse modo, a metodologia estabelecida permite parear os dois bancos de dados de acordo com diferentes graus de precisão estabelecidos pelos gestores, retornando um maior ou menor número de resultados encontrados de acordo com o grau de precisão estabelecido.

Além de apresentarem toda as etapas necessárias para a construção do algoritmo de pareamento, os pesquisadores demonstram o grau de eficiência dessa metodologia para as diferentes Unidades da Federação e para diferentes limiares de precisão estabelecidos. Nas tabelas de resultado apresentadas, é possível notar um bom grau de pareamento para a quase totalidade dos estados, o que indica que a metodologia construída está adequada para o objetivo proposto para o trabalho.

Ao apresentar esse trabalho, espera-se contribuir com a difusão dos conhecimentos e técnicas construídos, de tal modo que seja possível uma melhor integração das bases e registros administrativos disponíveis na administração pública, com o fim último de aprimorar os serviços e políticas oferecidos ao cidadão.

Finalmente, agora a Sagicad tem a oportunidade de divulgar esta pesquisa.
Boa leitura!

Diego Rodrigues Macedo[1],
Vagner Braga Nunes Coelho[2],
Guilherme Henrique Rodrigues Nascimento[3],
Fredy Sales Ribeiro[4],
Lanna Kallen Parreiras[5]

1 Professor Doutor do Instituto de Geociências da Universidade Federal de Minas Gerais (UFMG) – Coordenador responsável pela Sublinha de Pesquisa 1.1 da Chamada Pública CNPq/MC nº 30/2019.

2 Professor Doutor do Instituto de Geociências da UFMG.

3 Aluno de Pós-Graduação do Departamento de Ciência da Computação da UFMG.

4 Aluno de Graduação da Escola de Engenharia da UFMG.

5 Aluna de Graduação da Escola de Engenharia da UFMG.

Introdução

O mundo real pode ser entendido como um conjunto de feições cuja espacialização pode ser identificada, por exemplo, por ruas, praças, monumentos e demais entidades reais existentes em uma determinada localidade. A percepção desse conjunto, por meio de funções de mapeamento, permite instanciar essas representações em Bancos de Dados Geográficos (BDG), propiciando a armazenagem de uma coleção de dados coerentes e estruturados, visando permitir processamentos posteriores.

Uma vez que o mundo real é modelado por n produtores de dados, cada representação individual em um BDG pode ser similar, mas não necessariamente igual às dos demais produtores. Com isso, a partir dos modelos individuais dos vários produtores de dados, obtém-se uma multiplicidade de representações advindas de um mesmo domínio, gerando um contra-domínio não necessariamente igual (COELHO: 2010).

Dentre as representações no BDG, cada instância compreende os aspectos espaciais - relacionados à descrição dos atributos geométricos - e os aspectos semânticos, estabelecidos usualmente por meio de *strings* (cadeia de caracteres) que nomeiam/identificam a feição.

Ao se analisar os aspectos semânticos de uma determinada feição instanciada em um BDG, deve-se considerar equívocos recorrentes nos dados, advindos de diversos fatores, tais como erros de grafia, convenções de escrita e duplicidade de representações. Logo, pode-se pressupor que existam ambiguidades no BDG, uma vez que dados que tratam de uma mesma entidade real passam a ser relacionados como elementos diferentes ou, até mesmo, associados a mais de um elemento em um mesmo BDG. Isso ocorre, pois funções de mapeamento distintas podem gerar representações de nomes diferentes para a mesma feição.

Por esse motivo, ao analisar BDG diferentes, a fim de integrá-los, deve-se levar em consideração relações de semelhança entre seus elementos, de modo a avaliar e quantificar o quanto uma *string* é similar a outra e, mais além, inferir quando determinadas representações podem ser consideradas as mesmas e, consequentemente eliminando ambiguidades potenciais. Por conseguinte, determinar uma função de similaridade que permita avaliar representações, propondo um valor de proximidade, torna-se essencial no processo de identificação de entidades em um BDG. Assim, busca-se identificar e estabelecer um conjunto unívoco das representações a partir dos diversos conjuntos construídos por diferentes produtores de dados.

A existência de produtores de dados distintos no Brasil produz bancos de dados com instâncias diferentes, o que dificulta a correspondência ou pareamento entre *strings* que representam a mesma feição. Essas dificuldades poderiam ser mais bem geridas se existisse uma norma governamental que estabelecesse diretrizes e protocolos para a utilização de uma dada informação gerada por um único produtor de dado.

O projeto em voga procura estabelecer pareamento por similaridade entre BDG distintos que contenham endereços (dados semânticos) como instâncias. Neste caso, serão considerados o Cadastro Nacional de Endereços para Fins Estatísticos (CNEFE) criado e mantido pelo Instituto Brasileiro de Geografia e Estatística (IBGE) e o Cadastro Único (CadÚnico) criado para atender Programas Sociais do Governo Federal .

O BDG do CadÚnico e do CNEFE apresentam diferenças que não possibilitam a correspondência direta entre as representações. Com isso, tem-se uma dificuldade de acesso às informações georreferenciadas, pois existem instâncias duplicadas referentes à mesma realidade, devido à complexidade em se estabelecer relações de similaridades entre as strings.

No caso particular do endereço, este possui uma série de atributos que precisam ser considerados para a criação de uma string. Assim, informações quanto ao tipo de logradouro, nome, número, bairro, dentre outras, tornam-se essenciais para que haja condições técnicas para se identificar endereços semelhantes como representativos da mesma entidade física (feição).

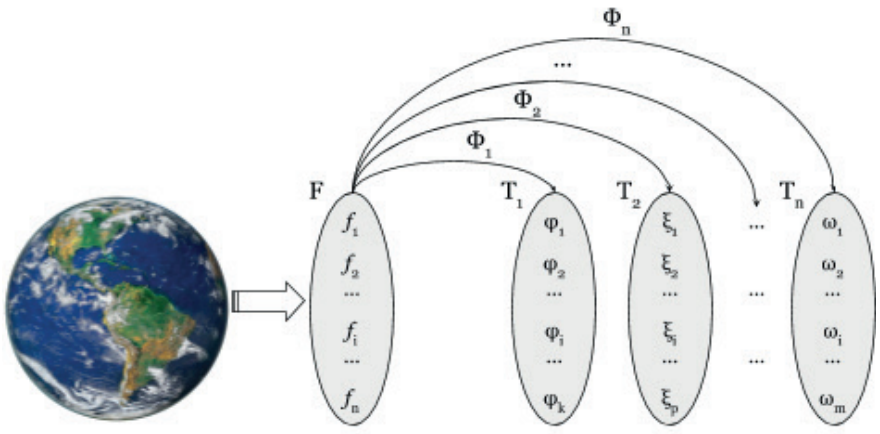
1. Fundamentação Teórica

Seja Fo conjunto das instâncias do mundo real. Tem-se que $F = \{f_1, f_2, ..., f_i\}$, onde f i é uma feição particular do mundo real. Um produtor qualquer de BDG modela *F* de tal maneira que as entidades reais f_i sejam instanciadas, no tempo e no espaço, através de uma função de mapeamento $\Phi_j(f_1)$ ou, simplesmente, Φ_j, para j = 1...*n* , sendo n o número de produtores de dados. Neste caso, Φ_j é uma visão particular de *F* que possibilita estabelecer outros conjuntos contendo representações Φ dessas entidades reais. Assim, tem-se:

$$\forall f_i \in F \Rightarrow \Phi_j(f_i) = \varphi_{ji} \quad (1)$$

Considerando φ apenas como o atributo não geométrico das representações, como, por exemplo, o *nome* de uma entidade real, seja o conjunto *T*, onde $T=\{\varphi_1, \varphi_2, ..., \varphi_i\}$. Assim, dados os n produtores de dados, obtém-se uma multiplicidade de representações advindas de um mesmo domínio, gerando um contra-domínio não necessariamente igual (Figura 1).

Figura 1 - Esquema conceitual da multiplicidade de representações



Da Figura 1, constata-se ser possível estabelecer uma relação biunívoca entre um conjunto T_j qualquer e um determinado subconjunto Q_j de *F*.

Isto ocorre porque cada feição f_i mapeada gera uma representação de modo único. Logo, é possível aplicar a função Φ_j à feição pertencente à Q_j e a função inversa à representação em T_j. Daí, tem-se:

$$\Phi_j (Q_j) = T_{j,} \; j = 1... \; n \wedge Qj \subset F \tag{2}$$

Neste projeto, propõe-se a construção teórica de um banco de dados resultante Q′ obtido como resultado da integração de dois BDG distintos (T_1 e T_2). Isso deve ser feito levando em consideração relações de similitudes entre seus elementos. Sob essa premissa, cada instância de Q′ advém dessas similitudes e das representações exclusivas de cada BDG, conforme Equação 3.

$$|Q'| = |T_1 \approx T_2| + |T_1 - T_2| + |T_2 - T_1| \tag{3}$$

Nesse sentido, $T_1 \approx T_2$ representa as instâncias comuns ou similares nos conjuntos, e $T_1 - T_2$ e $T_2 - T_1$ equivalem às representações de T_1 e T_2 onde não foram identificadas instâncias similares em T_2 e T_1, respectivamente.

Assim, procura-se estabelecer uma função de similaridade S capaz de propor um valor de proximidade entre as instâncias consideradas no BGD. Por conseguinte, um limite pré-estabelecido L serve como parâmetro definidor da similitude entre duas representações quaisquer φ_1 e φ_2, da seguinte forma:

$$S(\varphi_1, \varphi_2) \geq L \iff \varphi_1 \approx \varphi_2 \tag{4}$$

Além disso, a fim de se contemplar o espaço métrico (LIMA: 1993) normalizado entre [0, 1], no qual o valor 0 representa uma total dissimilaridade e o valor 1 uma similaridade perfeita, ou seja, a igualdade (identidade) entre as *strings* consideradas, essa função deve ser tal que, dado um métrica qualquer M, tem-se $0 \leq S \leq 1 - M$.

Neste projeto, entende-se por similaridade a similitude entre duas *strings* por meio da análise comparativa de suas respectivas cadeias de caracte-

res. Desse modo, quanto maior for a diferença semântica entre elas, menos similares elas serão. No mesmo sentido, a similitude máxima se dará quando as *strings* forem idênticas (LI: 1998). Por esse motivo, o método de análise de similaridade pode ser utilizado para resolver problemas de ambiguidades entre instâncias de um BDG, uma vez que podem existir representações distintas de uma mesma feição.

2. Metodologia

Diante da possibilidade de existência de representações potencialmente ambíguas, deve-se estabelecer uma função de similaridade, bem como um limiar de aceitação da função $S(\varphi_1, \varphi_2)$, capaz de quantificar a similitude entre as strings. Para isso, deve-se seguir os seguintes procedimentos:

- levantamento de dados;
- proposta de função de similaridade;
- estabelecimento do limiar de aceitação;
- criação de conjunto identificado Q';
- análise do resultado.

2.1 Levantamento de Dados

Para realizar o pareamento de endereços entre BDG distintos, a fim de se gerar um conjunto resultante Q', primeiro deve-se considerar o levantamento de dados, etapa primordial na qualificação dos resultados. Para tal, é necessário realizar a coleta de dados em BDG de produtores diferentes, mas que contemplam o mesmo campo de realidade, ou seja, que mapeiam um mesmo subconjunto Q real *F*.

Realizada a coleta, os dados obtidos devem ser armazenados em tuplas distintas *T*, sendo estas organizados por pares, de modo que por meio da função de similaridade S possam ser identificados os pontos de contato entre estas, bem como os dados exclusivos de cada BDG.

2.2 Proposta de Função de Similaridade

A função de similaridade permite um correlacionamento entre as instâncias do BDG, tanto interno quanto externamente. Assim, é possível estabelecer a interseção entre dois BDG distintos, bem como analisar as instâncias exclusivas presentes em apenas um desses conjuntos.

Para se determinar essa função, deve-se analisá-la quanto ao atendimento do espaço métrico. Dado isso, uma segunda avaliação é realizada de modo a analisar se ela atende aos requisitos deste trabalho.

2.3 Estabelecimento do Limiar de Aceitação

Proposta a função de similaridade, cabe estabelecer um limiar de aceitação para que essa função seja implementada, isto é, um valor capaz de considerar se as strings analisadas se referem à mesma realidade. Ressalta-se que o limiar é um valor propositivo. Neste caso, é possível modificar o valor proposto e, desta forma, alterar a quantidade de pareamentos obtidos. É importante ter conhecimento da base de dados, pois os valores do limiar podem gerar resultados superestimados, caso o limiar seja baixo a ponto de parear instâncias diferentes no mundo real, mas que possuem grafia próxima; ou subestimados, caso o limiar seja alto ao ponto de não parear a mesma instância do mundo real devido ao um pequeno detalhe de grafia.

2.4 Criação de Conjunto Identificado Q'

O pareamento entre tuplas de BDG distintos, bem como os registros únicos de cada um dos bancos de dados, permite a construção de um conjunto Q'. Observa-se, portanto, que Q' contempla a totalidade de feições mapeadas nos BDG analisados. Nesse caso, Q' é um conjunto cuja cardinalidade é maior ou igual do que cada um dos conjuntos T_1 e T_2 considerados individualmente.

2.5 Análise do Resultado

Realizado o correlacionamento entre as instâncias, avalia-se os resultados com base nas similaridades apresentadas. Com isso, determina-se a abrangência de endereços obtida pelo estudo e a qualidade do processo. Assim, será possível quantificar quão eficiente foi a função de similaridade e o limiar de aceitação utilizado.

A partir disso, o conjunto Q' pode ser utilizado pelos usuários para fins diversos, o que garante uma maior abrangência e menores equívocos, além de permitir a redução de gastos.

3. Bases de Dados

3.1 Cadastro Único (CadÚnico)

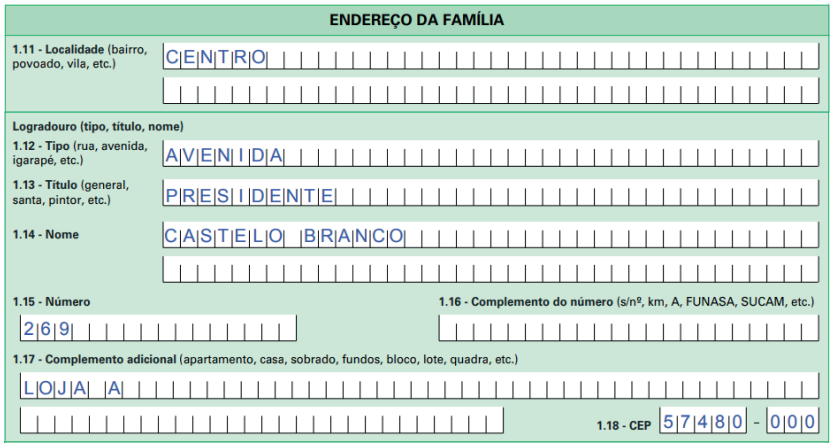
O Cadastro Único para Programas Sociais do Governo Federal (CadÚnico), criado em 2001 por meio do Decreto Federal nº 3.877, é um instrumento que identifica e caracteriza as famílias com renda mensal de até meio salário-mínimo per capita ou renda familiar mensal de até 3 salários mínimos no total, permitindo que o Governo conheça melhor a realidade

socioeconômica dessa população e possa direcionar políticas públicas a essas famílias vulneráveis. Nele são registradas informações como: características da residência, identificação de cada pessoa, escolaridade, situação de trabalho e renda, entre outras (MDS: 2011).

Por proporcionar em uma só base um retrato abrangente das condições de vida deste público, o CadÚnico é uma importante ferramenta de gestão nas três esferas de governo para implementação de programas sociais, ações e serviços voltados à população de baixa renda. Destaca-se sua relevância quanto a integração dos dados acerca das famílias de baixa renda, permitindo maior controle sobre a distribuição dos recursos, evitando a sobreposição e acúmulo de benefícios de forma irregular. Além disso, por meio do georreferenciamento das famílias é possível cruzar informações com outras bases espaciais, como por exemplo dados em nível de setores censitários, ou até mesmo acompanhar a migração de uma mesma família em escala municipal e até mesmo inframunicipal. Dessa maneira, pode-se ampliar as possibilidades analíticas do CadÚnico, tornando os processos dentro do Ministério da Cidadania mais eficientes e sofisticados.

A determinação do endereço das famílias para o CadÚnico é realizada através da coleta de dados em entrevista, via bloco 01 do formulário geral de coleta (Figura 2), feita pelo órgão público competente, em geral na esfera municipal. Para a coleta de dados, o ideal seria que todas fossem realizadas através de visitas domiciliares, de modo que os dados fossem coletados *in loco*, evitando erros, já que os dados são fornecidos pelo responsável familiar. Para aumentar a confiabilidade dos dados referente aos endereços das famílias, os entrevistadores são estimulados a solicitarem o comprovante de endereço, como a fatura de fornecimento de energia, na qual consta a maioria das informações necessárias. Além disso, se existisse um sistema de validação interativo entre o endereço cadastrado e a base de dados dos endereços, o processo seria mais eficiente.

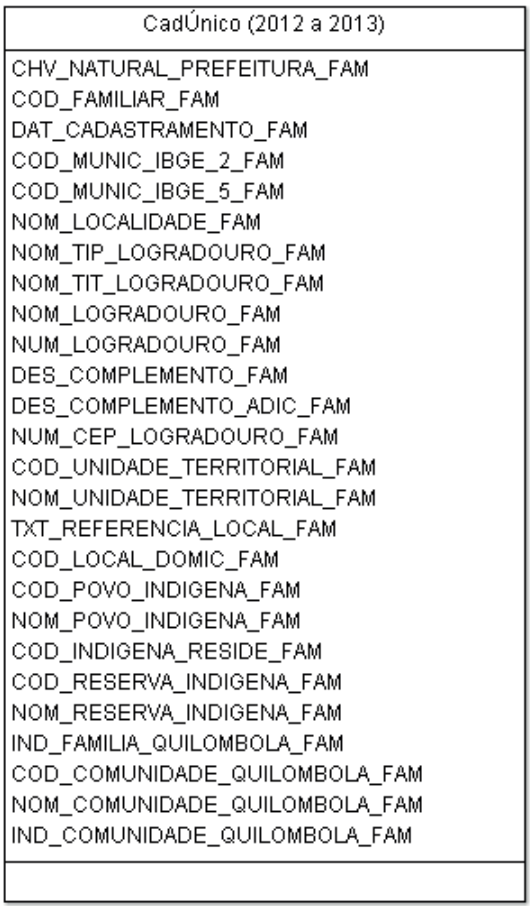
Figura 2 - Bloco do formulário de coleta de dados do Cadastro Único



Fonte: Manual do Entrevistador - Cadastro Único para Programas Sociais, 2011.

Para o seguinte trabalho foi utilizado como referência as informações que constam nos BDG do Cadastro Único referente aos anos de 2012 a 2019. Entretanto, para fins de viabilidade do projeto, esses dados foram convertidos em um BDG único, permitindo que as informações possam ser utilizadas como entrada no algoritmo sem necessidade de grandes adaptações. Os principais atributos considerados para execução do algoritmo estão apresentados no diagrama da Figura 3.

Figura 3 - Diagrama de Classes do CadÚnico nos anos de 2012 e 2013



O Diagrama de Classes do CadÚnico para os anos de 2014 a 2019 é semelhante aos dos anos de 2012 e 2013 com o acréscimo do atributo DAT_ATUALIZACAO_FAMILIA. Neste caso, para o processamento do pareamento, os atributos considerados permanecem os mesmos, alterando apenas sua posição relativa dentro do BDG. A Tabela 1 apresenta a descrição e exemplificação de alguns dos principais atributos que compõem o BDG analisado.

Tabela 1 - Principais atributos de composição dos endereços do Cadastro Único

Atributo	Descrição	Exemplo
NOM_LOCALIDADE_FAMILIA	Localidade	bairro, povoado, vila, etc.

NOM_TIP_LOGRADOURO_FAM	Tipo	rua, avenida, igarapé, etc.
NOM_TIT_LOGRADOURO_FAM	Título	general, santa, professor, etc.
NOM_LOGRADOURO_FAM	Logradouro	"Antônio Carlos", "Carlos Luz", etc.
NUM_LOGRADOURO_FAM	Número	“196”, “37”, etc.
DESC_COMPLEMENTO_FAM	Complemento	apartamento, casa, sobrado, etc.
NUM_CEP_LOGRADOURO_FAM	CEP - Cód. de Endereçamento Postal	“31270-400”, “30493-175”

3.2 Cadastro Nacional de Endereços para Fins Estatísticos (CNEFE)

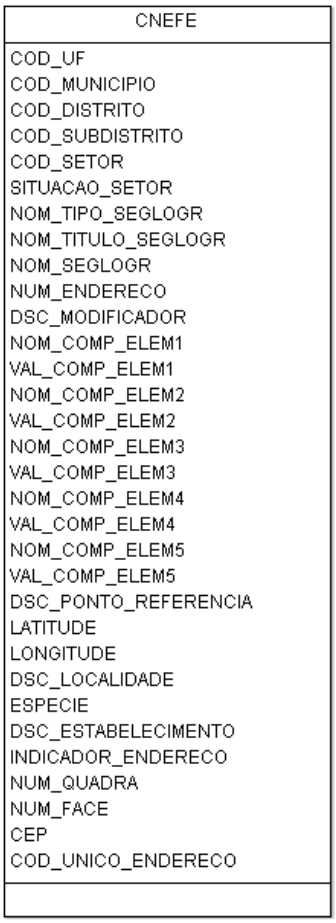
O Cadastro Nacional de Endereços para Fins Estatísticos (CNEFE) é um banco de dados com interface espacial implementado pelo IBGE para uso das pesquisas domiciliares, como o Censo Demográfico, a Pesquisa Nacional por Amostra de Domicílios (PNAD), Pesquisa de Orçamentos Familiares (POF) dentre outras (IBGE: 2013).

O banco do CNEFE contém diversos tipos de endereços no Brasil e proporciona a geração de informações bastante inovadoras e significativas para as pesquisas socioeconômicas em nível domiciliar, pois possui seus logradouros georreferenciados com base nos setores censitários, que são as áreas de coleta das pesquisas domiciliares do IBGE (IBGE: 2010). Todos os endereços do banco do CNEFE se encontram georreferenciados, na área urbana por face de logradouro e na área rural por coordenada da edificação. Além disso, este cadastro contém outros elementos que integram os endereços mapeados pelo IBGE, como número de porta, complemento e CEP, relacionados às quadras e às faces. No caso das zonas rurais, no qual muitas delas possuem um CEP único, os

endereços estão referenciados pelo nome da localidade, sendo este um aglomerado rural ou zonas em que houve uma apuração do censo vigente, como as áreas de proteção ambiental ou terras indígenas. Toda essa integração de dados é realizada pelo Sistema de Mapeamento da Base Territorial do IBGE, conhecido como SISMAP.

O diagrama da Figura 4 apresenta os principais atributos constituintes do BDG do CNEFE.

Figura 4 - Diagrama de Classes do CNEFE



A Tabela 2 abaixo apresenta os principais atributos, para o projeto, que compõem os endereços identificados pelo CNEFE.

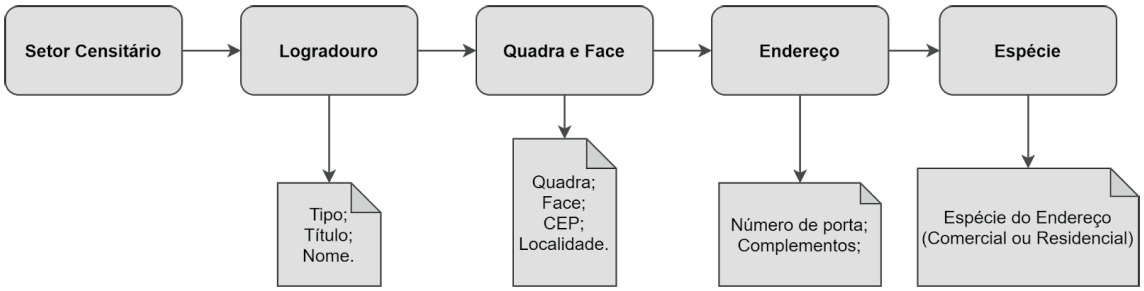
Tabela 2 - Principais atributos de composição dos endereços do CNEFE

Atributo	Descrição	Exemplo
NOM_TIPO_SEGLOGR	Tipo	rua, avenida, igarapé, etc.
NOM_TITULO_SEGLOGR	Título	general, santa, professor, etc.
NOM_SEGLOGR	Logradouro	"Pedro Pinto", "Afonso Pena", etc.
NUM_ENDERECO	Número	“237”, “46”, etc.
NOM_COMP_ELEM1	Complemento	apartamento, casa, sobrado, etc.
VAL_COMP_ELEM1	Valor do complemento	“1, “A-1”, etc...
CEP	CEP - Cód. de Endereçamento Postal	“31270-400”, “30493-175”

Cada endereço cadastrado no CNEFE possui um identificador único, no qual a chave primária é a unidade cadastrada ou Unidade Visitada - UV (i.e., domicílio, estabelecimento).

Essa chave de identificação está atrelada ao setor censitário original, e, caso um novo setor seja criado mediante subdivisão ou mudança de subordinação, o histórico é registrado. Nesse sentido, apenas se a UV deixar de existir (i.e., demolição para implantação de infraestrutura urbana), é que o endereço será excluído do banco de dados do CNEFE.
A Figura 5 apresenta as entidades e principais elementos que compõem o banco de dados do CNEFE.

Figura 5 - Esquema de funcionamento do BDG do CNEFE



Fonte: SKABA, 2017.

4. Similaridade

4.1 Conceito de Similaridade

A similaridade é o método utilizado para o pareamento dos dados entre os BDG do CadÚnico e do CNEFE, de modo a estabelecer uma relação entre as instâncias, gerando um novo conjunto Q'.

Para se criar o conjunto Q' é preciso estabelecer adequadamente o sub-conjunto referente à interseção entre os dois bancos de dados iniciais. A relação entre as representações desses BDG não são, necessariamente, uma identidade. Logo, a definição de quão similar as representações são é crucial para identificar aquelas que podem ser compreendidas como referindo-se à mesma feição.

4.2 Métricas de Similaridade

A literatura dispõe de várias propostas de funções que permitem quantificar a métrica de *strings*. Dentre as opções mais usuais, buscou-se analisar potenciais candidatos para a função de similaridade, de modo que fossem capazes de viabilizar uma avaliação percentual. Assim, tem-se as apresentadas por Santos et al (2017): Jaro, Jaro-Winkler, Jaccard e Dice.

Visando estabelecer a função de similaridade que melhor atenda às demandas desse trabalho, analisou-se as opções quanto ao atendimento do espaço métrico e ao seu comportamento perante o conjunto de dados amostral.

4.2.1 Métrica do Jaro

A distância de Jaro baseia-se na ordem e na quantidade de caracteres comuns entre duas strings, adequando-se a *strings* de mesma cardinalidade (JARO: 1989). A métrica de similaridade de Jaro é dada pela Equação 5.

$$d_j(\varphi_1, \varphi_2) = \begin{cases} 0, & se\ m = 0 \\ \frac{1}{3} \times \left(\frac{m}{|\varphi_1|} + \frac{m}{|\varphi_2|} + \frac{m-t}{m} \right), & se\ m \neq 0 \end{cases} \quad (5)$$

onde, m é o número de caracteres iguais dentro da janela de busca, t é o número de transposições necessárias e $|\varphi_1|$ e $|\varphi_2|$ são as normas – quantidade de caracteres – da *string* φ_1 e φ_2, respectivamente.

4.2.2 Métrica do Jaro-Winkler

A distância de Jaro-Winkler (WINKLER: 1990), por sua vez, é uma versão refinada da métrica de Jaro, dada pela Equação 6.

$$d_w(\varphi_1, \varphi_2) = d_j(\varphi_1, \varphi_2) + (l \times p \times (1 - d_j(\varphi_1, \varphi_2))) \quad (6)$$

onde, d_j é a distância de Jaro, l é o comprimento do prefixo comum da *string*, sendo no máximo quatro caracteres e p é uma constante de valor 0,1.

4.2.3 Métrica do Jaccard

A distância de Jaccard é calculada entre conjuntos de n-gramas (JACCARD: 1912), dada pela Equação 7.

$$d_{jc}(\varphi_1, \varphi_2) = \frac{|N_1 \cap N_2|}{|N_1 \cup N_2|} \quad (7)$$

onde $N_1 = \{\varphi_1\}$ e $N_2 = \{\varphi_2\}_{1 \leq j \leq |\varphi_2|}$ são os conjuntos de n-gramas das *strings* φ_1 e φ_2, respectivamente, com $|\varphi_1|$ e $|\varphi_2|$ correspondendo ao número de n-gramas formados em cada string.

4.2.4 Métrica de Dice

A distância de Dice mede a similaridade de acordo com o índice de Jaccard (JACCARD: 1912), com base no cálculo dos bigramas de caracteres adjacentes (DICE: 1945). Assim, dadas duas *strings* φ_1 e φ_1, o coeficiente de similaridade entre elas é dado pela Equação 8.

$$d_d(\varphi_1, \varphi_2) = \frac{2x\,|n_1 \cap n_2|}{|n_1| + |n_2|} \quad (8)$$

em que $|n_1 \cap n_2|$ são os bigramas comuns entre as duas *strings* e $|n_1|$ e $|n_2|$ representam a cardinalidade, isto é, o número total de bigramas das *strings* φ_1 e φ_2, respectivamente.

4.3 Espaço Métrico

Sobre o espaço métrico pode-se dizer que dado (X, d), X é um conjunto não vazio que possui uma métrica qualquer d, onde $d : X \times X \rightarrow \mathbb{R}$, associa a cada par ordenado $(r_1, r_2) \in X \times X$, um número real $d(r_1, r_2)$, chamado de distância de r_1 a r_2. Logo, a métrica d deve satisfazer as seguintes propriedades para que se tenha um espaço métrico:

i. $d(r_1, r_2) \geq 0$: positividade;

ii. Se $r_1 = r_2 \Rightarrow d(r_1, r_2) = 0$: identidade;

iii. $d(r_1, r_2) = d(r_1, r_2)$: simetria;

iv. $d(r_1, r_3) \leq d(r_1, r_2) + d(r_2, r_3)$: desigualdade triangular.

Como a função de similaridade é dada por $S \geq 1 - d$, dado duas representações r_1 e r_2, tem-se que $S(r_1, r_2) \geq 1 - d(r_1, r_2)$. Neste caso, a função S deve possuir as seguintes propriedades:

i. $0 \leq S(r_1, r_2) \leq 1$: positividade;

ii. Se $x \approx y \Rightarrow S(r_1, r_2) \geq L$: similaridade;

iii. $S(r_1, r_2) = S(r_1, r_2)$: simetria;

iv. $1 + S(r_1, r_3) \geq S(r_1, r_2) + S(r_2, r_3)$: desigualdade triangular.

4.4 Comparação entre os métodos

Visando estabelecer uma métrica de *string* - dentre as especificadas - a ser utilizada neste trabalho, foi realizada uma comparação entre os métodos apresentados, de modo a determinar a que melhor atende o espaço métrico e a avaliação de similitude entre as feições. Assim, sejam as análises demonstradas na Tabela 3.

Tabela 3 - Avaliação das métricas quanto às propriedades do espaço métrico

Método	Positividade	Simetria	Desig. Triangular	Ident.
Jaro	✔	✔	✔	✗
Jaro-Winkler	✔	✔	✔	✗
Jaccard	✔	✔	✔	✗
Dice	✔	✔	✔	✗

A partir dos resultados apresentados na Tabela 3 e pelas formulações das métricas apresentadas neste trabalho, é evidente que as propriedades de positividade, simetria e desigualdade triangular serão atendidas. Contudo, percebe-se que, a princípio, nenhum método atende a todas as propriedades do espaço métrico, devido à propriedade da identidade. Cabe, então, uma discussão acerca do comportamento dos métodos em relação a essa propriedade.

Sejam as *strings* abaixo e as análises demonstradas na Tabela 4.
φ_1 = 'rua Brasil'; φ_2 = 'beco Tefer'

Tabela 4 - Análise da propriedade da identidade

Método	Jaro	Jaro-Winkler	Jaccard	Dice
$d(\varphi_1, \varphi_1)$	1,000	1,000	1,000	1,000

Segundo a propriedade da identidade, quando comparadas duas *strings* iguais, $d(\varphi_1, \varphi_1)$, o resultado obtido pela métrica deve ser 0, indicando um afastamento nulo entre as *strings*. Entretanto, pelos métodos apresentados e mostrados na Tabela 4, obtêm-se valores opostos quanto a essa propriedade. Isso ocorre devido ao fato de que esses métodos não quantificam o afastamento entre as *strings*, mas a própria similaridade em si. Com isso, implica-se que esses métodos não podem ser considerados métricos, por não atenderem a propriedade da identidade.

Dessa forma, pode-se considerar esses métodos como possíveis candidatos à função de similaridade, desde que a dissimilaridade entre duas *strings* completamente diferentes seja 0. Então, dadas duas *strings*, a Tabela 5 avalia seus resultados em cada um dos métodos.

Ressalta-se, entretanto, que neste caso as representações propostas, φ_1 e φ_2 dissimilares.

Tabela 5 - Análise da dissimilaridade entre *strings*

Método	Jaro	Jaro-Winkler	Jaccard	Dice
$d(\varphi_1, \varphi_2)$	0,422	0,422	0,000	0,000

Percebe-se, pela análise da Tabela 5, que os métodos de Jaro e Jaro-Winkler não possuem seus valores variando dentro do intervalo normalizado [0, 1], uma vez que não apresentam o valor 0 para a dissimilaridade. Com isso, percebe-se que apenas os métodos de Jaccard e Dice são normalizados, pois atendem à dissimilaridade.

Sabendo-se que os métodos de Jaccard e Dice são os únicos que estão normalizados, dentre os métodos selecionados, cabe, então, discutir suas características quanto à aplicabilidade ao trabalho proposto. Assim, sejam as *strings* e análises abaixo.
φ_3 = 'ama'; φ_4 = 'mama'; $d_d(\varphi_3, \varphi_4) = 0{,}800$; $d_{jc}(\varphi_3, \varphi_4) = 0{,}000$

A partir dos resultados obtidos acima e sabendo que o método de Jaccard considera a interseção entre n-gramas, ele acaba sendo pouco sensível e apresentando valores destoantes em *strings* praticamente idênticas compostas por uma única palavra. Isso ocorre, pois esse método é baseado em *tokens*, considerando as *strings* de entrada como palavras separadas por espaços para realizar a comparação.

Desse modo, o método de Jaccard, pelo contra-exemplo apresentado, não se apresenta como conveniente para ser aplicado neste trabalho, pois ele não seria capaz de avaliar, com precisão, a similaridade de *strings* com pequenos desvios de escrita, por exemplo, o que não permitiria a integração desses endereços.

Diante do exposto, neste trabalho o método que atende às premissas propostas é, portanto, o método de Dice.

Para compreender o funcionamento do algoritmo de Dice, sejam as duas *strings* abaixo, referentes a endereços reais do município de Contagem, em Minas Gerais.

- φ_5 = 'rua␣antônio␣josé␣costa'
- φ_6 = 'rua␣antônio␣josé␣costinha'

A partir das *strings* φ_5 e φ_6, obtém-se os seguintes bigramas *B*:

- *B* (φ_5) = {ru, ua, a␣, ␣a, an, nt, tô, ôn, ni, io, o␣, ␣j, jo, os, sé, é␣, ␣c, co, os, st, ta}
- *B* (φ_6) = {ru, ua, a␣, ␣a, an, nt, tô, ôn, ni, io, o␣, ␣j, jo, os, sé, é␣, ␣c, co, os, st, ti, in, nh, ha}
- *B* ($\varphi_5 \cap \varphi_6$) = {ru, ua, a␣, ␣a, an, nt, tô, ôn, ni, io, o␣, ␣j, jo, os, sé, é␣, ␣c, co, os, st}

De posse dos bigramas *B* (φ_5) e *B* (φ_6), tem-se a seguinte cardinalidade:

- $|B(\varphi_5)| = 21$
- $|B(\varphi_5)| = 24$
- $|B(\varphi_5 \cap \varphi_6)| = 20$

Nesse caso,

$$d_d(\varphi_5, \varphi_6) = \frac{2x\,|n_1 \cap n_2|}{|n_1| + |n_2|} = \frac{2x\,|B(\varphi_1 \cap \varphi_2 c|}{|B(\varphi_1)| + |B(\varphi_2)|} = \frac{2\,x\,20}{21 + 24} = 0{,}8889$$

Percebe-se que, embora sejam strings muito parecidas e, portanto, com alto valor de similaridade (0,8889), trata-se de feições distintas. Isso demonstra a necessidade em analisar o limiar de aceitação capaz de maximizar as correspondências entre os BDG com o menor erro possível.

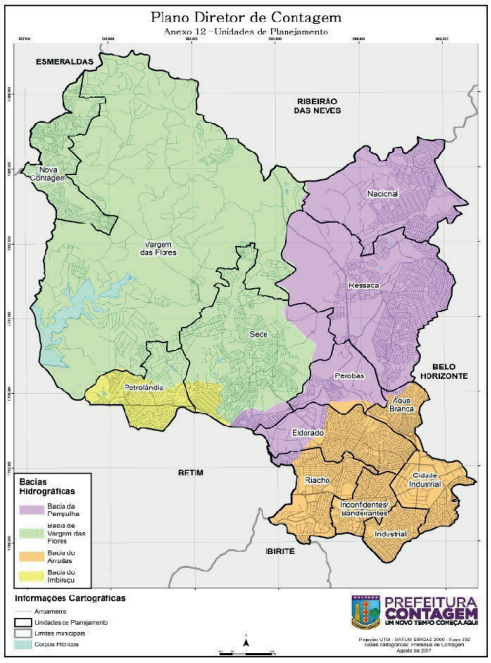
4.5 Limiar de Aceitação

Com a função de similaridade previamente definida, deve-se estabelecer um limiar de aceitação *L*, valor pelo qual deve ser julgada a similitude entre as *strings*. Visando estabelecer um limiar que melhor atenda a similitude entre representações, foi analisada um espaço amostral utilizando a função de similaridade a partir da métrica de Dice.

4.5.1 Espaço Amostral

A fim de se determinar o limiar de aceitação, utilizou-se a rede viária do município de Contagem, localizado no estado de Minas Gerais, com uma população estimada de 663.855 habitantes, distribuída em uma área de 194,746 km^2 (IBGE: 2020). Tal escolha se deve ao fato de o município estar localizado na Região Metropolitana de Belo Horizonte, permitindo que os pesquisadores possam realizar trabalhos de conferências *in loco*, caso fosse necessário. A Figura 6 apresenta o mapa representativo do município de estudo.

Figura 6 - Mapa representativo do município de Contagem



Fonte: Câmara Municipal de Contagem, 2017 (http://www.cmc.mg.gov.br/).

Para o estudo do limiar de aceitação utilizou-se o cadastro geral de endereços (*trechok*) do município, fornecido pela Secretaria de Planejamento. A partir da Classe de Endereços originais, selecionou-se apenas 5 (cinco) atributos dentre os 16 (dezesseis) existentes no BD para se conduzir a análise da similaridade (Figura 7).

Figura 7 - Diagrama de Classe simplificado do cadastro geral de endereços de Contagem

O banco é composto por 4.681 instâncias distintas e seus atributos foram selecionados do *trechock* como explicitados a seguir:

- **id:** refere-se à chave primária dentro da classe, sendo um número inteiro sequencial identificador da tupla correspondente dentro do Banco;
- **tipo:** refere-se a um código padronizado que estabelece as condições de uso de um determinado espaço físico (rua, avenida, beco, alameda etc). Foi selecionado a partir do atributo TIPOLOG;
- **logradouro:** identifica a representação da feição linear do terreno (rede viária). Foi selecionado a partir do atributo LOGOFICIAL;
- **bairro:** regionalização existente em um município. Foi selecionado a partir do atributo BAIRROPRIN;

- **município:** espaço territorial existente em um determinado estado, subordinado ao poder político de uma prefeitura - identifica um determinado *locus* geográfico -, no caso possui uma única instância (“Contagem”).

De posse desses atributos, define-se a composição ideal, a qual cria-se uma string única, sem espaço entre os caracteres e sem os sinais diacríticos, visando permitir a entrada do dado na função de similaridade S . A retirada dos espaços entre os caracteres e dos sinais diacríticos reduz a quantidade de bigramas a serem analisados pela métrica de Dice. Consequentemente, tem-se uma redução nos valores obtidos pela função de similaridade S, tornando o processo mais conservador. A composição definida contém os atributos da seguinte forma:

- **Composição P (*tipo, logradouro*)** = *tipologradouro*, sem espaços internos. Conforme pode ser observado no exemplo abaixo:
- **P (*rua, antônio josé costa*)** = *ruaantoniojosecosta*

O atributo município não foi considerado na Composição P devido ao fato do objeto de estudo se limitar somente ao município de Contagem. Já o atributo bairro será utilizado pelo algoritmo, juntamente com o Código de Endereçamento Postal (CEP), para filtrar e limitar a análise da função de similaridade S somente ao mesmo *locus* geográfico, não sendo incluído na string de análise de similaridade.

4.5.2 Cálculo do Limiar

Para avaliar o comportamento da função de similaridade S de acordo com as diferentes composições de *strings*, calculou-se os resultados obtidos pela métrica de Dice, utilizando o espaço amostral do município de Contagem conforme o Algoritmo 1. Neste caso, considerou-se o banco original em duplicidade, denominados como T_1 e T_2.

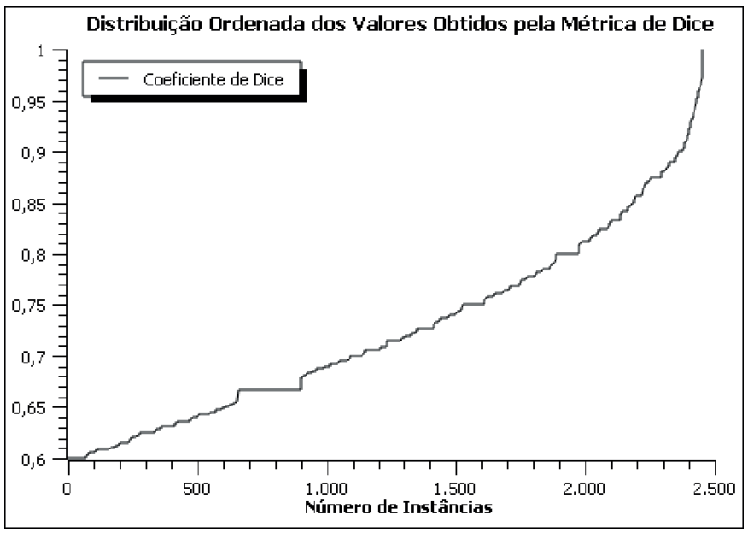
O Algoritmo 1 gera uma lista sequencial contendo em suas tuplas a identificação no banco T_1 (id_1), a identificação no banco T_2 (id_2) e o valor obtido pela função de similaridade (*Dice*($T_1[id_1]$), $T_2[id_2]$). Ressalta-se que a resposta obtida é ordenada a partir do valor de similaridade calculado (função **Ordenar**). Dessa forma, o registro final será o máximo valor obtido por *Dice*($T_1[id_1]$), $T_2[id_2]$. No caso de a instância possuir apenas um caractere, a função de similaridade atribui o valor 0 . Assim, procura-se ser conservador quanto à associação dos valores de similaridade aos pares das instâncias avaliadas e à identificação de possível similitude.

Algoritmo 1 - Pseudocódigo de Dice

Input: T_1 e T_2
Output: Matriz contendo tuplas contruídas a partir dos pares id_1, id_2 com respectivos valores de similaridade
begin
    $resposta = [\,]$
    $n \leftarrow |T_1|$ `/* |T1| = cardinalidade de T1 */`
    $m \leftarrow |T_2|$ `/* |T2| = cardinalidade de T2 */`
    for $id_1 \leftarrow 1$ ***to*** n **do**
        for $id_2 \leftarrow 1$ ***to*** m **do**
            $max \leftarrow 0$
            if $|T_1[id_1]|$ ***and*** $|T_2[id_2]| \neq 1$ **then**
                $valor \leftarrow$ `Dice`$(T_1[id_1], T_2[id_2])$
                if $valor \geq max$ **then** $max \leftarrow valor$
            end
        end
        $resposta.append([id_1, id_2, max])$
    end
    `Ordenar`$(resposta)$
end
return *resposta*

Para determinar o limiar de aceitação, deve-se realizar uma análise conjunta da distribuição dos valores obtidos pela métrica de Dice no espaço amostral, conforme gráfico da Figura 8, e da qualidade dos resultados obtidos, por meio da avaliação dos percentuais de aproveitamento, como mostrado na Tabela 6.

Figura 8 - Coeficiente de Dice ordenado conforme a Composição *P*



O gráfico da Figura 8 apresenta os valores ordenados obtidos pela função de similaridade S, desenvolvida conforme Algoritmo 1, sendo que para tal considerou-se apenas os valores iguais ou superiores a 0,6 para o coeficiente de Dice. Tal divisão é justificável pelo fato do limiar de aceitação estar dentro desse intervalo, tendo em vista que os valores abaixo de 0,6 apresentam pareamentos bastante destoante entre as *strings*, do ponto de vista analítico.

Tabela 6 - Avaliação de qualidade da métrica de Dice para a composição *P*

Atributo		Classificação			Percentual		Registrados
Limite inferior	Limite superior	A	B	C	Positivo	Negativo	
0.60	0.65	5	25	577	4,942%	95,058%	607
0.65	0.70	20	12	449	6,653%	93,347%	481
0.70	0.75	39	26	371	14,908%	85,092%	436
0.75	0.80	71	37	254	29,834%	70,166%	362
0.80	0.85	107	39	146	50,000%	50,000%	292
0.85	0.90	61	49	71	60,773%	39,227%	181
0.90	0.95	30	20	19	72,464%	27,536%	69
0.95	1.00	22	4	1	96,296%	3,704%	27
Totais		355	212	1.888	-	-	2455

A Tabela 6 apresenta a classificação de qualidade do pareamento entre as *strings* do BDG do município de Contagem - MG. Essa classificação foi realizada conforme os seguintes parâmetros: A - as *strings* se referem à mesma entidade (o pareamento é considerado correto); B - as *strings* não permitem assegurar a identidade, mas permite inferir uma probabilidade de se referirem à mesma feição (o pareamento é considerado provável); C - as *strings* não se referem à mesma entidade (o pareamento é considerado equivocado). Os resultados foram compilados percentualmente, sendo que para os resultados positivos foram consideradas as classificações A e B e para os negativos a classificação C.

É possível identificar, pela análise da Figura 8, uma mudança no comportamento da curva com a Métrica de Dice igual a 0,67, o que permite inferir que o limiar de aceitação deve partir desse valor. Concomitantemente, a Tabela 6 mostra que no intervalo [0,65, 0,70] a quantidade de erros é maior que a quantidade de acertos no pareamento das *strings*, o que não torna interessante utilizar um limite de Dice igual a 0,67.

Pela análise da Tabela 6, é possível perceber que a partir do coeficiente de Dice igual a 0,85 tem-se uma quantidade de acertos maior que a de erros. E, pela Figura 8, a partir desse valor, os dados ficam exponencialmente mais aderentes. Além disso, a análise dos percentuais de aproveitamento permite visualizar que valores superiores a 0,95 garantem um aproveitamento de 96,296 %, embora a custo de um menor número de registros. Com isso, é possível estabelecer que o limiar de aceitação deve ser igual ou superior a 0,85 ($L \geq 85\%$), sendo escolhido conforme a demanda solicitada, levando em consideração a porcentagem de aproveitamento e o número de registros.

5. Algoritmo

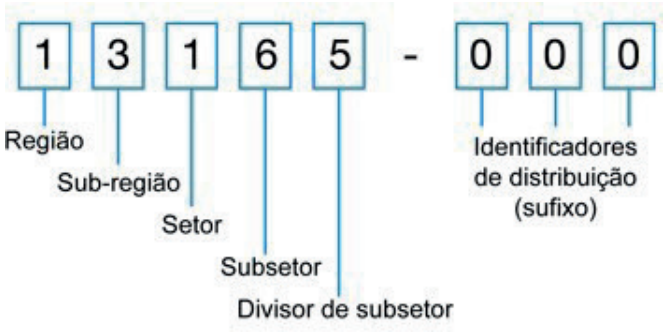
O Algoritmo 2 em pseudocódigo realiza a pareamento das instâncias e possui como entrada as tabelas *C* e *N* que são os conjuntos de dados referentes ao CadÚnico e do CNEFE, respectivamente. A ideia central do algoritmo é identificar a melhor semelhança entre as instâncias das famílias presentes na tabela *C* com o os endereços presentes na tabela *N.*

Em virtude do código implementado, cada instância de *C* é reduzida a um Id (f_i . *id*) - funcionando como Chave Primária -, o tipo do endereço e o nome do logradouro que agregados integram uma string (*rb_cad*) e o número da residência (*n_cad*), obtidos pelos atributos NOM_TIP_LOGRADOURO_FAM, NOM_LOGRADOURO_FAM e NUM_LOGRADOURO_FAM, respectivamente.

Analogamente, as instâncias do CNEFE foram reduzidas a um Id específico (e_j . *id*) - funcionando como Chave Primária -, o tipo do endereço e o nome do logradouro que juntos estarão presentes em uma *string* (*rb_cnefe*) e o número da residência (*n_cnefe*), obtidos pelos atributos NOM_TIPO_SEGLOGR, NOM_SEGLOGR e NUM_ENDERECO, respectivamente.

Em linhas gerais para cada instância f_i em C (linha 3) encontra-se uma melhor instância e_j em *N*′ que possua o nome do bairro, rua, e número o mais similar possível. Tem-se que *N*′ é um subconjunto de *N*, gerado pela função **get_endereco_cep** cujos endereços relacionados estejam na mesma cidade e possuam os mesmos quatro primeiros dígitos do CEP, fato que garante a localização dos endereços no mesmo subsetor de uma cidade (Figura 9).

Figura 9 - Esquema do CEP



Fonte: Correios[1].

Para computar a similaridade entre os endereços das instâncias a função **dice_coeficiente** recebe como entrada as *strings* com o nome do bairro e da rua das instâncias f_i e e_i. As combinações com os melhores coeficientes é armazenada (linha 20) e, em seguida, a qualidade da resposta é avaliada (linhas 22-45) ao se atribuir ao par identificado um grau de precisão que vai de 0 a 5 (apresentado na seção 7.4).

Após identificar o melhor pareamento entre f_i e e_i, o algoritmo gera uma tupla *n*, contendo o Id da família do CadÚnico (f_{id}), o *id* do melhor endereço (e_{id}) do CNEFE e o coeficiente de Dice (função de similaridade) entre essas instâncias e a qualificação do pareamento. Cada uma das tupla é armazenada em um conjunto resposta T , cuja codificação é apresentada na seção 7.3.

1 https://www.correios.com.br/enviar-e-receber/ferramentas/cep/estrutura-do-cep, capturado em 13 de dezembro de 2020.

Algoritmo 2 - Método de correlação das Tabelas

Input: C: Tabela do Cadastro Único. N: Tabela do CNEFE
Output: T: Tabela resposta.

```
i ← 0                                          /* Contador usado para percorrer as famílias de C' */
T ← {}
for f_i ∈ C do
    cid_cad ← f_i.cidade                                  /* Código da cidade da família f_i */
    rb_cad ← f_i.endereco                          /* String - nome bairro e rua do CadÚnico */
    n_cad ← f_i.numero                              /* Número da casa no Cad. Único */
    rb_cnefe ← null                                /* String com nome bairro e rua do CNEFE */
    n_cnefe ← null                                          /* Número da casa em N */
    f_id ← f_i.id                                  /* ID da família do Cad. Único */
    e_id ← null                                         /* ID do endereço do CNEFE */
    c ← 0                                      /* Classificação da qualidade do Pareamento */
    j ← 0                                  /* Contador usado para percorrer os endereços de N' */
    p_cep ← f.cep div 1000                          /* Obtém 4 primeiros digitos do CEP */
    N' ← get_enderecos_cep(N, p_cep, cid_cad)               /* Retorna os endereços */
    dc ← dice_coeficente(end_cnefe, end_cad)                        /* Calcula DICE */
    while not dc = 1 and n_cad = n_cnefe do
        rb_cnefe ← e_j ∈ N'
        new_dc ← dice_coeficente(rb_cad, rb_cnefe)
        if dc < new_dc then
            dc ← new_dc
        end
        if dc ≥ 0.95 and n_cad = n_cnefe and n_cad ∉ {null, 0} and n_cnefe ∉ {null, 0} and c ≤ 5 then
            c ← 5
            e_id ← e_j.id
        end
        if dc ≥ 0.95 and n_cad = n_cnefe and n_cad ∈ {null, 0} and n_cnefe ∈ {null, 0} and c ≤ 4 then
            c ← 4
            e_id ← e_j.id
        end
        if dc ≥ 0.95 and n_cad ≠ n_cnefe and n_cad ∉ {null, 0} and n_cnefe ∉ {null, 0} and c ≤ 3 then
            c ← 3
            e_id ← e_j.id
        end
        if dc < 0.95 and n_cad = n_cnefe and n_cad ∉ {null, 0} and n_cnefe ∉ {null, 0} and c ≤ 2 then
            c ← 2
            e_id ← e_j.id
        end
        if dc < 0.95 and n_cad = n_cnefe and n_cad ∈ {null, 0} and n_cnefe ∈ {null, 0} and c ≤ 1 then
            c ← 1
            e_id ← e_j.id
        end
        if dc < 0.95 and n_cad ≠ n_cnefe and n_cad ∈ {null, 0} and n_cnefe ∈ {null, 0} then
            c ← 0
            e_id ← e_j.id
        end
        i ← i+1
    end
    j ← j+1
    n ← (f_id, e_id, dc, c)
    T ← T ∪ {n}
end
return T
```

5.1 Especificações Técnicas de Aplicativos utilizados

Para o desenvolvimento do trabalho foram utilizados os seguintes aplicativos:

- PostgreSQL 12.x;
- Sistema operacional Windows 10;
- Python 3.6.

Obs.: A importação do BDG foi operacionalizada pelo Sistema Gerenciador de Banco de Dados (SGBD) pgAdmin do próprio PostgreSQL. Para mais informações acessar o link: https://blog.tecnospeed.com.br/backup-e-restore-postgresql/

5.2 Bases de Entrada de Dados

A base do CNEFE encontra-se subdividida pelas unidades da federação. Assim, cada um dos estados corresponde a um nome de arquivo contendo o termo "cnefe" acrescido da sigla da Unidade da Federação e o código do estado de acordo com o IBGE, conforme exemplificado abaixo:

- Base de dados do CNEFE dos estados de São Paulo e Roraima:
    - cnefe_sp_35.backup;
    - cnefe_rr_14.backup.

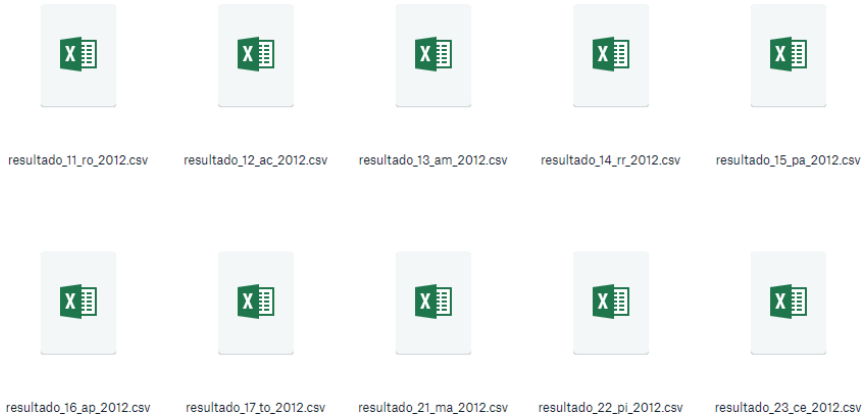
A base do CadÚnico encontra-se com instâncias de todo o país. Entretanto, há um banco para cada ano, ou seja, há 8 (oito) BDG referentes ao CadÚnico, considerando os anos 2012 a 2019. Encontram-se esquematizados, com o termo "cad_unic" acrescido do ano em questão, conforme exemplo abaixo:

- Base de dados do CadÚnico do ano de 2019:
    - cad_unic_2019.backup.

5.3 Tabelas de Resultados

As tabelas de resultado (Figura 10) encontram-se armazenadas no esquema principal do PostgreSQL *public* e possui uma codificação para o nome do arquivo correspondente ao registro do processamento. O termo "resultado" é acrescido ao código do estado segundo o IBGE[2], a sigla da Unidade da Federação e o ano referente ao CadÚnico. Dessa forma, as tabelas de resultados relacionam para cada estado, sua tabela do CNEFE a uma tabela do CadÚnico de um determinado ano. Por exemplo, no arquivo "resultado_41_pr_2012.csv" tem-se a junção da tabela do CNEFE para o estado do Paraná (PR) com a tabela do CadÚnico do ano de 2012.

Figura 10 - Exemplo de Tabelas Respostas



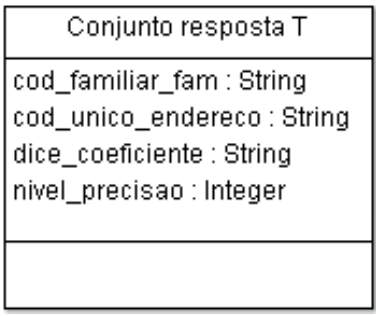
Cada uma das tabelas resultados apresentam os seguintes atributos (Figura 11):

- cod_familiar_fam
    - Código familiar presente nas bases do CadÚnico.
- cod_unico_endereco
    - Código do endereço presente nas bases do CNEFE.

2 https://www.ibge.gov.br/explica/codigos-dos-municipios.php, capturado em 14 de dezembro de 2020.

- dice_coeficiente
    - Representa o grau de similaridade (%) entre os endereços da base do CNEFE e do CadÚnico (exceto o número da casa). Esse coeficiente é gerado pelo algoritmo de Dice (seção 6.2.1).
- nível_precisao
    - nível de precisão do pareamento obtido.

Figura 11 - Diagrama de Classe da Tabela Resultado



5.4 Nível de Precisão

O processo de pareamento entre instâncias de duas bases de dados pode apresentar níveis de precisão diferentes, uma vez que o algoritmo de Dice apresenta o grau de similaridade entre ruas, avenidas, estradas, rodovias, bairros e distritos que compõem o endereço em porcentagem e, caso fique abaixo do limiar de aceitação, pode produzir pareamentos anômalos.

Em paralelo ao grau de similaridade dos tipos de logradouros presentes nas bases do CadÚnico e do CNEFE, é levado em consideração os números das residências que compõem o endereço. Entretanto, podem haver instâncias do CadÚnico e do CNEFE em que o número residencial não esteja presente como parte dos endereços.

Tendo em vista esse cenário, pode-se ter diferentes níveis de precisão nos pareamentos, por exemplo, pode-se fazer o pareamento entre dois endereços com o nome da rua e do bairro 100% de similares, bem como números residenciais iguais. Em paralelo, pode-se ter o pareamento entre dois endereços com o nome da rua e do bairro 100% similares, porém sem número da residência compatíveis, seja porque esse número é nulo na base do Cadastro Único, ou que esse número pode não existir na base do CNEFE, impossibilitando um pareamento mais preciso. Logo, dependendo do nível de precisão do pareamento de uma dada instância do CadÚnico ou do CNEFE, pode ser rejeitado, aceito com ressalvas ou aceito integralmente pelo analista de dados ou qualquer responsável técnico, levando em consideração as regras de negócio institucionais vigentes.

Com isso, dispor de uma maneira rápida de avaliar os resultados e, consequentemente, ajudar os gestores no processo de aceitação dos pareamentos, foi decidido organizar em grupos os resultados de tal forma que representassem esses níveis. Esses grupos abrangem todos os tipos de pareamento possíveis, levando em consideração o limiar de aceitação de similaridade e os números das residências. Esses níveis de precisão vão de uma escala de 0 a 5, sendo 0 (zero) com menor nível de precisão ou menor qualidade das respostas, e 5 maior nível de precisão ou melhor qualidade na resposta.

Partindo do pressuposto que o estado, a cidade e os quatro primeiros dígitos do CEP de um pareamento entre uma instância do CadÚnico e do CNEFE são iguais, além de considerar o coeficiente de Dice com maior peso na definição dos níveis, tem-se seguinte definição:

- nível 0: $(D < L)$ e $(Eu \neq Ec)$ e $(Eu \neq 0)$ e $(Ec \neq 0)$
- nível 1: $(D < L)$ e $(Eu = Ec)$ e $(Eu = 0)$ e $(Ec = 0)$
- nível 2: $(D < L)$ e $(Eu = Ec)$ e $(Eu \neq 0)$ e $(Ec \neq 0)$
- nível 3: $(D \geq L)$ e $(Eu \neq Ec)$ e $(Eu \neq 0)$ e $(Ec \neq 0)$

- nível 4: *(D ≥ L)* e *(Eu =/ Ec)* e *(Eu = 0 ou NULL)* e *(Ec = 0 ou NULL)*
- nível 5: *(D ≥ L)* e *(Eu = Ec)* e *(Eu =/ 0)* e *(Ec =/ 0)*

Onde,

- *D*= igual ao Coeficiente de Dice calculado para o par estabelecido;
- *L* = limiar de aceitação;
- *Eu*= número do endereço no CadÚnico;
- *Ec*= número do endereço no CNEFE.

Observando a definição acima tem-se que o nível 5 é o melhor, pois o coeficiente de Dice é maior ou igual ao limiar de aceitação e os números residenciais são iguais e não são nulos, apresentando um maior grau de certeza do quão correto o pareamento está. Pareamentos que estão nesse nível podem ser aceitos como pareamentos corretos.

Abaixo do nível 5 tem-se o nível 4, que possui similaridade maior ou igual ao limiar de aceitação, porém os números das residências não são iguais e o número no Cadastro Único é igual a 0 ou nulo. Esses casos tendem a ocorrer em localidades fora da zona urbana. Nessa situação, pareamentos dentro desse grupo podem ser aceitos desde que o gestor avalie que realmente o número nesses casos não é importante.

O nível 3 é similar aos níveis 4 e 5, mas os números das residências não podem ser nulos em ambas as bases. Nesse caso, há grandes chances que o endereço com um número aleatório x qualquer para uma dada instância do CadÚnico não conste na base do CNEFE, nesse caso pode ser aceito desde que o gestor avalie que realmente a numeração nestes casos não é importante ou ainda seja feita uma avaliação se realmente aquela residência existe.

Os níveis 2, 1 e 0 possuem o grau similaridade menor que o limiar de aceitação, consequentemente, são casos que os gestores podem rejeitar, pois os pareamentos que encontram nesses níveis são tidos como não

confiáveis, exigindo que sejam feitas análises mais profundas que podem resultar desde simples ajustes nas bases de dados até mesmo de novos trabalhos de campo para o levantamento dos endereços corretos.

Para efeito de avaliação do total de pareamento aceito, considerou-se como adequado os níveis de precisão 3, 4 e 5. Por conseguinte, sempre que forem identificados níveis de precisão nesses valores, o pareamento foi considerado correto e foi considerado na estatística do processamento.

6. Resultados e Discussão

6.1 Dados do CNEFE

No BDG do CNEFE encontram-se 88.247.900 instâncias, subdivididas pelas Unidades da Federação de acordo com a Tabela 7.

Tabela 7 - Quantidade de instâncias no CNEFE

UF	Número de Instâncias
AC	340.444
AL	1.333.907
AM	250.261
AP	1.401.988
BA	6.875.499
CE	3.579.908
DF	1.093.568
ES	1.773.492

GO	3.127.016
MA	2.581.244
MG	1.567.767
MS	1.224.041
MT	9.634.538
PA	2.917.413
PB	1.829.881
PE	4.928.690
PI	3.881.503
PR	1.425.202
RJ	1.493.961
RN	5.401.889
RO	7.418.919
RR	765.304
RS	205.255
SC	3.345.911
SE	18.196.842
SP	973.409
TO	680.048

6.2 Dados do CadÚnico

No BDG do CadÚnico encontram-se por ano o total de instâncias conforme apresentado na Tabela 8:

Tabela 8 - Quantidade de instâncias no CadÚnico

Ano	Número de Instâncias
2012	25.068.130
2013	27.198.312
2014	29.172.341
2015	27.325.069
2016	26.456.063
2017	26.946.898
2018	26.913.731
2019	28.884.000

A distribuição das instâncias para cada Unidade da Federação por ano pode ser verificada na Tabela 9 abaixo:

Tabela 9 - Distribuição das instâncias do CadÚnico por ano

UF	2012	2013	2014	2015	2016	2017	2018	2019
AC	109.370	122.157	133.925	124.890	123.098	125.106	127.235	138.241
AL	639.675	673.356	707.776	663.918	641.604	646.449	647.960	670.096
AM	496.566	550.839	605.644	577.613	583.440	599.323	618.548	672.788
AP	81.177	87.804	101.301	94.439	98.239	107.371	119.950	137.448
BA	2.867.966	3.057.955	3.255.124	3.042.783	2.961.792	3.031.154	2.996.335	3.168.401

CE	1.717.739	1.795.089	1.883.225	1.794.637	1.759.041	1.787.127	1.769.940	1.865.205
DF	239.918	250.696	265.904	233.888	184.168	174.051	148.848	170.094
ES	420.889	454.131	493.438	459.492	430.497	415.626	414.733	465.352
GO	740.628	822.667	889.579	839.801	788.170	807.198	775.396	837.302
MA	1.352.778	1.461.346	1.560.802	1.456.984	1.460.571	1.477.575	1.467.439	1.540.090
MG	2.574.708	2.745.616	2.933.390	2.622.362	2.451.519	2.501.280	2.457.946	2.631.830
MS	318.694	379.666	432.704	431.407	413.517	401.418	388.699	419.387
MT	394.800	456.145	511.520	499.108	472.683	474.308	473.854	512.928
PA	1.176.231	1.292.222	1.420.561	1.388.542	1.395.281	1.462.972	1.504.898	1.641.310
PB	781.539	832.184	880.471	802.672	780.377	790.890	788.580	836.331
PE	1.804.978	1.905.297	2.003.942	1.837.981	1.792.501	1.844.034	1.885.096	1.998.072
PI	687.762	730.974	770.233	709.782	676.769	670.292	670.155	707.570
PR	1.157.732	1.254.548	1.324.843	1.243.316	1.187.275	1.202.206	1.177.125	1.268.294
RJ	1.422.250	1.570.652	1.725.631	1.604.720	1.541.366	1.593.909	1.633.366	1.822.666
RN	619.790	659.267	697.317	632.827	601.508	600.230	604.802	640.847
RO	223.553	249.893	272.790	244.850	232.711	228.207	225.504	248.109
RR	71.692	78.400	88.039	92.497	91.721	89.239	92.753	100.636
RS	1.033.387	1.150.111	1.249.986	1.111.069	1.015.986	979.621	929.860	994.150
SC	439.013	486.954	532.891	461.245	421.848	395.391	371.528	411.361
SE	403.713	437.337	466.925	444.761	441.787	448.526	459.463	485.370
SP	3.039.042	3.416.986	3.664.677	3.622.697	3.632.216	3.811.135	3.881.379	4.206.308
TO	252.540	276.020	299.703	286.788	276.378	282.260	282.339	293.814

Após a conclusão do processamento, identificou-se uma quantidade de pareamentos nos níveis de precisão igual ou superior a 3, por ano. Visando favorecer a análise da eficiência do processo de pareamento, foram construídas tabelas que apresentam o percentual de pareamento para

limiares de aceitação iguais ou superior a 65% (Tabela 10), 75% (Tabela 11), 85% (Tabela 12) e 95% (Tabela 13).

Tabela 10 - Percentual de pareamento com limiar de 65%

UF	2012	2013	2014	2015	2016	2017	2018	2019
AC	92,407%	96,152%	97,487%	99,070%	99,758%	99,865%	99,879%	99,813%
AL	98,760%	99,154%	99,311%	99,663%	99,830%	99,710%	99,483%	99,433%
AM	90,439%	94,541%	96,417%	98,649%	99,424%	99,579%	99,611%	99,538%
AP	94,749%	95,933%	96,761%	97,820%	97,981%	97,980%	98,138%	98,779%
BA	95,881%	97,106%	97,694%	98,811%	99,126%	99,100%	98,887%	98,614%
CE	96,064%	96,856%	97,179%	97,949%	98,064%	97,980%	97,816%	97,576%
DF	79,676%	81,507%	83,338%	89,060%	91,726%	92,271%	91,356%	90,814%
ES	95,631%	96,660%	97,364%	98,724%	99,136%	99,227%	99,263%	99,303%
GO	82,804%	87,733%	91,076%	96,134%	97,548%	97,245%	97,495%	97,017%
MA	95,681%	97,317%	97,905%	99,265%	99,648%	99,658%	99,667%	99,702%
MG	95,099%	96,464%	97,083%	99,096%	99,431%	99,178%	98,924%	98,451%
MS	96,835%	97,949%	98,372%	99,124%	99,379%	99,247%	99,145%	98,994%
MT	93,200%	95,288%	96,407%	98,687%	99,602%	99,626%	99,694%	99,664%
PA	93,731%	95,371%	95,788%	96,723%	96,935%	96,962%	97,162%	97,175%
PB	94,178%	96,382%	97,291%	99,105%	99,638%	99,811%	99,857%	99,859%
PE	94,904%	95,934%	96,649%	98,021%	98,307%	98,115%	97,785%	97,559%
PI	97,657%	98,275%	98,600%	99,490%	99,792%	99,769%	99,682%	99,673%
PR	96,500%	97,063%	97,521%	98,361%	98,566%	98,640%	98,567%	98,525%
RJ	91,363%	90,811%	91,267%	93,904%	95,106%	95,879%	96,157%	96,079%
RN	99,010%	99,246%	99,415%	99,750%	99,837%	99,889%	99,531%	99,243%
RO	61,353%	72,156%	77,563%	93,563%	98,588%	98,795%	99,119%	98,564%
RR	86,851%	91,832%	95,538%	98,507%	99,548%	99,666%	99,787%	99,767%

RS	93,786%	95,217%	96,068%	97,960%	98,618%	98,653%	98,248%	97,533%
SC	96,315%	97,277%	97,873%	98,954%	99,448%	99,664%	99,720%	99,574%
SE	94,846%	96,989%	97,844%	99,248%	99,660%	99,784%	99,760%	99,824%
SP	95,555%	96,781%	97,724%	98,800%	99,134%	99,027%	98,961%	98,586%
TO	92,678%	94,670%	95,743%	97,521%	97,844%	96,813%	97,543%	97,258%

Tabela 11 - Percentual de pareamento com limiar de 75%

UF	2012	2013	2014	2015	2016	2017	2018	2019
AC	88,488%	93,826%	95,622%	98,007%	98,949%	98,960%	99,242%	98,866%
AL	95,234%	96,675%	97,223%	98,421%	98,944%	98,837%	98,522%	98,509%
AM	84,328%	90,614%	93,460%	97,063%	98,250%	98,363%	98,482%	98,265%
AP	88,507%	91,491%	93,289%	95,601%	96,061%	96,089%	95,764%	96,535%
BA	91,532%	93,660%	94,721%	96,774%	97,357%	97,370%	97,121%	96,800%
CE	92,006%	93,735%	94,487%	95,967%	96,274%	96,312%	96,030%	95,791%
DF	57,806%	61,052%	64,424%	72,656%	77,937%	79,685%	78,253%	78,492%
ES	90,594%	92,797%	94,283%	96,946%	97,817%	98,032%	98,152%	98,249%
GO	75,623%	81,930%	86,144%	92,517%	94,414%	93,591%	94,673%	93,589%
MA	91,674%	94,457%	95,408%	97,482%	97,991%	97,929%	97,925%	97,859%
MG	91,062%	93,468%	94,586%	97,873%	98,503%	98,195%	97,848%	97,346%
MS	91,386%	94,343%	95,555%	97,537%	98,270%	98,194%	98,156%	98,061%
MT	87,603%	90,968%	92,780%	96,392%	97,862%	97,813%	98,235%	98,070%
PA	89,202%	92,163%	93,093%	94,746%	95,206%	95,226%	95,490%	95,432%
PB	90,414%	93,713%	94,946%	97,515%	98,302%	98,543%	98,653%	98,606%
PE	89,532%	91,958%	93,367%	96,077%	96,766%	96,699%	96,428%	96,264%
PI	94,176%	95,557%	96,256%	97,855%	98,451%	98,520%	98,401%	98,386%
PR	91,478%	93,251%	94,551%	96,874%	97,478%	97,616%	97,662%	97,686%

RJ	83,177%	82,651%	83,864%	89,206%	91,983%	93,513%	94,506%	94,360%
RN	95,794%	96,746%	97,413%	98,573%	98,879%	99,108%	98,418%	98,341%
RO	59,341%	69,551%	74,613%	89,892%	94,646%	93,769%	95,130%	93,536%
RR	77,829%	86,144%	91,703%	96,426%	98,160%	98,308%	98,701%	98,510%
RS	89,727%	91,868%	93,157%	96,306%	97,215%	97,190%	96,840%	95,993%
SC	92,698%	94,388%	95,601%	97,838%	98,656%	98,894%	98,990%	98,837%
SE	92,258%	94,909%	95,916%	97,750%	98,185%	97,920%	98,161%	97,791%
SP	90,286%	93,042%	95,170%	97,480%	98,284%	98,117%	98,181%	97,698%
TO	86,245%	89,214%	90,294%	92,862%	93,113%	92,389%	92,580%	93,567%

Tabela 12 - Percentual de pareamento com limiar de 85%

UF	2012	2013	2014	2015	2016	2017	2018	2019
AC	71,058%	79,646%	82,648%	87,073%	88,482%	88,372%	88,832%	88,164%
AL	76,763%	80,479%	81,940%	84,878%	86,134%	84,843%	85,168%	84,168%
AM	65,383%	76,110%	81,092%	87,346%	89,254%	88,759%	89,397%	88,533%
AP	65,138%	73,327%	76,724%	81,638%	82,680%	80,362%	80,824%	80,270%
BA	71,937%	75,664%	77,760%	81,626%	82,797%	82,427%	82,744%	81,751%
CE	74,553%	78,318%	80,113%	83,233%	84,039%	84,281%	84,076%	83,710%
DF	21,705%	24,894%	28,241%	34,794%	40,918%	45,372%	43,210%	45,135%
ES	71,877%	76,741%	80,210%	85,821%	87,917%	87,725%	89,029%	88,784%
GO	58,174%	65,491%	70,238%	77,668%	80,005%	78,380%	80,134%	78,309%
MA	73,729%	79,069%	80,923%	84,456%	84,562%	84,235%	83,790%	84,092%
MG	76,443%	80,835%	83,121%	89,116%	90,324%	89,546%	89,702%	88,600%
MS	74,721%	80,757%	83,528%	87,601%	89,129%	88,520%	89,879%	89,150%
MT	69,360%	74,974%	77,918%	83,859%	86,094%	85,096%	86,460%	85,337%
PA	70,217%	76,040%	78,230%	81,390%	82,341%	82,105%	82,814%	82,423%

PB	73,034%	77,242%	78,749%	81,909%	82,738%	81,384%	83,052%	81,675%
PE	69,414%	74,448%	77,322%	82,288%	83,870%	83,361%	84,908%	83,987%
PI	74,680%	76,542%	77,689%	80,158%	81,067%	80,863%	80,802%	80,543%
PR	74,205%	78,640%	81,983%	87,690%	89,368%	89,398%	90,450%	90,073%
RJ	61,438%	61,479%	64,103%	73,630%	78,730%	81,383%	83,829%	83,504%
RN	80,693%	83,817%	86,010%	89,494%	90,429%	89,515%	89,308%	88,458%
RO	45,778%	52,845%	56,588%	68,219%	71,836%	71,774%	74,186%	73,579%
RR	55,805%	68,254%	77,758%	85,555%	88,554%	88,114%	89,455%	88,780%
RS	73,454%	77,165%	79,611%	85,509%	87,017%	86,374%	86,867%	85,539%
SC	76,915%	80,729%	83,645%	89,419%	91,406%	91,547%	92,622%	92,369%
SE	80,910%	84,435%	85,855%	88,606%	89,186%	89,392%	89,064%	89,174%
SP	72,980%	79,007%	83,852%	88,881%	90,719%	89,729%	90,897%	89,560%
TO	66,399%	70,384%	71,943%	75,873%	76,432%	75,113%	75,680%	75,945%

Tabela 13 - Percentual de pareamento com limiar de 95%

UF	2012	2013	2014	2015	2016	2017	2018	2019
AC	30,393%	36,987%	39,300%	42,573%	43,820%	46,276%	45,680%	46,861%
AL	36,096%	39,701%	40,582%	43,087%	43,709%	43,560%	42,463%	42,992%
AM	29,498%	39,202%	44,327%	50,013%	51,978%	53,345%	53,077%	53,841%
AP	19,945%	26,730%	27,781%	30,737%	31,192%	30,880%	30,087%	32,365%
BA	31,693%	34,411%	36,405%	39,622%	40,495%	41,599%	40,788%	41,462%
CE	35,101%	38,805%	40,845%	43,977%	44,805%	45,727%	44,965%	45,377%
DF	2,851%	3,671%	4,433%	5,831%	7,438%	9,236%	8,177%	9,505%
ES	32,368%	38,208%	42,599%	48,840%	51,886%	53,195%	53,973%	55,930%
GO	28,377%	33,728%	36,957%	42,241%	43,590%	43,610%	43,562%	43,989%
MA	33,828%	38,673%	40,554%	43,683%	43,638%	44,685%	42,702%	44,506%

MG	40,700%	46,335%	49,572%	57,516%	59,278%	59,996%	59,291%	59,650%
MS	38,816%	45,015%	47,736%	52,202%	53,801%	55,290%	55,980%	57,207%
MT	29,870%	34,360%	36,878%	41,565%	43,327%	43,283%	42,975%	43,137%
PA	30,515%	35,882%	38,069%	41,159%	42,135%	43,909%	42,893%	44,539%
PB	32,444%	35,083%	35,714%	37,442%	37,526%	38,356%	37,611%	38,726%
PE	30,316%	34,786%	37,623%	42,099%	43,532%	45,244%	45,620%	46,710%
PI	31,079%	31,383%	31,721%	33,089%	33,689%	34,727%	33,559%	34,458%
PR	36,744%	42,403%	46,680%	53,743%	55,916%	57,204%	57,774%	58,315%
RJ	27,728%	28,478%	31,437%	41,121%	46,428%	50,941%	52,189%	53,478%
RN	43,200%	48,836%	52,677%	58,473%	59,867%	59,158%	57,635%	58,301%
RO	19,655%	22,235%	23,664%	28,579%	30,106%	32,639%	34,387%	36,919%
RR	22,375%	35,881%	46,594%	55,640%	59,508%	59,913%	61,008%	61,515%
RS	34,390%	38,494%	41,512%	48,100%	49,813%	50,718%	50,622%	51,431%
SC	37,839%	43,119%	47,206%	55,823%	58,830%	60,960%	62,585%	64,784%
SE	43,018%	46,619%	48,402%	51,392%	51,760%	52,219%	50,939%	51,294%
SP	34,321%	42,003%	48,100%	54,733%	57,340%	57,954%	58,406%	58,443%
TO	32,056%	35,312%	36,356%	39,391%	39,629%	39,785%	38,648%	39,583%

A partir da análise dos valores de eficiência de pareamento, percebe-se que o limiar de aceitação impacta na quantidade de endereços pareados e percebe-se que é possível atribuir um limiar visando obter quantidades distintas de pareamento. Entretanto, de acordo com o exposto na Seção 6.5.2, entende-se que com limiar superior a 85% o pareamento é considerado adequado. Com índices de eficiência que permitem considerar os pareamentos encontrados como corretos.

A partir da qualificação da eficiência de pareamento (Tabela 14) para cada limiar permite inferir que o Distrito Federal (DF) não propicia um

pareamento compatível com as demais Unidades da Federação em função do seu sistema de endereçamento baseado em “quadras” ao invés de “logradouros”, e por isso o pareamento do DF é consideravelmente inferior a qualquer outro estado. Em contrapartida, o estado de Santa Catarina (SC) pode ser considerado o melhor para fins de pareamento.

Tabela 14 - Qualificação da eficiência pelo métrica de Dice

Limiar	Piores	Melhores
65%	DF, RJ, GO, PA, TO	PB, SE, AC, RR, MA
75%	DF, RO, TO, GO, RJ	AL, SC, PB, RR, AL
85%	DF, RO, TO, GO, AP	SC, PR, SP, SE, MS
95%	DF, AP, PI, RO, PB	SC, RR, MG, SP, PR

Conclusão

Para se obter um pareamento que possa ser identificado a partir de uma identidade entre instâncias de BDG distintos, é necessário que cada conjunto de caracteres (string) seja exatamente igual. Entretanto, os registros encontrados nos BDG encontram-se, invariavelmente, eivados de erros. Tais erros são inerentes ao processo de inserção dos dados. Aqui, é possível constatar que há registros equivocados, falta de informação, digitação de caracteres invertidos, por exemplo. Esses pequenos equívocos inviabilizam a busca de instâncias idênticas em boa parte dos casos.

Visando mitigar o problema de não identificação da identidade entre registros, procurou-se desenvolver o conceito de similaridade, buscando encontrar instâncias similares com potencial para serem identificadas como iguais mediante um parâmetro de aceitação. Destarte, aplicou-se a métrica de Dice como função de similaridade entre as bases do CadÚnico e do CNEFE.

A partir da função desenvolvida e do limiar de aceitação estabelecido, foi desenvolvido um programa em Python que permitiu processar todas as bases do CadÚnico com o CNEFE. O processamento foi desenvolvido por Unidades da Federação com o intuito de diminuir o tamanho do banco para reduzir o tempo para obtenção da resposta. O algoritmo desenvolvido procurou geolocalizar os registros de endereços a partir da identificação dos municípios e do CEP.

Diante dos dados processados a partir de um BDG teste, percebe-se que o limiar de aceitação da função de similaridade que retorna a quantidade de pareamentos de modo mais eficiente é de 85%. Entretanto, quando se almeja dados sem qualquer tipo de incerteza é preciso aumentar o limiar para 95%. Possíveis decréscimo no valor do limiar é possível, mas

o aumento de falsos-positivos encontrados induz a uma queda na eficiência do processamento.

A resposta obtida permite inferir que Santa Catarina possui os melhores resultados, seguido do Paraná. Em contrapartida, o Distrito Federal apresenta os piores resultados. Neste caso, esta Unidade da Federação possui valores muito abaixo dos demais estados. Esse fato ocorre em função de os endereços do DF possuírem uma lógica distinta. Por outro lado, o DF apresenta condições para o desenvolvimento de um projeto piloto para a padronização do preenchimento e validação dos endereços no momento do cadastro, devido as características de endereçamento intrínsecas, se constituir de apenas um município e ser a sede do Ministério da Cidadania.

Os bancos e as tabelas respostas ocupam um espaço de cerca de 66 Gb. Daí conclui-se que o processamento é lento. Com o programa desenvolvido, é possível utilizar equipamentos tipo notebook. Entretanto, sugere-se que o aplicativo seja instalado em máquinas com maior capacidade de processamento visando reduzir o tempo para se obter o pareamento. Ressalta-se que o algoritmo foi implementado sob o paradigma da programação em paralelo e, no contexto deste projeto, rodou em computadores com no mínimo 6 (seis) *threads*. Fator que proporciona considerável redução de tempo de processamento.

Referências Bibliográficas

COELHO, V. B. N. **Processamento de consultas em banco de dados geográficos ambíguos.** Rio de Janeiro: UFRJ/COPPE, 2010.

DICE, Lee R. 1945. **Measures of the Amount of Ecologic Association between Species.** Ecology 26 (3): 297–302.

D. LI. **An Information-Theoretic Definition of Similarity.** Peking University, 1998.

E. L. LIMA. **Espaços Métricos**. Edição 3. Rio de Janeiro, Instituto de Matemática Pura e Aplicada, CNPq: Editora S. A., Abril de 1993, 299 p.

IBGE – Instituto Brasileiro de Geografia e Estatística (2011). **Censo Demográfico 2010.** IBGE, Rio de Janeiro, RJ.

IBGE – Instituto Brasileiro de Geografia e Estatística (2013). **Cadastro Nacional de Endereços para Fins Estatísticos.** IBGE, Rio de Janeiro, RJ.

IBGE – Instituto Brasileiro de Geografia e Estatística. **Contagem.** Disponível em: <https://cidades.ibge.gov.br/brasil/mg/contagem/panorama>. Acesso em: 23 de junho de 2020.

JACCARD, Paul. 1912. **The Distribution of the Flora in the Alpine Zone.** New Phytologist 11 (2): 37–50.

JARO, Matthew A. 1989. **Advances in Record-linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida.** Journal of the American Statistical Association 84 (406): 414–420.

MDS. **Cadastro Único para Programas Sociais – Manual do Entrevistador.** 3a Edição. Brasília. 2011.

SANTOS, R., Murrieta-Flores, P., and Martins, B., 2017. **Learning to combine multiple string similarity metrics for effective toponym matching.** International Journal of Digital Earth, 1–26. doi:10.1080/17538947.2017.1371253.

SKABA, D. A. **PRODUTO 1: Análise comparativa entre o padrão de endereçamento do Cadastro Nacional de Endereços para Fins Estatísticos (CNEFE) e o Diretório Nacional de Endereços (DNE) dos Correios, apontando as diferenças de captação, preenchimento e estrutura das tabelas auxiliares.** Ministério do Desenvolvimento Social e Agrário – MDSA, Brasília, DF. 2017.

WINKLER, William E. 1990. **String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of RecordLinkage.** Paper presented at the annual meeting of the American Statistical Association – Section on Survey Research Methods, Anaheim, CA, USA, August 6–9. AMS.

Apêndice Referente ao Limiar de Aceitação de 95%

Segue a estatística referente aos pareamentos, para o limiar de aceitação igual a 95%, por estado da Federação, considerando o processamento entre a base do CNEFE e a do CadÚnico por ano:

1. Quantidade de pareamentos feitos no ano de 2012

UF	nível 0	nível 1	nível 2	nível 3	nível 4	nível 5
AC	50.410	25.534	185	19.575	9.069	4.597
AL	217.561	189.194	2.023	91.711	83.771	55.415
AM	246.980	101.483	1.624	82.015	38.458	26.006
AP	55.703	9.063	220	12.076	2.226	1.889
BA	1.101.546	847.335	10.128	387.892	359.200	161.865
CE	669.601	441.486	3.712	273.047	205.425	124.468
DF	35.769	196.637	671	26	6.808	7
ES	143.262	139.733	1.662	52.041	51.861	32.330
GO	193.547	333.922	2.990	42.412	148.481	19.276
MA	409.627	480.519	5.011	135.645	247.292	74.684
MG	1.143.312	369.493	13.984	523.713	115.261	408.945
MS	156.423	37.216	1.352	70.736	13.938	39.029
MT	176.158	98.673	2.044	56.652	35.047	26.226
PA	470.957	342.745	3.598	149.729	131.778	77.424
PB	246.643	279.547	1.786	84.619	118.178	50.766
PE	906.651	343.236	7.889	281.281	115.965	149.956
PI	204.000	268.564	1.447	67.346	122.975	23.430
PR	522.959	203.379	5.997	193.404	81.892	150.101
RJ	793.750	222.361	11.772	199.555	67.929	126.883
RN	264.070	84.835	3.133	137.034	35.979	94.739
RO	152.629	26.312	672	24.684	3.946	15.310

RR	38.105	17.329	217	7.252	4.091	4.698
RS	567.222	106.437	4.347	195.320	56.229	103.832
SC	157.467	113.859	1.570	63.708	60.568	41.841
SE	156.798	72.194	1.053	91.565	41.799	40.304
SP	1.862.544	110.842	22.641	653.204	24.470	365.341
TO	61.279	109.550	756	21.095	47.229	12.631

2. Quantidade de pareamentos feitos no ano de 2013

UF	nível 0	nível 1	nível 2	nível 3	nível 4	nível 5
AC	50.164	26.622	189	25.896	12.179	7.107
AL	214.907	189.109	2.014	108.740	90.351	68.235
AM	232.225	101.088	1.585	120.980	47.959	47.002
AP	54.805	9.363	166	16.541	2.112	4.817
BA	1.122.715	872.904	10.064	471.958	368.655	211.659
CE	657.639	437.299	3.565	320.679	220.741	155.166
DF	35.436	205.439	617	35	9.157	12
ES	139.417	139.603	1.595	67.641	59.934	45.941
GO	198.844	343.142	3.215	52.430	198.930	26.106
MA	396.966	494.245	4.982	173.036	285.422	106.695
MG	1.075.459	385.473	12.511	619.674	113.804	538.695
MS	165.522	41.906	1.333	95.522	17.544	57.839
MT	192.292	104.820	2.301	75.928	42.592	38.212
PA	462.178	362.724	3.650	196.814	152.008	114.848
PB	242.446	295.970	1.811	103.662	122.356	65.939
PE	908.811	326.396	7.316	345.849	116.175	200.750
PI	205.677	294.502	1.393	79.497	116.951	32.954
PR	513.112	203.929	5.538	241.369	88.856	201.744
RJ	883.178	226.986	13.199	222.558	82.215	142.516
RN	258.810	75.467	3.029	164.396	31.944	125.621
RO	160.677	32.803	850	31.965	5.029	18.569
RR	33.438	16.632	199	12.081	4.787	11.263
RS	590.958	112.079	4.352	242.727	58.041	141.954
SC	157.591	117.894	1.497	80.101	71.489	58.382
SE	159.423	73.082	952	110.079	44.279	49.522
SP	1.850.753	109.924	21.088	870.056	27.775	537.390
TO	62.406	115.431	716	25.524	55.684	16.259

3. Quantidade de pareamentos feitos no ano de 2014

UF	nível 0	nível 1	nível 2	nível 3	nível 4	nível 5
AC	52.630	28.466	196	30.234	14.028	8.371
AL	222.439	196.075	2.029	119.196	94.762	73.275
AM	229.307	106.246	1.625	146.808	57.699	63.959
AP	62.459	10.527	173	19.927	2.345	5.870
BA	1.156.576	902.930	10.598	543.707	388.656	252.657
CE	675.584	434.772	3.658	359.862	231.128	178.221
DF	35.224	218.303	590	50	11.719	18
ES	139.469	142.202	1.565	84.698	65.015	60.489
GO	210.026	347.330	3.458	60.468	237.060	31.237
MA	406.954	515.759	5.123	197.147	310.764	125.055
MG	1.062.916	404.375	11.949	699.773	114.039	640.338
MS	179.971	44.799	1.379	115.425	18.966	72.164
MT	208.941	111.445	2.497	91.843	48.598	48.196
PA	496.219	379.545	4.000	234.185	166.723	139.889
PB	252.854	311.379	1.789	115.899	125.250	73.300
PE	926.023	316.929	7.054	395.108	122.218	236.610
PI	211.823	312.660	1.428	87.053	119.736	37.533
PR	501.970	199.212	5.227	282.157	91.120	245.157
RJ	939.540	229.537	14.074	266.582	96.137	179.761
RN	255.647	71.448	2.898	185.077	30.546	151.701
RO	169.788	37.525	925	37.321	5.847	21.384
RR	29.176	17.637	205	16.351	6.157	18.513
RS	610.646	116.029	4.420	282.293	59.637	176.961
SC	158.007	121.847	1.478	96.593	79.249	75.717
SE	166.776	73.121	1.027	123.302	46.534	56.165
SP	1.777.628	105.694	18.646	1.048.272	29.743	684.694
TO	63.899	126.135	709	28.534	61.767	18.659

4. Quantidade de pareamentos feitos no ano de 2015

UF	nível 0	nível 1	nível 2	nível 3	nível 4	nível 5
AC	45.625	25.928	168	30.435	13.856	8.878
AL	195.254	180.713	1.887	121.483	89.998	74.583
AM	186.451	100.930	1.349	156.586	60.358	71.939
AP	55.579	9.695	137	20.518	2.164	6.346
BA	1.018.780	809.499	8.906	571.086	357.866	276.646
CE	611.300	390.887	3.224	376.274	220.364	192.588
DF	20.233	199.730	288	52	13.568	17
ES	111.198	122.714	1.165	92.336	62.993	69.086
GO	175.441	306.803	2.817	63.817	256.151	34.772
MA	340.692	475.610	4.230	200.861	303.502	132.089
MG	755.408	350.584	8.092	704.536	92.997	710.745
MS	165.141	39.932	1.129	125.272	18.843	81.090
MT	190.777	98.583	2.292	101.729	49.652	56.075
PA	457.542	355.817	3.672	249.695	164.974	156.842
PB	213.796	286.898	1.443	114.353	110.855	75.327
PE	800.869	257.794	5.550	408.190	109.170	256.408
PI	183.104	290.571	1.250	86.937	107.342	40.578
PR	411.050	160.217	3.854	306.054	83.442	278.699
RJ	750.863	184.042	9.938	310.158	107.638	242.081
RN	209.464	50.937	2.392	186.599	22.497	160.938
RO	134.771	39.146	958	40.715	6.167	23.093
RR	24.075	16.785	172	19.392	6.989	25.084
RS	481.278	92.023	3.340	290.415	46.624	197.389
SC	114.593	88.268	901	100.217	71.743	85.523
SE	150.829	64.409	951	126.727	42.545	59.300
SP	1.540.505	84.491	14.901	1.158.895	28.372	795.533
TO	53.573	119.729	518	29.578	63.132	20.258

5. Quantidade de pareamentos feitos no ano de 2016

UF	nível 0	nível 1	nível 2	nível 3	nível 4	nível 5
AC	43.227	25.793	136	30.859	13.602	9.481
AL	185.071	174.273	1.820	121.966	86.575	71.899
AM	176.137	102.744	1.299	161.825	63.907	77.528
AP	57.587	9.856	153	21.625	2.153	6.865
BA	976.505	777.208	8.687	577.140	344.586	277.666
CE	592.398	375.441	3.066	379.833	214.466	193.837
DF	13.585	156.721	163	62	13.620	17
ES	96.439	109.750	940	93.178	59.719	70.471
GO	158.364	283.733	2.513	63.795	244.664	35.101
MA	354.298	464.686	4.222	202.893	301.451	133.021
MG	668.585	323.019	6.696	676.340	81.766	695.113
MS	153.710	36.366	965	124.640	17.435	80.401
MT	175.434	90.325	2.126	100.611	48.183	56.004
PA	451.980	351.704	3.692	256.935	167.125	163.845
PB	202.228	283.957	1.346	113.185	105.117	74.544
PE	766.830	240.231	5.137	415.401	106.876	258.026
PI	163.958	283.713	1.100	84.558	102.414	41.026
PR	377.757	142.131	3.511	308.831	76.087	278.958
RJ	654.767	163.903	7.070	328.792	112.054	274.780
RN	195.984	43.227	2.192	183.868	19.248	156.989
RO	123.349	38.330	971	41.223	6.063	22.775
RR	19.940	17.069	131	20.091	7.763	26.727
RS	428.480	78.479	2.934	277.224	39.230	189.639
SC	97.783	75.167	723	97.541	65.216	85.418
SE	149.942	62.285	891	128.112	41.227	59.330
SP	1.459.562	76.138	13.821	1.215.958	27.179	839.558
TO	48.498	117.816	538	28.110	61.769	19.647

6. Quantidade de pareamentos feitos no ano de 2017

UF	nível 0	nível 1	nível 2	nível 3	nível 4	nível 5
AC	41.491	25.580	141	33.361	13.639	10.894
AL	185.735	177.302	1.818	123.492	84.798	73.304
AM	171.590	106.823	1.201	163.682	71.161	84.866
AP	63.362	10.689	164	23.114	2.385	7.657
BA	993.112	768.438	8.666	609.444	353.970	297.524
CE	597.526	369.422	2.985	392.158	221.906	203.130
DF	14.863	142.933	179	58	16.005	13
ES	91.122	102.552	858	92.714	58.351	70.029
GO	158.723	294.062	2.394	65.939	247.838	38.242
MA	364.717	448.182	4.425	204.183	318.283	137.785
MG	684.398	309.635	6.580	688.851	81.377	730.439
MS	146.024	32.523	928	124.539	16.127	81.277
MT	176.279	90.573	2.159	99.947	49.269	56.081
PA	459.887	356.864	3.847	277.004	179.707	185.663
PB	199.118	286.932	1.490	115.628	109.781	77.941
PE	768.504	236.098	5.123	442.758	114.277	277.274
PI	156.023	280.307	1.188	84.108	106.670	41.996
PR	376.117	134.925	3.450	321.278	74.169	292.267
RJ	617.092	158.973	5.889	366.973	120.936	324.046
RN	203.882	39.029	2.233	182.889	17.064	155.133
RO	115.327	37.503	892	43.480	6.312	24.693
RR	18.009	17.645	119	19.534	8.664	25.268
RS	408.221	71.636	2.917	270.226	36.707	189.914
SC	88.040	65.670	651	95.189	58.438	87.403
SE	151.231	62.200	880	130.100	42.557	61.558
SP	1.511.855	75.737	14.852	1.281.702	28.599	898.390
TO	51.714	117.545	705	28.612	63.000	20.684

7. Quantidade de pareamentos feitos no ano de 2018

UF	nível 0	nível 1	nível 2	nível 3	nível 4	nível 5
AC	42.459	26.474	181	33.140	13.384	11.597
AL	190.674	180.311	1.831	125.942	79.224	69.978
AM	172.786	116.205	1.250	166.233	73.681	88.393
AP	71.807	11.880	174	25.110	2.516	8.463
BA	1.002.012	763.544	8.632	601.271	338.874	282.002
CE	596.874	374.305	2.912	389.884	212.110	193.855
DF	15.378	121.081	217	74	12.089	9
ES	92.331	97.675	883	96.733	54.097	73.014
GO	151.486	283.830	2.300	63.590	238.255	35.935
MA	374.948	461.242	4.618	202.161	296.774	127.696
MG	703.100	291.157	6.352	681.184	70.989	705.164
MS	142.369	27.788	947	124.612	13.853	79.130
MT	179.806	88.192	2.215	102.761	46.579	54.301
PA	484.199	371.206	3.998	285.255	175.479	184.761
PB	195.441	295.065	1.479	115.922	105.559	75.114
PE	784.675	235.189	5.258	461.918	114.839	283.217
PI	155.119	288.958	1.181	83.545	100.746	40.606
PR	368.355	125.292	3.411	324.132	66.912	289.023
RJ	614.843	160.844	5.248	386.181	124.690	341.560
RN	217.575	36.304	2.343	183.890	14.917	149.773
RO	110.142	37.016	801	44.035	6.165	27.345
RR	17.273	18.767	126	20.977	8.847	26.763
RS	391.673	64.710	2.764	258.565	31.454	180.694
SC	82.776	55.554	678	94.264	49.142	89.114
SE	158.595	65.897	926	133.756	41.501	58.788
SP	1.524.836	74.209	15.377	1.318.359	27.328	921.270
TO	57.931	114.227	1.063	27.814	62.090	19.214

8. Quantidade de pareamentos feitos no ano de 2019

UF	nível 0	nível 1	nível 2	nível 3	nível 4	nível 5
AC	45.342	27.920	198	37.171	14.229	13.381
AL	196.320	183.822	1.868	130.848	81.777	75.461
AM	186.586	122.659	1.308	182.305	80.021	99.909
AP	79.182	13.588	193	29.884	3.186	11.415
BA	1.065.658	780.051	9.017	645.135	360.998	307.542
CE	632.829	382.986	3.012	411.771	226.668	207.939
DF	22.429	131.134	364	98	16.062	7
ES	102.232	101.874	975	113.619	59.851	86.801
GO	163.617	302.964	2.402	68.418	259.274	40.627
MA	386.891	462.889	4.880	216.341	327.369	141.720
MG	763.916	291.221	6.809	723.524	73.025	773.335
MS	150.627	27.799	1.041	137.434	14.436	88.050
MT	196.185	92.950	2.530	109.429	51.188	60.646
PA	514.849	391.041	4.402	320.511	194.923	215.584
PB	202.873	308.124	1.454	125.841	114.706	83.333
PE	820.231	239.092	5.446	498.509	124.633	310.161
PI	166.295	296.206	1.257	91.067	107.980	44.765
PR	399.021	125.890	3.775	352.919	68.750	317.939
RJ	667.314	174.955	5.664	437.862	144.217	392.654
RN	231.614	33.194	2.416	194.770	14.263	164.590
RO	114.643	41.064	803	50.590	7.015	33.994
RR	18.116	20.503	111	22.865	10.591	28.450
RS	415.991	64.075	2.783	277.362	32.991	200.948
SC	88.978	55.121	765	107.785	50.266	108.446
SE	168.299	67.167	940	139.871	44.891	64.202
SP	1.654.884	76.730	16.385	1.422.509	30.316	1.005.484
TO	61.586	114.682	1.247	29.482	66.144	20.673

Apêndice Referente ao Limiar de Aceitação de 85%

Segue a estatística referente aos pareamentos, para o limiar de aceitação igual a 85%, por estado da Federação, considerando o processamento entre a base do CNEFE e a do CadÚnico por ano:

1. Quantidade de pareamentos feitos no ano de 2012

UF	nível 0	nível 1	nível 2	nível 3	nível 4	nível 5
AC	22.499	9.092	63	47.486	25.511	4.719
AL	75.121	72.830	688	234.151	200.135	56.750
AM	132.567	38.539	792	196.428	101.402	26.838
AP	23.684	4.532	84	44.095	6.757	2.025
BA	488.263	312.274	4.305	1.001.175	894.261	167.688
CE	263.034	172.744	1.333	679.614	474.167	126.847
DF	34.688	152.514	642	1.107	50.931	36
ES	61.438	56.264	663	133.865	135.330	33.329
GO	122.757	185.185	1.833	113.202	297.218	20.433
MA	193.443	159.677	2.265	351.829	568.134	77.430
MG	440.748	161.300	4.466	1.226.277	323.454	418.463
MS	63.805	16.277	481	163.354	34.877	39.900
MT	79.387	40.558	1.022	153.423	93.162	27.248
PA	222.166	126.651	1.495	398.520	347.872	79.527
PB	98.345	111.942	460	232.917	285.783	52.092
PE	395.994	152.936	3.145	791.938	306.265	154.700
PI	72.253	101.336	552	199.093	290.203	24.325
PR	212.606	83.840	2.194	503.757	201.431	153.904
RJ	425.887	116.603	5.953	567.418	173.687	132.702

RN	83.615	35.068	979	317.489	85.746	96.893
RO	106.367	14.645	203	70.946	15.613	15.779
RR	22.640	8.933	111	22.717	12.487	4.804
RS	238.262	34.708	1.357	524.280	127.958	106.822
SC	61.356	39.481	509	159.819	134.946	42.902
SE	55.329	21.438	300	193.034	92.555	41.057
SP	758.621	54.017	8.520	1.757.127	81.295	379.462
TO	33.700	50.741	414	48.674	106.038	12.973

2. Quantidade de pareamentos feitos no ano de 2013

UF	nível 0	nível 1	nível 2	nível 3	nível 4	nível 5
AC	17.469	7.344	51	58.591	31.457	7.245
AL	64.610	66.231	603	259.037	213.229	69.646
AM	100.694	30.299	602	252.511	118.748	47.985
AP	18.961	4.403	56	52.385	7.072	4.927
BA	448.987	291.253	3.955	1.145.686	950.306	217.768
CE	233.683	154.395	1.134	744.635	503.645	157.597
DF	34.355	153.341	592	1.116	61.255	37
ES	54.279	50.776	571	152.779	148.761	46.965
GO	115.604	166.567	1.726	135.670	375.505	27.595
MA	166.039	137.920	1.914	403.963	641.747	109.763
MG	367.361	155.313	3.510	1.327.772	343.964	547.696
MS	56.584	16.084	391	204.460	43.366	58.781
MT	75.432	37.712	1.010	192.788	109.700	39.503
PA	192.005	116.282	1.326	466.987	398.450	117.172
PB	76.381	112.562	448	269.727	305.764	67.302
PE	355.907	128.324	2.610	898.753	314.247	205.456
PI	66.085	104.840	547	219.089	306.613	33.800
PR	189.359	76.853	1.763	565.122	215.932	205.519
RJ	485.559	112.079	7.391	620.177	197.122	148.324
RN	75.754	30.027	910	347.452	77.384	127.740
RO	99.386	18.182	269	93.256	19.650	19.150
RR	17.192	7.613	84	28.327	13.806	11.378
RS	228.696	32.611	1.319	604.989	137.509	144.987
SC	56.097	37.306	438	181.595	152.077	59.441
SE	48.347	19.493	233	221.155	97.868	50.241
SP	661.039	49.266	7.026	2.059.770	88.433	551.452
TO	31.474	49.902	369	56.456	121.213	16.606

3. Quantidade de pareamentos feitos no ano de 2014

UF	nível 0	nível 1	nível 2	nível 3	nível 4	nível 5
AC	16.002	7.185	52	66.862	35.309	8.515
AL	61.361	65.910	553	280.274	224.927	74.751
AM	87.309	26.661	548	288.806	137.284	65.036
AP	18.896	4.639	44	63.490	8.233	5.999
BA	431.897	288.211	3.844	1.268.386	1.003.375	259.411
CE	227.478	145.916	1.115	807.968	519.984	180.764
DF	34.040	156.204	566	1.234	73.818	42
ES	48.651	48.487	514	175.516	158.730	61.540
GO	115.875	147.076	1.806	154.619	437.314	32.889
MA	160.835	135.049	1.875	443.266	691.474	128.303
MG	334.495	157.517	3.108	1.428.194	360.897	649.179
MS	55.086	15.819	368	240.310	47.946	73.175
MT	75.100	36.829	1.023	225.684	123.214	49.670
PA	195.676	112.184	1.394	534.728	434.084	142.495
PB	70.066	116.645	395	298.687	319.984	74.694
PE	335.136	116.966	2.348	985.995	322.181	241.316
PI	63.345	108.000	501	235.531	324.396	38.460
PR	167.071	70.160	1.460	617.056	220.172	248.924
RJ	503.467	108.419	7.560	702.655	217.255	186.275
RN	68.837	27.940	775	371.887	74.054	153.824
RO	97.836	20.302	286	109.273	23.070	22.023
RR	12.667	6.831	84	32.860	16.963	18.634
RS	222.038	31.573	1.244	670.901	144.093	180.137
SC	50.773	35.987	394	203.827	165.109	76.801
SE	47.325	18.477	244	242.753	101.178	56.948
SP	543.053	43.570	5.157	2.282.847	91.867	698.183
TO	30.388	53.347	352	62.045	134.555	19.016

4. Quantidade de pareamentos feitos no ano de 2015

UF	nível 0	nível 1	nível 2	nível 3	nível 4	nível 5
AC	10.602	5.506	36	65.458	34.278	9.010
AL	43.577	56.383	437	273.160	214.328	76.033
AM	51.310	21.476	304	291.727	139.812	72.984
AP	13.372	3.947	22	62.725	7.912	6.461
BA	322.492	233.931	2.669	1.267.374	933.434	282.883
CE	181.245	118.794	874	806.329	492.457	194.938
DF	19.385	132.851	273	900	80.447	32
ES	29.577	35.298	278	173.957	150.409	69.973
GO	82.644	103.667	1.236	156.614	459.287	36.353
MA	112.913	112.295	1.266	428.640	666.817	135.053
MG	159.574	124.291	1.546	1.300.370	319.290	717.291
MS	40.962	12.311	216	249.451	46.464	82.003
MT	53.957	25.871	735	238.549	122.364	57.632
PA	161.790	95.450	1.161	545.447	425.341	159.353
PB	43.413	101.526	271	284.736	296.227	76.499
PE	243.991	80.123	1.426	965.068	286.841	260.532
PI	44.941	95.490	403	225.100	302.423	41.425
PR	106.047	46.260	743	611.057	197.399	281.810
RJ	346.998	71.728	4.442	714.023	219.952	247.577
RN	47.633	18.273	581	348.430	55.161	162.749
RO	56.547	20.968	301	118.939	24.345	23.750
RR	8.138	5.171	52	35.329	18.603	25.204
RS	140.889	19.411	705	630.804	119.236	200.024
SC	28.016	20.616	171	186.794	139.395	86.253
SE	36.169	14.312	197	241.387	92.642	60.054
SP	370.881	28.693	3.233	2.328.519	84.170	807.201
TO	21.513	47.457	224	61.638	135.404	20.552

5. Quantidade de pareamentos feitos no ano de 2016

UF	nível 0	nível 1	nível 2	nível 3	nível 4	nível 5
AC	8.999	5.146	33	65.087	34.249	9.584
AL	36.356	52.199	407	270.681	208.649	73.312
AM	42.191	20.265	238	295.771	146.386	78.589
AP	13.023	3.973	19	66.189	8.036	6.999
BA	290.750	216.325	2.436	1.262.895	905.469	283.917
CE	170.055	109.908	795	802.176	479.999	196.108
DF	12.746	95.922	142	901	74.419	38
ES	22.232	29.605	178	167.385	139.864	71.233
GO	71.168	85.428	1.002	150.991	442.969	36.612
MA	118.449	105.801	1.233	438.742	660.336	136.010
MG	123.744	112.339	1.129	1.221.181	292.446	700.680
MS	34.373	10.436	143	243.977	43.365	81.223
MT	44.139	20.956	638	231.906	117.552	57.492
PA	154.770	90.440	1.187	554.145	428.389	166.350
PB	35.977	98.469	264	279.436	290.605	75.626
PE	219.329	68.557	1.240	962.902	278.550	261.923
PI	36.898	90.928	310	211.618	295.199	41.816
PR	87.079	38.634	519	599.509	179.584	281.950
RJ	267.915	57.227	2.703	715.644	218.730	279.147
RN	41.989	15.080	504	337.863	47.395	158.677
RO	44.760	20.474	306	119.812	23.919	23.440
RR	5.682	4.787	29	34.349	20.045	26.829
RS	116.233	15.122	548	589.471	102.587	192.025
SC	20.029	16.122	104	175.295	124.261	86.037
SE	34.427	13.173	174	243.627	90.339	60.047
SP	311.146	23.310	2.654	2.364.374	80.007	850.725
TO	18.176	46.757	205	58.432	132.828	19.980

6. Quantidade de pareamentos feitos no ano de 2017

UF	nível 0	nível 1	nível 2	nível 3	nível 4	nível 5
AC	9.063	5.450	34	65.789	33.769	11.001
AL	39.131	58.438	412	270.096	203.662	74.710
AM	44.683	22.429	255	290.589	155.555	85.812
AP	16.364	4.691	31	70.112	8.383	7.790
BA	313.112	217.010	2.531	1.289.444	905.398	303.659
CE	178.200	101.924	801	811.484	489.404	205.314
DF	13.793	81.130	157	1.128	77.808	35
ES	21.962	28.878	177	161.874	132.025	70.710
GO	75.784	97.698	1.032	148.878	444.202	39.604
MA	134.191	97.274	1.472	434.709	669.191	140.738
MG	145.882	114.423	1.169	1.227.367	276.589	735.850
MS	36.379	9.597	107	234.184	39.053	82.098
MT	47.139	22.834	717	229.087	117.008	57.523
PA	165.934	94.608	1.257	570.957	441.963	188.253
PB	38.751	108.222	259	275.995	288.491	79.172
PE	236.388	69.215	1.225	974.874	281.160	281.172
PI	36.982	90.963	330	203.149	296.014	42.854
PR	89.569	37.321	562	607.826	171.773	295.155
RJ	238.829	56.172	1.737	745.236	223.737	328.198
RN	47.666	14.710	557	339.105	41.383	156.809
RO	43.197	20.901	316	115.610	22.914	25.269
RR	5.700	4.877	30	31.843	21.432	25.357
RS	119.359	13.568	557	559.088	94.775	192.274
SC	19.200	14.144	80	164.029	109.964	87.974
SE	35.518	11.846	216	245.813	92.911	62.222
SP	364.936	23.150	3.351	2.428.621	81.186	909.891
TO	22.257	47.650	338	58.069	132.895	21.051

7. Quantidade de pareamentos feitos no ano de 2018

UF	nível 0	nível 1	nível 2	nível 3	nível 4	nível 5
AC	9.233	4.924	52	66.366	34.934	11.726
AL	38.698	57.054	355	277.918	202.481	71.454
AM	42.702	22.644	238	296.317	167.242	89.405
AP	18.097	4.883	22	78.820	9.513	8.615
BA	300.996	213.696	2.366	1.302.287	888.722	288.268
CE	172.498	108.578	765	814.260	477.837	196.002
DF	14.097	70.248	186	1.355	62.922	40
ES	19.933	25.384	185	169.131	126.388	73.712
GO	66.083	87.096	859	148.993	434.989	37.376
MA	130.339	106.244	1.288	446.770	651.772	131.026
MG	148.299	103.735	1.075	1.235.985	258.411	710.441
MS	31.749	7.470	121	235.232	34.171	79.956
MT	43.096	20.412	653	239.471	114.359	55.863
PA	163.408	93.945	1.278	606.046	452.740	187.481
PB	32.337	101.035	275	279.026	299.589	76.318
PE	221.601	61.736	1.167	1.024.992	288.292	287.308
PI	36.075	92.268	312	202.589	297.436	41.475
PR	77.620	34.296	499	614.867	157.908	291.935
RJ	211.087	51.990	1.059	789.937	233.544	345.749
RN	51.147	12.962	558	350.318	38.259	151.558
RO	38.081	19.865	266	116.096	23.316	27.880
RR	4.859	4.889	33	33.391	22.725	26.856
RS	108.924	12.741	458	541.314	83.423	183.000
SC	15.538	11.809	63	161.502	92.887	89.729
SE	35.988	14.091	169	256.363	93.307	59.545
SP	328.447	22.043	2.839	2.514.748	79.494	933.808
TO	25.324	42.809	533	60.421	133.508	19.744

8. Quantidade de pareamentos feitos no ano de 2019

UF	nível 0	nível 1	nível 2	nível 3	nível 4	nível 5
AC	10.560	5.755	47	71.953	36.394	13.532
AL	42.732	62.942	414	284.436	202.657	76.915
AM	50.884	25.971	292	318.007	176.709	100.925
AP	21.090	5.992	37	87.976	10.782	11.571
BA	351.374	224.109	2.715	1.359.419	916.940	313.844
CE	194.723	108.337	781	849.877	501.317	210.170
DF	20.523	72.474	325	2.004	74.722	46
ES	24.034	27.947	215	191.817	133.778	87.561
GO	78.286	102.380	957	153.749	459.858	42.072
MA	141.848	101.629	1.527	461.384	688.629	145.073
MG	188.372	110.452	1.192	1.299.068	253.794	778.952
MS	37.261	8.114	128	250.800	34.121	88.963
MT	51.075	23.331	806	254.539	120.807	62.370
PA	185.082	101.948	1.466	650.278	484.016	218.520
PB	37.184	115.825	248	291.530	307.005	84.539
PE	252.140	66.561	1.251	1.066.600	297.164	314.356
PI	40.532	96.842	301	216.830	307.344	45.721
PR	90.486	34.827	585	661.454	159.813	321.129
RJ	239.732	59.734	1.196	865.444	259.438	397.122
RN	60.532	12.830	605	365.852	34.627	166.401
RO	42.212	23.049	293	123.021	25.030	34.504
RR	5.947	5.304	40	35.034	25.790	28.521
RS	130.631	12.659	473	562.722	84.407	203.258
SC	19.058	12.242	91	177.705	93.145	109.120
SE	39.648	12.690	207	268.522	99.368	64.935
SP	412.485	23.192	3.467	2.664.908	83.854	1.018.402
TO	28.127	41.904	647	62.941	138.922	21.273

Apêndice Referente ao Limiar de Aceitação de 75%

Segue a estatística referente aos pareamentos, para o limiar de aceitação igual a 75%, por estado da Federação, considerando o processamento entre a base do CNEFE e a do CadÚnico por ano:

1. Quantidade de pareamentos feitos no ano de 2012

UF	nível 0	nível 1	nível 2	nível 3	nível 4	nível 5
AC	10.551	2.025	15	59.434	32.578	4.767
AL	17.417	12.913	158	291.855	260.052	57.280
AM	66.735	10.736	350	262.260	129.205	27.280
AP	7.953	1.369	8	59.826	9.920	2.101
BA	181.006	60.504	1.353	1.308.432	1.146.031	170.640
CE	95.751	41.164	400	846.897	605.747	127.780
DF	27.222	73.538	472	8.573	129.907	206
ES	22.612	16.736	242	172.691	174.858	33.750
GO	77.474	102.007	1.063	158.485	380.396	21.203
MA	80.003	31.640	990	465.269	696.171	78.705
MG	191.397	37.344	1.381	1.475.628	447.410	421.548
MS	22.384	4.894	175	204.775	46.260	40.206
MT	34.911	13.536	497	197.899	120.184	27.773
PA	98.699	27.862	453	521.987	446.661	80.569
PB	53.265	21.576	78	277.997	376.149	52.474
PE	147.896	39.930	1.128	1.040.036	419.271	156.717
PI	22.928	16.967	158	248.418	374.572	24.719
PR	75.175	22.864	628	641.188	262.407	155.470
RJ	186.824	49.879	2.569	806.481	240.411	136.086

RN	17.776	8.081	209	383.328	112.733	97.663
RO	88.744	2.122	29	88.569	28.136	15.953
RR	11.993	3.846	56	33.364	17.574	4.859
RS	97.293	8.514	356	665.249	154.152	107.823
SC	22.430	9.513	113	198.745	164.914	43.298
SE	26.676	4.444	134	221.687	109.549	41.223
SP	275.249	17.265	2.690	2.240.499	118.047	385.292
TO	16.547	17.992	198	65.827	138.787	13.189

2. Quantidade de pareamentos feitos no ano de 2013

UF	nível 0	nível 1	nível 2	nível 3	nível 4	nível 5
AC	6.332	1.203	7	69.728	37.598	7.289
AL	12.694	9.592	105	310.953	269.868	70.144
AM	44.447	7.008	249	308.758	142.039	48.338
AP	6.301	1.166	4	65.045	10.309	4.979
BA	145.271	47.441	1.153	1.449.402	1.194.118	220.570
CE	79.153	33.016	293	899.165	625.024	158.438
DF	27.020	70.183	439	8.451	144.413	190
ES	18.787	13.727	196	188.271	185.810	47.340
GO	65.618	82.202	838	185.656	459.870	28.483
MA	58.421	21.849	727	511.581	757.818	110.950
MG	148.471	29.915	955	1.546.662	469.362	550.251
MS	17.310	4.039	130	243.734	55.411	59.042
MT	29.542	11.225	431	238.678	136.187	40.082
PA	80.379	20.501	389	578.613	494.231	118.109
PB	34.569	17.692	61	311.539	400.634	67.689
PE	123.575	28.825	826	1.131.085	413.746	207.240
PI	18.303	14.037	134	266.871	397.416	34.213
PR	64.603	19.591	474	689.878	273.194	206.808
RJ	223.286	45.888	3.322	882.450	263.313	152.393
RN	14.845	6.436	172	408.361	100.975	128.478
RO	73.517	2.529	45	119.125	35.303	19.374
RR	8.180	2.641	42	37.339	18.778	11.420
RS	85.804	7.405	319	747.881	162.715	145.987
SC	18.919	8.312	99	218.773	181.071	59.780
SE	18.815	3.358	94	250.687	114.003	50.380
SP	221.136	14.565	2.054	2.499.673	123.134	556.424
TO	13.910	15.699	163	74.020	155.416	16.812

3. Quantidade de pareamentos feitos no ano de 2014

UF	nível 0	nível 1	nível 2	nível 3	nível 4	nível 5
AC	4.845	1.009	9	78.019	41.485	8.558
AL	10.846	8.731	79	330.789	282.106	75.225
AM	34.042	5.351	216	342.073	158.594	65.368
AP	5.730	1.064	4	76.656	11.808	6.039
BA	128.210	42.555	1.075	1.572.073	1.249.031	262.180
CE	74.330	29.234	265	961.116	636.666	181.614
DF	26.836	67.337	424	8.438	162.685	184
ES	15.955	12.086	171	208.212	195.131	61.883
GO	61.303	61.179	774	209.191	523.211	33.921
MA	51.738	19.284	651	552.363	807.239	129.527
MG	130.525	27.478	818	1.632.164	490.936	651.469
MS	15.459	3.657	118	279.937	60.108	73.425
MT	26.590	9.928	416	274.194	150.115	50.277
PA	80.312	17.425	382	650.092	528.843	143.507
PB	27.411	17.032	53	341.342	419.597	75.036
PE	107.994	24.241	677	1.213.137	414.906	242.987
PI	15.880	12.839	115	282.996	419.557	38.846
PR	54.946	16.866	378	729.181	273.466	250.006
RJ	232.381	42.674	3.389	973.741	283.000	190.446
RN	12.268	5.627	143	428.456	96.367	154.456
RO	66.408	2.796	49	140.701	40.576	22.260
RR	5.125	2.141	39	40.402	21.653	18.679
RS	78.558	6.671	303	814.381	168.995	181.078
SC	15.944	7.417	80	238.656	193.679	77.115
SE	15.884	3.101	86	274.194	116.554	57.106
SP	163.879	11.789	1.341	2.662.021	123.648	701.999
TO	12.495	16.452	142	79.938	171.450	19.226

4. Quantidade de pareamentos feitos no ano de 2015

UF	nível 0	nível 1	nível 2	nível 3	nível 4	nível 5
AC	1.971	511	7	74.089	39.273	9.039
AL	4.986	5.466	31	311.751	265.245	76.439
AM	13.520	3.368	77	329.517	157.920	73.211
AP	3.439	715	0	72.658	11.144	6.483
BA	72.681	24.893	577	1.517.185	1.142.472	284.975
CE	52.711	19.519	151	934.863	591.732	195.661
DF	15.629	48.113	212	4.656	165.185	93
ES	7.559	6.414	62	195.975	179.293	70.189
GO	34.393	28.071	378	204.865	534.883	37.211
MA	23.914	12.514	261	517.639	766.598	136.058
MG	40.712	14.787	268	1.419.232	428.794	718.569
MS	8.400	2.177	47	282.013	56.598	82.172
MT	12.849	4.938	220	279.657	143.297	58.147
PA	61.797	10.860	298	645.440	509.931	160.216
PB	9.312	10.610	24	318.837	387.143	76.746
PE	60.268	11.601	237	1.148.791	355.363	261.721
PI	6.843	8.343	41	263.198	389.570	41.787
PR	30.025	8.705	138	687.079	234.954	282.415
RJ	149.662	21.627	1.919	911.359	270.053	250.100
RN	6.171	2.783	77	389.892	70.651	163.253
RO	21.848	2.855	47	153.638	42.458	24.004
RR	2.183	1.104	19	41.284	22.670	25.237
RS	38.263	2.648	135	733.430	135.999	200.594
SC	7.003	2.945	26	207.807	157.066	86.398
SE	8.088	1.865	54	269.468	105.089	60.197
SP	84.970	5.731	582	2.614.430	107.132	809.852
TO	6.694	13.690	88	76.457	169.171	20.688

5. Quantidade de pareamentos feitos no ano de 2016

UF	nível 0	nível 1	nível 2	nível 3	nível 4	nível 5
AC	948	341	5	73.138	39.054	9.612
AL	2.519	4.238	17	304.518	256.610	73.702
AM	7.502	2.672	39	330.460	163.979	78.788
AP	3.174	696	0	76.038	11.313	7.018
BA	58.106	19.676	494	1.495.539	1.102.118	285.859
CE	48.762	16.643	129	923.469	573.264	196.774
DF	10.449	30.080	104	3.198	140.261	76
ES	4.742	4.635	19	184.875	164.834	71.392
GO	28.369	15.373	286	193.790	513.024	37.328
MA	18.924	10.217	203	538.267	755.920	137.040
MG	24.881	11.700	121	1.320.044	393.085	701.688
MS	5.651	1.482	19	272.699	52.319	81.347
MT	7.174	2.801	130	268.871	135.707	58.000
PA	58.034	8.562	292	650.881	510.267	167.245
PB	4.727	8.506	14	310.686	380.568	75.876
PE	49.746	8.066	149	1.132.485	339.041	263.014
PI	3.777	6.679	25	244.739	379.448	42.101
PR	23.321	6.561	66	663.267	211.657	282.403
RJ	108.849	13.717	1.008	874.710	262.240	280.842
RN	4.740	1.952	51	375.112	60.523	159.130
RO	9.528	2.891	41	155.044	41.502	23.705
RR	945	733	10	39.086	24.099	26.848
RS	26.497	1.701	94	679.207	116.008	192.479
SC	3.715	1.937	18	191.609	138.446	86.123
SE	6.384	1.588	45	271.670	101.924	60.176
SP	58.298	3.780	253	2.617.222	99.537	853.126
TO	5.011	13.951	71	71.597	165.634	20.114

6. Quantidade de pareamentos feitos no ano de 2017

UF	nível 0	nível 1	nível 2	nível 3	nível 4	nível 5
AC	956	340	5	73.896	38.879	11.030
AL	3.535	3.955	28	305.692	258.145	75.094
AM	7.535	2.238	38	327.737	175.746	86.029
AP	3.526	670	3	82.950	12.404	7.818
BA	61.744	17.533	457	1.540.812	1.104.875	305.733
CE	52.495	13.289	123	937.189	578.039	205.992
DF	11.267	23.984	108	3.654	134.954	84
ES	4.278	3.882	20	179.558	157.021	70.867
GO	31.832	19.582	319	192.830	522.318	40.317
MA	21.989	8.407	211	546.911	758.058	141.999
MG	33.595	11.431	128	1.339.654	379.581	736.891
MS	6.126	1.108	14	264.437	47.542	82.191
MT	7.287	2.950	134	268.939	136.892	58.106
PA	60.775	8.795	269	676.116	527.776	189.241
PB	3.300	8.217	10	311.446	388.496	79.421
PE	53.977	6.726	174	1.157.285	343.649	282.223
PI	3.956	5.945	17	236.175	381.032	43.167
PR	22.380	6.226	53	675.015	202.868	295.664
RJ	91.079	11.837	486	892.986	268.072	329.449
RN	3.942	1.365	46	382.829	54.728	157.320
RO	9.888	4.287	45	148.919	39.528	25.540
RR	865	636	9	36.678	25.673	25.378
RS	25.988	1.461	77	652.459	106.882	192.754
SC	2.806	1.557	12	180.423	122.551	88.042
SE	8.163	1.110	57	273.168	103.647	62.381
SP	68.091	3.313	372	2.725.466	101.023	912.870
TO	7.485	13.849	149	72.841	166.696	21.240

7. Quantidade de pareamentos feitos no ano de 2018

UF	nível 0	nível 1	nível 2	nível 3	nível 4	nível 5
AC	741	215	9	74.858	39.643	11.769
AL	4.840	4.711	27	311.776	254.824	71.782
AM	6.785	2.579	28	332.234	187.307	89.615
AP	4.236	845	0	92.681	13.551	8.637
BA	65.361	20.443	463	1.537.922	1.081.975	290.171
CE	53.737	16.401	120	933.021	570.014	196.647
DF	11.340	20.905	125	4.112	112.265	101
ES	4.072	3.567	24	184.992	148.205	73.873
GO	26.091	14.995	217	188.985	507.090	38.018
MA	19.698	10.539	213	557.411	747.477	132.101
MG	40.497	12.276	127	1.343.787	349.870	711.389
MS	6.093	1.060	16	260.888	40.581	80.061
MT	5.667	2.597	100	276.900	132.174	56.416
PA	58.765	8.790	322	710.689	537.895	188.437
PB	2.867	7.739	13	308.496	392.885	76.580
PE	60.329	6.848	150	1.186.264	343.180	288.325
PI	4.034	6.659	20	234.630	383.045	41.767
PR	21.764	5.698	58	670.723	186.506	292.376
RJ	79.035	10.548	149	921.989	274.986	346.659
RN	8.028	1.464	73	393.437	49.757	152.043
RO	7.567	3.375	39	146.610	39.806	28.107
RR	625	573	7	37.625	27.041	26.882
RS	27.877	1.451	56	622.361	94.713	183.402
SC	2.320	1.428	6	174.720	103.268	89.786
SE	6.560	1.852	39	285.791	105.546	59.675
SP	67.236	3.115	232	2.775.959	98.422	936.415
TO	9.564	11.140	246	76.181	165.177	20.031

8. Quantidade de pareamentos feitos no ano de 2019

UF	nível 0	nível 1	nível 2	nível 3	nível 4	nível 5
AC	1.274	290	4	81.239	41.859	13.575
AL	5.734	4.217	43	321.434	261.382	77.286
AM	9.127	2.507	42	359.764	200.173	101.175
AP	3.897	860	6	105.169	15.914	11.602
BA	81.597	19.332	446	1.629.196	1.121.717	316.113
CE	63.662	14.737	115	980.938	594.917	210.836
DF	16.350	19.994	240	6.177	127.202	131
ES	4.651	3.467	30	211.200	158.258	87.746
GO	32.561	20.884	235	199.474	541.354	42.794
MA	23.659	9.076	235	579.573	781.182	146.365
MG	56.782	12.912	149	1.430.658	351.334	779.995
MS	7.129	988	14	280.932	41.247	89.077
MT	6.993	2.785	124	298.621	141.353	63.052
PA	65.134	9.516	321	770.226	576.448	219.665
PB	2.789	8.866	7	325.925	413.964	84.780
PE	67.925	6.553	161	1.250.815	357.172	315.446
PI	4.936	6.467	19	252.426	397.719	46.003
PR	23.954	5.346	54	727.986	189.294	321.660
RJ	90.108	12.526	160	1.015.068	306.646	398.158
RN	9.482	1.091	60	416.902	46.366	166.946
RO	10.498	5.490	49	154.735	42.589	34.748
RR	748	745	6	40.233	30.349	28.555
RS	38.442	1.336	59	654.911	95.730	203.672
SC	3.433	1.327	23	193.330	104.060	109.188
SE	9.372	1.293	57	298.798	110.765	65.085
SP	93.345	3.091	397	2.984.048	103.955	1.021.472
TO	9.238	9.426	238	81.830	171.400	21.682

Apêndice Referente ao Limiar de Aceitação de 65%

Segue a estatística referente aos pareamentos, para o limiar de aceitação igual a 65%, por estado da Federação, considerando o processamento entre a base do CNEFE e a do CadÚnico por ano:

1. Quantidade de pareamentos feitos no ano de 2012

UF	nível 0	nível 1	nível 2	nível 3	nível 4	nível 5
AC	7.807	491	6	62.178	34.112	4.776
AL	5.405	2.478	51	303.867	270.487	57.387
AM	43.468	3.802	208	285.527	136.139	27.422
AP	3.824	433	6	63.955	10.856	2.103
BA	103.552	14.018	558	1.385.886	1.192.517	171.435
CE	56.095	11.295	221	886.553	635.616	127.959
DF	18.327	30.156	278	17.468	173.289	400
ES	11.456	6.830	103	183.847	184.764	33.889
GO	55.424	71.235	702	180.535	411.168	21.564
MA	48.899	8.937	588	496.373	718.874	79.107
MG	116.720	8.898	581	1.550.305	475.856	422.348
MS	8.408	1.607	73	218.751	49.547	40.308
MT	20.495	6.038	312	212.315	127.682	27.958
PA	66.372	7.228	133	554.314	467.295	80.889
PB	40.139	5.345	20	291.123	392.380	52.532
PE	79.453	11.993	527	1.108.479	447.208	157.318
PI	11.501	4.526	88	259.845	387.013	24.789
PR	33.537	6.798	181	682.826	278.473	155.917
RJ	95.961	25.792	1.080	897.344	264.498	137.575

RN	4.247	1.829	60	396.857	118.985	97.812
RO	86.216	171	9	91.097	30.087	15.973
RR	7.812	1.575	40	37.545	19.845	4.875
RS	60.812	3.266	141	701.730	159.400	108.038
SC	12.716	3.417	43	208.459	171.010	43.368
SE	19.501	1.213	93	228.862	112.780	41.264
SP	128.093	6.026	963	2.387.655	129.286	387.019
TO	9.528	8.856	107	72.846	147.923	13.280

2. Quantidade de pareamentos feitos no ano de 2013

UF	nível 0	nível 1	nível 2	nível 3	nível 4	nível 5
AC	4.380	319	2	71.680	38.482	7.294
AL	3.893	1.766	35	319.754	277.694	70.214
AM	27.748	2.182	141	325.457	146.865	48.446
AP	3.247	321	3	68.099	11.154	4.980
BA	77.798	10.278	427	1.516.875	1.231.281	221.296
CE	47.360	8.930	156	930.958	649.110	158.575
DF	18.323	27.777	262	17.148	186.819	367
ES	9.575	5.512	82	197.483	194.025	47.454
GO	44.874	55.513	532	206.400	486.559	28.789
MA	33.101	5.695	405	536.901	773.972	111.272
MG	89.997	6.689	395	1.605.136	492.588	550.811
MS	6.477	1.255	55	254.567	58.195	59.117
MT	16.358	4.893	242	251.862	142.519	40.271
PA	54.752	4.949	115	604.240	509.783	118.383
PB	26.124	3.973	11	319.984	414.353	67.739
PE	68.618	8.490	364	1.186.042	434.081	207.702
PI	8.986	3.544	77	276.188	407.909	34.270
PR	30.996	5.715	138	723.485	287.070	207.144
RJ	119.697	23.225	1.411	986.039	285.976	154.304
RN	3.424	1.500	48	419.782	105.911	128.602
RO	69.380	193	6	123.262	37.639	19.413
RR	5.347	1.026	31	40.172	20.393	11.431
RS	52.081	2.810	122	781.604	167.310	146.184
SC	10.363	2.855	43	227.329	186.528	59.836
SE	12.280	836	54	257.222	116.525	50.420
SP	104.394	4.857	725	2.616.415	132.842	557.753
TO	7.686	6.939	87	80.244	164.176	16.888

3. Quantidade de pareamentos feitos no ano de 2014

UF	nível 0	nível 1	nível 2	nível 3	nível 4	nível 5
AC	3.089	274	3	79.775	42.220	8.564
AL	3.219	1.641	20	338.416	289.196	75.284
AM	19.970	1.608	125	356.145	162.337	65.459
AP	3.025	254	2	79.361	12.618	6.041
BA	65.926	8.757	367	1.634.357	1.282.829	262.888
CE	45.445	7.544	134	990.001	658.356	181.745
DF	18.410	25.645	249	16.864	204.377	359
ES	8.284	4.657	67	215.883	202.560	61.987
GO	40.635	38.300	447	229.859	546.090	34.248
MA	27.709	4.650	345	576.392	821.873	129.833
MG	79.287	5.958	318	1.683.402	512.456	651.969
MS	5.931	1.070	43	289.465	62.695	73.500
MT	14.038	4.127	213	286.746	155.916	50.480
PA	55.710	4.011	113	674.694	542.257	143.776
PB	20.294	3.552	8	348.459	433.077	75.081
PE	60.034	6.847	281	1.261.097	432.300	243.383
PI	7.722	2.995	69	291.154	429.401	38.892
PR	27.918	4.835	96	756.209	285.497	250.288
RJ	127.945	21.292	1.461	1.078.177	304.382	192.374
RN	2.770	1.270	39	437.954	100.724	154.560
RO	60.987	212	7	146.122	43.160	22.302
RR	3.158	747	23	42.369	23.047	18.695
RS	46.684	2.352	112	846.255	173.314	181.269
SC	8.803	2.497	35	245.797	198.599	77.160
SE	9.294	724	48	280.784	118.931	57.144
SP	79.295	3.629	478	2.746.605	131.808	702.862
TO	6.567	6.119	72	85.866	181.783	19.296

4. Quantidade de pareamentos feitos no ano de 2015

UF	nível 0	nível 1	nível 2	nível 3	nível 4	nível 5
AC	1.067	93	1	74.993	39.691	9.045
AL	1.367	865	7	315.370	269.846	76.463
AM	7.025	738	38	336.012	160.550	73.250
AP	1.934	125	0	74.163	11.734	6.483
BA	32.224	3.810	158	1.557.642	1.163.555	285.394
CE	33.109	3.658	49	954.465	607.593	195.763
DF	11.160	14.307	121	9.125	198.991	184
ES	4.044	1.803	15	199.490	183.904	70.236
GO	19.120	13.183	161	220.138	549.771	37.428
MA	8.466	2.163	87	533.087	776.949	136.232
MG	21.377	2.252	85	1.438.567	441.329	718.752
MS	3.268	499	13	287.145	58.276	82.206
MT	5.000	1.471	80	287.506	146.764	58.287
PA	43.343	2.072	85	663.894	518.719	160.429
PB	5.935	1.246	2	322.214	396.507	76.768
PE	33.827	2.492	57	1.175.232	364.472	261.901
PI	2.381	1.218	21	267.660	396.695	41.807
PR	18.241	2.112	31	698.863	241.547	282.522
RJ	88.101	8.931	797	972.920	282.749	251.222
RN	981	586	15	395.082	72.848	163.315
RO	15.562	192	7	159.924	45.121	24.044
RR	1.057	314	10	42.410	23.460	25.246
RS	22.057	571	37	749.636	138.076	200.692
SC	4.069	749	7	210.741	159.262	86.417
SE	2.946	379	19	274.610	106.575	60.232
SP	41.937	1.378	169	2.657.463	111.485	810.265
TO	2.845	4.222	42	80.306	178.639	20.734

5. Quantidade de pareamentos feitos no ano de 2016

UF	nível 0	nível 1	nível 2	nível 3	nível 4	nível 5
AC	258	39	1	73.828	39.356	9.616
AL	486	606	0	306.551	260.242	73.719
AM	2.890	457	11	335.072	166.194	78.816
AP	1.855	128	0	77.357	11.881	7.018
BA	23.330	2.420	123	1.530.315	1.119.374	286.230
CE	31.304	2.721	33	940.927	587.186	196.870
DF	7.998	7.184	56	5.649	163.157	124
ES	2.638	1.078	5	186.979	168.391	71.406
GO	15.091	4.125	111	207.068	524.272	37.503
MA	3.802	1.299	39	553.389	764.838	137.204
MG	12.379	1.547	16	1.332.546	403.238	701.793
MS	2.331	237	1	276.019	53.564	81.365
MT	1.498	361	20	274.547	138.147	58.110
PA	41.302	1.386	75	667.613	517.443	167.462
PB	2.091	735	1	313.322	388.339	75.889
PE	28.869	1.467	19	1.153.362	345.640	263.144
PI	758	645	7	247.758	385.482	42.119
PR	15.666	1.350	8	670.922	216.868	282.461
RJ	70.511	4.521	406	913.048	271.436	281.444
RN	581	390	7	379.271	62.085	159.174
RO	3.099	183	5	161.473	44.210	23.741
RR	255	154	6	39.776	24.678	26.852
RS	13.771	251	20	691.933	117.458	192.553
SC	1.916	406	6	193.408	139.977	86.135
SE	1.218	272	12	276.836	103.240	60.209
SP	30.657	743	42	2.644.863	102.574	853.337
TO	1.863	4.067	28	74.745	175.518	20.157

6. Quantidade de pareamentos feitos no ano de 2017

UF	nível 0	nível 1	nível 2	nível 3	nível 4	nível 5
AC	130	39	0	74.722	39.180	11.035
AL	1.364	511	0	307.863	261.589	75.122
AM	2.159	354	10	333.113	177.630	86.057
AP	2.041	128	0	84.435	12.946	7.821
BA	24.702	2.492	100	1.577.854	1.119.916	306.090
CE	33.733	2.348	21	955.951	588.980	206.094
DF	8.665	4.727	61	6.256	154.211	131
ES	2.388	816	8	181.448	160.087	70.879
GO	17.442	4.678	121	207.220	537.222	40.515
MA	4.109	912	27	564.791	765.553	142.183
MG	19.094	1.448	17	1.354.155	389.564	737.002
MS	2.866	158	0	267.697	48.492	82.205
MT	1.527	228	19	274.699	139.614	58.221
PA	43.010	1.367	73	693.881	535.204	189.437
PB	1.056	438	2	313.690	396.275	79.429
PE	33.615	1.111	25	1.177.647	349.264	282.372
PI	933	613	3	239.198	386.364	43.181
PR	15.103	1.250	1	682.292	207.844	295.716
RJ	62.466	3.064	148	921.599	276.845	329.787
RN	402	256	7	386.369	55.837	157.359
RO	2.383	359	7	156.424	43.456	25.578
RR	173	122	3	37.370	26.187	25.384
RS	12.936	240	23	665.511	108.103	192.808
SC	1.028	298	1	182.201	123.810	88.053
SE	893	72	6	280.438	104.685	62.432
SP	36.416	641	38	2.757.141	103.695	913.204
TO	3.196	5.732	69	77.130	174.813	21.320

7. Quantidade de pareamentos feitos no ano de 2018

UF	nível 0	nível 1	nível 2	nível 3	nível 4	nível 5
AC	131	22	1	75.468	39.836	11.777
AL	2.796	550	3	313.820	258.985	71.806
AM	2.071	333	5	336.948	189.553	89.638
AP	2.064	170	0	94.853	14.226	8.637
BA	30.585	2.670	109	1.572.698	1.099.748	290.525
CE	35.948	2.669	30	950.810	583.746	196.737
DF	8.655	4.152	59	6.797	129.018	167
ES	2.351	698	6	186.713	151.074	73.891
GO	15.800	3.540	81	199.276	518.545	38.154
MA	3.672	1.175	40	573.437	756.841	132.274
MG	24.604	1.830	14	1.359.680	360.316	711.502
MS	3.186	138	1	263.795	41.503	80.076
MT	1.226	214	11	281.341	134.557	56.505
PA	41.349	1.292	65	728.105	545.393	188.694
PB	562	567	1	310.801	400.057	76.592
PE	40.454	1.284	19	1.206.139	348.744	288.456
PI	1.147	982	3	237.517	388.722	41.784
PR	15.557	1.309	6	676.930	190.895	292.428
RJ	60.142	2.600	30	940.882	282.934	346.778
RN	2.602	227	9	398.863	50.994	152.107
RO	1.804	177	6	152.373	43.004	28.140
RR	102	92	4	38.148	27.522	26.885
RS	16.069	210	8	634.169	95.954	183.450
SC	748	290	2	176.292	104.406	89.790
SE	840	260	4	291.511	107.138	59.710
SP	39.778	518	28	2.803.417	101.019	936.619
TO	3.506	3.352	79	82.239	172.965	20.198

8. Quantidade de pareamentos feitos no ano de 2019

UF	nível 0	nível 1	nível 2	nível 3	nível 4	nível 5
AC	221	37	0	82.292	42.112	13.579
AL	3.314	485	2	323.854	265.114	77.327
AM	2.720	377	11	366.171	202.303	101.206
AP	1.518	160	0	107.548	16.614	11.608
BA	40.824	2.977	110	1.669.969	1.138.072	316.449
CE	42.413	2.786	19	1.002.187	606.868	210.932
DF	12.115	3.369	141	10.412	143.827	230
ES	2.622	611	12	213.229	161.114	87.764
GO	20.094	4.796	88	211.941	557.442	42.941
MA	3.691	877	26	599.541	789.381	146.574
MG	38.748	1.994	19	1.448.692	362.252	780.125
MS	4.089	129	1	283.972	42.106	89.090
MT	1.536	173	14	304.078	143.965	63.162
PA	44.922	1.374	72	790.438	584.590	219.914
PB	763	415	1	327.951	422.415	84.786
PE	47.666	1.077	25	1.271.074	362.648	315.582
PI	1.536	776	3	255.826	403.410	46.019
PR	17.433	1.265	4	734.507	193.375	321.710
RJ	68.248	3.198	26	1.036.928	315.974	398.292
RN	4.679	167	5	421.705	47.290	167.001
RO	3.087	471	6	162.146	47.608	34.791
RR	102	132	0	40.879	30.962	28.561
RS	24.274	240	14	669.079	96.826	203.717
SC	1.503	246	2	195.260	105.141	109.209
SE	788	63	3	307.382	111.995	65.139
SP	58.900	521	36	3.018.493	106.525	1.021.833
TO	4.547	3.394	116	86.521	177.432	21.804

MINISTÉRIO DO
DESENVOLVIMENTO
E ASSISTÊNCIA SOCIAL,
FAMÍLIA E COMBATE À FOME

GOVERNO FEDERAL

BRASIL

UNIÃO E RECONSTRUÇÃO