Geocodificação dos Endereços do CadÚnico de Campinas

Observatório
PUC-Campinas

PUC
CAMPINAS

fundação feac

Índice

1 • Introdução

Este documento descreve os passos metodológicos do processo de geocodificação da base do Cadastro Único (CadÚnico) do Município de Campinas*. Esse processo é parte fundamental para o Índice de Vulnerabilidade Social Intramunicipal (IVSIM – Campinas), em desenvolvimento no âmbito do Acordo de Cooperação Técnica (ACT) firmado entre a Fundação FEAC e a Pontifícia Universidade Católica de Campinas (PUC-Campinas), por meio do Observatório PUC-Campinas (OPC).

O IVSIM, como o nome afirma, pretende ser um índice capaz de proporcionar uma leitura territorial intramunicipal tendo, além dessa qualidade, a vantagem de atualização periódica. Em caso de sucesso, seria um avanço diante de índices que avaliam a vulnerabilidade intramunicipal com base nos levantamentos censitários, a exemplo do Índice Paulista de Vulnerabilidade Social (IPVS).

Para nos aproximarmos dessa pretensão, o primeiro desafio foi encontrar uma base de dados territorializável com atualização constante. Com esse objetivo, acedeu-se aos microdados do Cadastro Único. O CadÚnico é, segundo a Secretaria de Avaliação, Gestão da Informação e Cadastro Único (Sagicad), uma ferramenta de identificação e caracterização socioeconômica das famílias brasileiras de baixa renda. É um importante instrumento de planejamento de políticas públicas, pois registra informações sobre características dos domicílios, composição familiar, escolaridade, trabalho e renda, dentre outras. Sua utilização para fins de estudos e pesquisas está regulada por meio de decreto e requer o atendimento de exigências éticas no manuseio das informações.

A empreitada subsequente foi efetuar a geocodificação das famílias constantes no cadastro, ou seja, segundo Eichelberger (1993), atribuir um código geográfico a um elemento cartográfico ou do mundo real, no caso, textos de endereços.

Para além da obtenção dos microdados do cadastro mencionado com a autoridade municipal, o processo de geocodificação incluiu diversas fases, desde a compra do Diretório Nacional de Endereços (DNE) até a utilização de interfaces de aplicativos de programação (APIs)[1] para a efetiva geocodificação dos endereços. Como produto final dessa etapa do projeto, tem-se a entrega das bases do Cadastro Único com as coordenadas aproximadas dos domicílios das famílias cadastradas.

Esta nota está dividida de acordo com os principais passos empregados para a geocodificação. Assim, além desta introdução, discorreremos sobre algumas referências utilizadas, sobre o DNE, a respeito dos softwares e bases utilizados, do processo de eleição de pares de coordenadas quando mais de um estava disponível, além de dedicarmos uma seção a explicar como o Diretório serviu ao georreferenciamento do Cadastro e como conduzimos diante de endereços inconsistentes. Finalmente, em um brevíssimo fechamento, apresentamos o resultado dessa empreitada através de um mapa de distribuição das famílias do CadÚnico.

[1]Application Programming Interface se definem como rotinas disponibilizadas ao usuário em formato simplificado, sem contato direto com o código, por meio de um programa mediante apresentação de credencial, também chamada de chave API.

*PMC.2023.00007972-20

2• Algumas referências

Atualmente, existem diferentes ferramentas de geocodificação automática que permitem a associação entre endereços em formato textual e suas coordenadas geográficas. A literatura documenta pelo menos duas dificuldades em sua utilização: limites de cobertura e de precisão (QUINTEIROS et al., 2022; FEITOSA et al., 2021).

A variação de desempenho das ferramentas nesses quesitos estaria associada, dentre outros fatores, à própria existência dos endereços no banco em que se efetua a busca, mas também à qualidade dos elementos textuais que dão entrada na pesquisa (FEITOSA et al., 2021; SILVEIRA et al., 2017).

Em Quinteros et al. (2022) foram comparados Google Maps, Google Earth e Bing para duas cidades chilenas. Em relação à cobertura, as ferramentas Google tiveram melhor desempenho em ambas as localidades, enquanto Bing falhou totalmente naquela de base rural e menor densidade demográfica. A precisão foi avaliada com um confronto entre as coordenadas obtidas de forma automática e aquelas levantadas por procedimento manual. Nesse tocante, o resultado revelou que, uma vez encontrado o endereço, a exatidão foi maior com Bing.

Diante da heterogeneidade dos resultados entre as cidades, os autores concluem pela necessidade de estabelecer protocolos de geocodificação específicos para a localidade de interesse do pesquisador.

Por sua vez, em uma empreitada bastante parecida àquela enfrentada aqui, Feitosa et al. (2021) optaram pela combinação de fornecedores de coordenadas para geocodificar endereços constantes no CadÚnico de seis cidades espalhadas pelas cinco regiões do Brasil, num total de 2.272.185 endereços. Foram utilizadas as ferramentas Galileo, Here e Geocoding with Google Sheets.

As duas primeiras classificam as coordenadas pelo nível de precisão. Galileo estratifica em quatro níveis, enquanto Here em nove (considerando de mesmo nível as coordenadas com precisão de tipo house number e building). Já o procedimento Geocoding with Google Sheets não possuía métrica de avaliação. Diante dessa especificidade e do tempo observado para processamento, os pesquisadores deram prioridade ao Galileo, aceitando todas as coordenadas com máxima avaliação de precisão, nível 4, cerca de 58,5%.

Os endereços remanescentes foram geocodificados por meio de Here e Google e, novamente, a equipe optou por aceitar a métrica de acurácia disponível, atribuindo as coordenadas do Here a todos os endereços classificados com a máxima precisão.

Os registros restantes, cerca de 600 mil endereços, tiveram as coordenadas obtidas por Google e Here confrontadas, sendo a eleição entre elas dependente da escolha do analista, mas com prerrogativa do Google. Este apresentou maior alcance territorial, especialmente nas áreas de expansão urbana, o que pôde ser aferido pelo espalhamento dos pontos geocodificados nos territórios de interesse.

Tendo em vista essas referências e considerando as limitações, como a inviabilidade de levantamento manual de coordenadas e a dificuldade de acessar o Galileo, além de especificidades dos endereços do CadÚnico a que tivemos acesso, desenhou-se uma estratégia própria de geocodificação descrita a seguir. Dentre as vantagens que ela apresenta, estão o contorno das inconsistências nos endereços que desejamos geocodificar e a criação de critérios de seleção internos para as coordenadas retornadas por variados serviços.

3. Uso do Diretório Nacional de Endereços (DNE)

Visando minimizar uma das dificuldades apresentadas pela literatura, a interferência de erros de grafia e abreviações no texto dos endereços, em vez de geocodificar diretamente a base do CadÚnico, optou-se por geocodificar os endereços de Campinas a partir do Diretório Nacional de Endereços (DNE), também conhecido por Base de CEP. Essa é uma base de dados restrita, oficialmente vendida de maneira exclusiva pelos próprios Correios em portal online[2]. Para a finalidade anunciada, utilizaram-se dois dos arquivos do DNE do tipo fixo, aqueles que contêm os códigos postais no maior nível de desagregação, logradouros e grandes usuários (DNE_GU_LOGRADOUROS e DNE_GU_GRANDES_USUARIOS).

O arquivo de logradouro possui 10.848 endereços únicos para o município de Campinas, com informações de bairro, tipo (a exemplo de rua, avenida, viela, praça etc.), complemento de título (como doutor, professor, governador) e, finalmente, nome do logradouro, cada um relacionado a um CEP exclusivo. Já o arquivo de grandes usuários apresenta 278 registros únicos para o município em questão. Nesse arquivo se dispõe de informações como o nome do grande usuário (que pode ser um estabelecimento comercial, como um shopping, ou residencial, edifícios e condomínios, sobre os quais recaem nosso interesse), bairro e CEP.

4. Diferentes bases e softwares envolvidos

Por critérios de disponibilidade e ampla utilização, foram selecionadas três diferentes bases de geocódigos para essa atividade: Google Maps, Bing e ArcGIS. O acesso a elas pode se dar por meio de uma variedade de softwares de processamento e análise de dados. Neste trabalho, a base do Google Maps foi acedida por meio do Qgis, um software livre de SIG (Sistema de Informação Geográfica) que pode ser conectado a um complemento que viabiliza a geocodificação direta e reversa, o MMQGIS. Esse complemento acessa a API Geocoding através de uma credencial obtida pelo cadastro de um projeto na plataforma Google Cloud. Há uma cota de uso sem custo que se mostrou suficiente para nossa atividade. Cumpre ainda dizer que os endereços precisam estar livres de caracteres especiais e acentos para o correto funcionamento da API.

A base do ArcGIS e Bing foi acedida por meio do software de programação e estatística R Cran. O pacote tidygeocoder foi utilizado como facilitador no processo de georreferenciamento, uma vez que disponibiliza por meio de interface unificada suporte para geocodificação de diversas plataformas (Cambon et al., 2021). O serviço Bing, através da licença de desenvolvedor, fornece 125 mil transações de forma gratuita, mostrando-se suficiente para os dados em questão. O ArcGIS, por sua vez, não necessita de qualquer chave de acesso, quando utilizado através do pacote mencionado. Para ambos os requerimentos, o método de padronização de endereços foi igual, seguindo o mesmo padrão da pesquisa realizada no Google.

[2] *Disponível em: https://shopping.correios.com.br/. Acesso em: 11 ago. 2023.*

5• Estudo de similaridade entre os resultados obtidos

A geocodificação por meio do ArcGIS retornou coordenadas para todos os endereços do DNE pesquisados. Google Maps e Bing deixaram de retornar para apenas um e três endereços, respectivamente. Os retornos apresentam as latitudes e longitudes, o endereço que fez correspondência com aquele que deu entrada e uma métrica que avalia a precisão das coordenadas.

No caso do Google Maps, essa métrica estratifica os resultados em quatro níveis: rooftop (geocódigo preciso), range interpolated (em geral, uma via), geometric center (centro geométrico de um polígono ou polilinha), approximate (aproximação com menor nível de precisão). Os dados apontaram que 85,5% dos endereços estariam geolocalizados na categoria geometric center, a terceira em nível acurácia. Esse resultado não surpreendeu, uma vez que as informações que deram entrada tinham a exatidão máxima do logradouro.

No entanto, 282 resultados foram classificados como rooftop, 221 dos quais com a entrada do endereço sendo apenas o nome do logradouro e o CEP (sem, sequer, o nome do edifício ou do condomínio).

A impossibilidade desse tipo de retorno, uma vez que os dados de entrada não tinham qualidade para tanto, depôs contra o uso da métrica para comparar os resultados que fossem obtidos por meio das diferentes bases, procedimento que, caso executado, se aproximaria daquele apresentado em Feitosa et al. (2021).

Cabe dizer que, na atividade enfrentada pelo grupo do Instituto de Pesquisa Econômica Aplicada (Ipea), contava-se com o número dos domicílios, informação ausente neste trabalho.

Assim, ao que parece, o indicador automático de acurácia não cabe quando a geocodificação pretende identificar coordenadas associadas ao ponto médio de um logradouro. As métricas de Bing e ArcGIS apresentaram comportamento semelhante.

Diante do exposto, seria necessário encontrar outra maneira de avaliar a precisão das coordenadas, bem como eleger o par mais preciso entre os três disponíveis. Partiu-se, então, para a comparação entre os resultados e a proposição de eleição por dupla confirmação, conforme exposto adiante.

Antes de iniciar a comparação, todavia, procedeu-se a exclusão de resultados que se tratavam de erros grosseiros, assim, foram desconsideradas as coordenadas dos endereços dos correios que indicavam pontos fora dos limites do município. Isso ocorreu em 439 dos requerimentos feitos ao Bing, em 54 do ArcGIS e em 125 do Google, sendo atribuídos NA[3] para esses endereços. Após esse filtro, apenas cinco dos 11.126 ficaram sem latitude e longitude para os três serviços.

[3]Código para ausência de informação, também chamado de missing ou omissão.

5.1 Comparação entre CEPs fornecidos e retornados

Com as três bases de retornos foi feita a união para iniciar a comparação. Preliminarmente, com o objetivo de detectar retornos com alta qualidade, foi feito o processo de comparação dos CEPs, sendo eles os fornecidos no requerimento e o que foi retornado pelos serviços de API.

Portanto, quando o CEP fornecido foi igual ao retornado pelo API, entendemos estar diante de um retorno de boa qualidade. Tal processo foi implementado nos três serviços de georreferenciamento e serve como filtro inicial para a escolha final das coordenadas. A Tabela 1 traz os resultados para cada um dos fornecedores, tendo como base os endereços do DNE. O ArcGIS foi a base com o melhor resultado nessa primeira avaliação; 73,5% das 11.126 requisições retornaram CEPs iguais aos que deram entrada.

Tabela 1 – Número de compatibilidades entre o CEP de entrada e o CEP de saída

	Google	Bing	ArcGIS
Frequência	5.994	5.836	8.177
Frequência %	53,9	52,5	73,5

Conforme mencionado, cada serviço de georreferenciamento apresenta um nível de confiança ou score que determina automaticamente a qualidade da geolocalização, contudo, além da limitação apontada na seção anterior, para cada serviço, a qualidade é medida em diferentes categorias e métricas, dificultando a comparação entre elas.

A comparação entre CEPs de entrada e saída fornece uma padronização, mantendo os três resultados em uma base comparativa, permitindo a implementação de medidas adicionais que colaborem para a escolha do melhor par de coordenadas.

Nesse sentido, não excluímos um georreferenciamento em detrimento de outro pautado apenas em uma única métrica de modo inicial. Por exemplo, caso um georreferenciamento do Google apresentasse o melhor score, que seria ROOFTOP, e o mesmo endereço apresentasse no ArcGIS um score de 99 pontos e no Bing uma "confiança" alta, não teríamos uma escolha arbitrária em selecionar um par de coordenadas de algum dos serviços em detrimento dos outros. Isso porque outras etapas irão pautar a escolha entre essas três coordenadas, sendo, além da igualdade entre CEPs fornecidos e retornados, a comparação de distância entre os pontos e o ranking de dupla confirmação (MDC), apresentados a seguir.

5.1 Comparação de distâncias e Método de Dupla Confirmação (MDC)

Se, por um lado, a verificação de compatibilidade entre CEP de entrada e saída ajuda em uma primeira leitura da qualidade dos resultados, por outro, ainda não aporta elementos suficientes para eleger entre um ou outro par de coordenadas encontrados nas diferentes bases.

Diante da ausência de um critério definitivo que indicasse a escolha de uma base em detrimento das outras, buscou-se critérios de avaliação internos, com base nos resultados retornados. Partiu-se para o confronto entre as coordenadas retornadas para um mesmo endereço pelos diferentes serviços por meio do cálculo da distância entre elas. Assim, calculou-se a distância entre os resultados dados por Bing e ArcGIS, Bing e Google e, por fim, ArcGIS e Google.

Com base nessas distâncias, cunhou-se o que chamamos de Método de Dupla Confirmação (MDC). Seu objetivo é escalonar os três serviços de geocodificação, servindo como critério final para a escolha entre um par de coordenadas ou outro. A seguir, descrevemos como o MDC foi criado e também, efetivamente, como o utilizamos.

Etapa I – Criando o MDC

Na etapa I, buscou-se hierarquizar Bing, Google e ArcGIS, identificando suas capacidades de retornar resultados que pudessem ser chancelados por outra base de geocodificação. Partiu-se da base gerada no item 5.1, com todos os pontos retornados pelas três bases de geocodificação.

Primeiro, quando havia pares de coordenadas nas três bases para um mesmo endereço, analisava-se a distância linear entre os três pontos. Identificados os pares de coordenadas que mais se aproximavam, assumia-se que tal proximidade indicava a confirmação desses pares específicos, em detrimento do mais distante (não houve casos em que os três pontos eram exatamente equidistantes).

Dada a confirmação, computou-se um ponto para as duas bases que compunham esse par de pontos próximos. No caso de haver apenas dois serviços com coordenadas disponíveis, entendeu-se que havia confirmação entre eles, atribuindo-se um ponto para ambos, embora não houvesse, evidentemente, qualquer comparativo de distâncias. Por fim, na presença de um único par de coordenadas, a confirmação não se efetuaria, não se atribuindo pontuação.

Finalmente, com os pontos atribuídos, computou-se a frequência geral com que cada uma das três bases de geocodificação obteve "sucesso". Essa frequência de sucessos foi utilizada para ordenar as bases em relação à qualidade na geocodificação. A Tabela 2 resume a ocorrência de sucessos.

Tabela 2 – Frequência de sucessos das bases (MDC)

Base	Frequência de sucessos	Ranking
ArcGIS	9.022	1º
Google	8.376	2º
Bing	4.760	3º

\Portanto, com base nessa tabela de frequência, estabeleceu-se um critério geral de qualificação das bases utilizadas para geocodificação neste trabalho.

A etapa II do MDC utiliza esses resultados, em conjunto com o confronto entre CEPs de entrada e saída, além da comparação de distância entre pontos geocodificados para atribuir um par de coordenadas final a cada endereço.

Etapa II – Utilizando o MDC na escolha das coordenadas

Nessa etapa foi feita a atribuição final do par de coordenadas a cada endereço da base do Correio. Para cada pesquisa de endereço nas bases de geocodificação, é possível a ocorrência das seguintes situações: obtenção de 3, 2, 1 ou 0 pares de coordenadas geográficas.

Para o caso de a pesquisa retornar três pares de coordenadas, a escolha entre os diferentes pares é feita da seguinte maneira: primeiro é verificado se os três pares possuem CEPs retornados iguais aos fornecidos. Caso essa condição seja verdadeira, os dois serviços que contenham a menor distância entre os pontos geolocalizados serão analisados e a escolha final entre as coordenadas dadas por eles será feita com base no ranking do MDC.

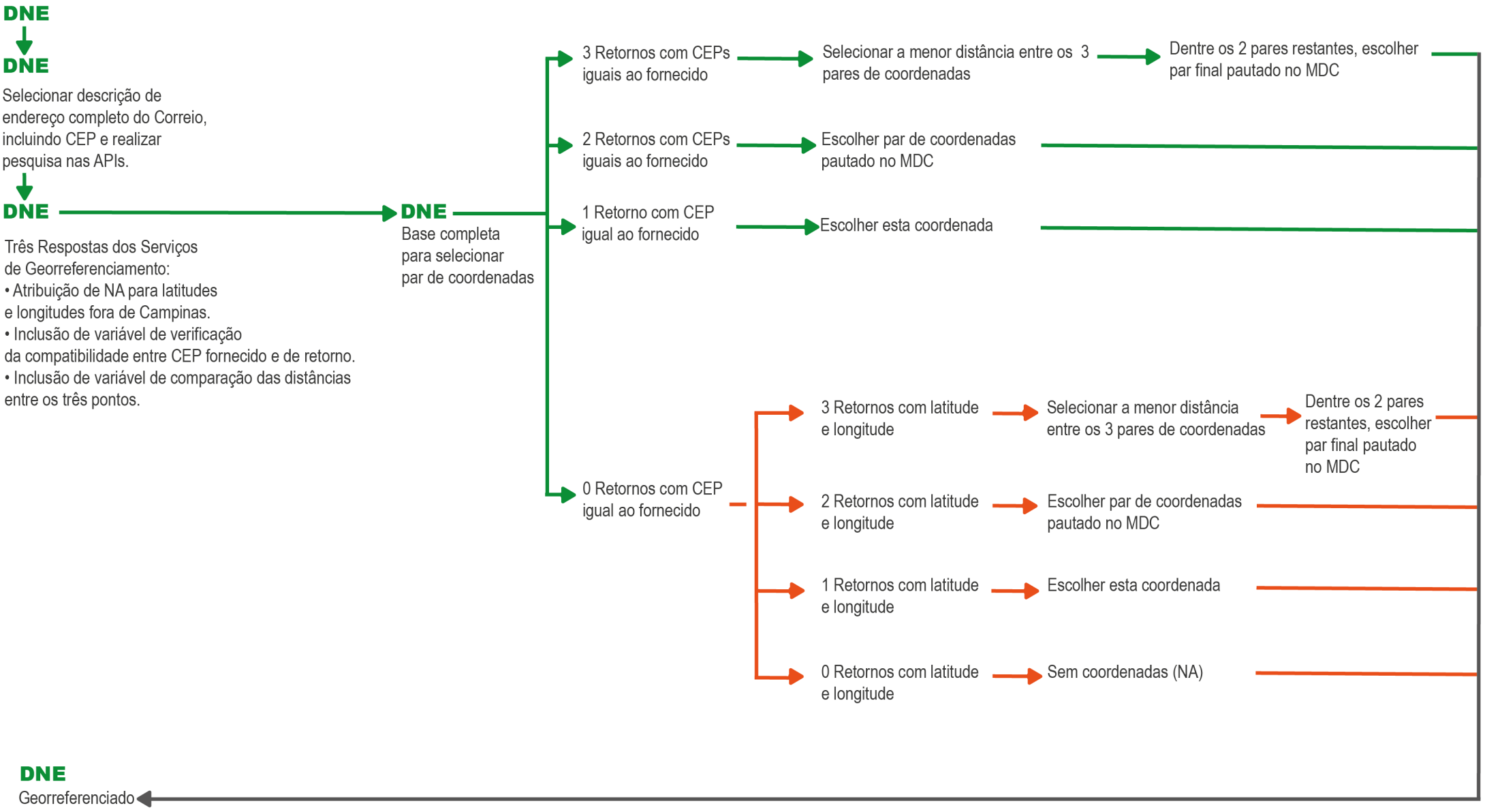
Outro cenário possível é aquele em que apenas dois pares de coordenadas tenham CEPs retornados iguais aos fornecidos. Nesse caso, o par de coordenadas final será elegido diretamente pela observação do ranking. Por exemplo, se apenas os pares de coordenadas de Google e ArcGIS possuem CEPs retornados iguais aos fornecidos, a coordenada final será a do ArcGIS, pois ele é o primeiro no ranking do MDC.

Seguindo a mesma lógica, caso apenas um serviço tenha o par de coordenadas com CEP retornado igual ao fornecido, este será escolhido como final. Essa etapa seleciona os melhores candidatos, uma vez que o filtro inicial é a igualdade de CEPs.

Resta o caso em que nenhum serviço retornou CEP idêntico àquele que deu entrada. Nessa situação, partimos para observar os dois serviços que apresentam a menor distância entre os pontos geolocalizados; aquele melhor posicionado no ranking do MDC terá o par de coordenadas selecionado como final. De forma semelhante, caso existam apenas dois pares de coordenadas, avalia-se diretamente pelo MDC. Finalmente, na eventualidade de haver apenas retorno de um serviço, essa coordenada é escolhida como final.

A métrica final traz que, para os 11.126 endereços do Correio, 8.691 foram georreferenciados pelo ArcGIS, 2.062 pelo Google e 368 pelo Bing. Apenas cinco endereços ficaram sem qualquer geolocalização[4]. O Fluxograma 1 resume o processo de geocodificação do DNE.

[4] No confronto desses resultados com o ranking do MDC, é fácil perceber que a distribuição das coordenadas finais entre os serviços não surpreende. O Bing, por exemplo, como última opção do MDC, só é utilizado em duas situações. A primeira, caso apenas ele possua geolocalização com CEP retornado igual ao fornecido. A segunda, quando apesar de não ter CEP retornado igual ao fornecido, ele seja o único ponto disponível.


DNE
DNE
Selecionar descrição de endereço completo do Correio, incluindo CEP e realizar pesquisa nas APIs.
DNE
DNE
Três Respostas dos Serviços de Georreferenciamento:
• Atribuição de NA para latitudes e longitudes fora de Campinas.
• Inclusão de variável de verificação da compatibilidade entre CEP fornecido e de retorno.
• Inclusão de variável de comparação das distâncias entre os três pontos.
DNE
Base completa para selecionar par de coordenadas
3 Retornos com CEPs iguais ao fornecido
Selecionar a menor distância entre os 3 pares de coordenadas
Dentre os 2 pares restantes, escolher par final pautado no MDC
2 Retornos com CEPs iguais ao fornecido
Escolher par de coordenadas pautado no MDC
1 Retorno com CEP igual ao fornecido
Escolher esta coordenada
0 Retornos com CEP igual ao fornecido
3 Retornos com latitude e longitude
Selecionar a menor distância entre os 3 pares de coordenadas
Dentre os 2 pares restantes, escolher par final pautado no MDC
2 Retornos com latitude e longitude
Escolher par de coordenadas pautado no MDC
1 Retornos com latitude e longitude
Escolher esta coordenada
0 Retornos com latitude e longitude
Sem coordenadas (NA)
DNE
Georreferenciado

6• Georreferenciando a Base do Cadastro Único

Recordando o objetivo finalístico dessa empreitada, ainda precisamos discorrer, efetivamente, sobre como o trabalho de geocodificação do DNE serviu para o CadÚnico. Porém, antes de mais nada, procedeu-se ao estudo detalhado das variáveis de localização disponíveis no cadastro.

Isso foi fundamental para esclarecer que os endereços dessa base deveriam ser divididos em dois grupos, um que possui endereçamento quase padrão, podendo receber as coordenadas obtidas via DNE, e outro que apresenta um descolamento entre o nome do logradouro e o código postal (lembremo-nos de que, em Campinas, o CEP adere ao logradouro), o que exigiria uma estratégia distinta.

6.1 Comparação entre CEPs fornecidos e retornados

A base do CadÚnico utilizada para georreferenciamento refere-se a extração em 12 de novembro de 2022. Ela é originada de um cadastro que sofre incremento contínuo, em que o responsável pela família responde sobre a situação no momento da entrevista, não há, portanto, uma data única de referência para a informação.

Todavia, para que o registro se mantenha ativo, é necessário que a família faça a atualização dos dados a cada dois anos (BRASIL, 2017). Sendo assim, o CadÚnico 2022 contém informações inseridas em até dois anos antes da data de extração.

Ainda assim, a frequência da variável que informa a data da última atualização do cadastro (variável dat_atual_fam) mostra a presença de registros com última modificação feita há mais de 24 meses (contados desde 12 de novembro de 2022). Entendidos como remanescentes e visando resguardar a aderência das informações à realidade atual, excluímos esses cadastros da base.

Outra remoção efetuada foram das famílias com renda per capita maior que R$ 606, meio salário mínimo em 2022 (variável vlr_renda_media_fam), pois a prioridade de cadastramento é daquelas com renda igual ou inferior a esse valor (BRASIL, 2017), o que desaconselha a leitura da realidade de famílias sem essa condição por meio do cadastro.

Essas exclusões não têm relação direta com a geocodificação, mas prepara a base para os desenvolvimentos futuros relativos ao índice, uma vez que consiste na supressão de dados sem interesse. Além disso, respondem parcialmente pela explicação da discrepância entre o número de famílias constantes no cadastro originalmente recebido da autoridade municipal e aquele que deu entrada no georreferenciamento.

Finalmente, como o interesse recai sobre a geocodificação do domicílio das famílias do cadastro, também procedeu-se à exclusão de pessoas em situação de rua (variável marc_sit_rua), que dão entrada com o endereço do serviço socioassistencial de referência ou da instituição de acolhimento informada pelo cadastrado (BRASIL, 2017). Assim, a base inicial continha 295.288 indivíduos em 126.809 famílias.

Após aplicação dos filtros descritos, restaram 174.505 indivíduos em 71.466 famílias.
Para essas 71.466 unidades familiares, o cadastro recebido continha as variáveis de endereço constantes na Tabela 3.

Tabela 3 – Códigos, definições e frequências de omissões nas variáveis de endereçamento do CadÚnico, 2022

Código	Definição	N. de missings
cd_ibge	código do IBGE do município	0
nom_localidade_fam	nome da localidade da família	0
nom_tip_logradouro_fam	tipo do logradouro da família	0
nom_titulo_logradouro_fam	título do logradouro da família	61.598
nom_logradouro_fam	nome do logradouro da família	0
num_cep_logradouro_fam	CEP	0

Fonte: Elaboração própria a partir dos dados do CadÚnico (2022).

Nota-se que, após os filtros indicados, não houve missings para as variáveis de endereço, exceto para o caso da variável com o título do logradouro da família. O elevado volume de supostas omissões nesse quesito não surpreende, nem indica uma dificuldade, afinal, a minoria dos logradouros no Brasil apresenta título. É válido ressaltar que o número da residência dos domicílios cadastrados não foi fornecido por questões éticas e legais.

Dada sua natureza numérica e específica, optou-se por utilizar como variável principal no processo de georreferenciamento o Código de Endereçamento Postal (CEP) variável – num_cep_logradouro_fam. As demais variáveis de endereço foram utilizadas de forma secundária para a maior parte do processo.

A leitura de que estaríamos diante de dois grupos com qualidades distintas de dados de endereçamento no CadÚnico apareceu na análise da frequência de famílias por CEPs. Encontrou-se grande concentração em alguns códigos. Por um lado, essa concentração poderia representar simplesmente a presença de muitas famílias em uma mesma localidade, o que, dados os padrões de segregação socioespacial, não seria improvável. No entanto, também havia a possibilidade de estarmos diante de problemas no preenchimento das informações. Cabe dizer que, diferente de um cadastro vinculado a uma base de dados dos correios, em que CEP e logradouro têm uma associação inequívoca, no CadÚnico, pode-se dar entrada em um CEP que não corresponde ao logradouro informado. Portanto, inconsistências no preenchimento são inevitáveis.

O diálogo com a Administração Pública revelou que, no momento do cadastro, algumas situações podem deliberadamente levar ao descolamento entre CEP e logradouro, quais sejam, quando o logradouro ou CEP informados pelo responsável não são localizados no site de buscas dos correios, ou ainda quando a residência está em loteamento irregular em que não há endereço oficial. Nesses casos existe a prática de inserir um CEP qualquer, especialmente, códigos do centro da cidade.

De posse do conhecimento dessa prática, arbitrariamente convencionou-se que os CEPs com duzentas ou mais famílias registradas passariam pela análise descrita a seguir para que se pudesse avaliar sua condição de CEPs genéricos (ou seja, códigos inseridos em discordância com o logradouro oficialmente associado). Oito códigos cumpriam a condição de apresentar duzentos ou mais cadastros, por isso, foram selecionados, perfazendo o total de 7.540 unidades familiares. Quatro deles se confirmaram como genéricos; a eles estão associadas 6.507 famílias, sendo que 4.272 não residem no endereço associado ao CEP cadastrado. A Tabela 4 apresenta esses números, sendo a supressão do sufixo dos códigos utilizada como recurso para reforçar a desidentificação.

Tabela 4 – CEPs genéricos no CadÚnico 2022

CEPs	Frequência no CadÚnico	Frequência de descolamento
13010-**a	4.846	3.266
13053-***	1.153	671
13010-**b	305	199
13100-***	203	136
Subtotal	6.507	4.272
13058-***	208	96
13059-**a	200	14
13059-**b	349	3
13059-**c	276	2
Total	7.540	4.387

Fonte: Elaboração própria a partir dos dados do CadÚnico (2022).

Convém esclarecer como a avaliação foi realizada. Procedeu-se à pesquisa no ArcGIS dos elementos textuais de endereços do CadÚnico associados aos oito CEPs selecionados, ou seja, buscou-se pela combinação de: tipo, título e nome do logradouro, mais localidade (frise-se, sem CEP).

Quando o endereço de retorno continha um CEP diferente daquele associado à combinação no registro do CadÚnico, atribuiu-se o valor 1; na ocasião em que o CEP de retorno era igual àquele associado à combinação, o valor atribuído foi igual a 0. Em seguida, observou-se a frequência do valor 1, que representa o número de vezes com que o CEP associado à combinação de elementos textuais de endereço no CadÚnico foi diferente do CEP encontrado no ArcGIS[5] para essa mesma combinação. Em outras palavras, esta pode ser entendida como a frequência com que determinado CEP entrou no CadÚnico com um logradouro diferente daquele a que está oficialmente vinculado. Assim, estamos diante de uma frequência de descolamento.

Todos os códigos destacados em azul na Tabela 4 foram identificados como genéricos, uma vez que frequentemente descolam do logradouro a que deveriam estar associados. Resta comunicar que o código 13058-*** não foi assim identificado pois a análise mostrou que, mesmo nos 96 casos de descolamento, a maioria dos retornos do ArcGIS tinha elementos textuais de endereço idênticos aos de entrada, além de um CEP igual ao do CadÚnico até o nível de subsetor[6].

[5] *O serviço de geocodificação com maior MDC para Campinas.*

Adicionalmente à situação descrita, identificou-se famílias no Cadastro Único com CEPs de unidades operacionais dos correios. Essa estratégia possivelmente tem a mesma origem da problemática que se vem tratando até aqui: na ausência ou no desconhecimento do CEP, o código de uma unidade dos correios foi imputado para cumprir a exigência de preenchimento do campo no formulário de cadastro. Uma vez que esses códigos tampouco respondem pela localização da família, eles também foram classificados como CEPs genéricos.

Diante do exposto, para o georreferenciamento, o banco do CadÚnico foi dividido em dois grupos: um com famílias que apresentam endereços em formato quase padrão e outro com famílias que foram cadastradas com CEPs genéricos.

6.2 CadÚnico A: endereços em formato quase padrão

Famílias cadastradas com endereços quase padrão são aquelas em que não se identificou inconsistência entre os elementos textuais de endereço e o código postal. É evidente que, uma vez que os campos "nome do logradouro" e "localidade" são preenchidos de forma manual, aparentemente sem qualquer validação instantânea, há infinitos exemplos de registros com erros de digitação, abreviações e apelidos que dificultariam a geocodificação direta, conforme nos havia adiantado a literatura.

No caso desses endereços, a maneira mais simples de georreferenciar é pela geocodificação do DNE, etapa já descrita, chaveamento por meio do código postal e doação das coordenadas. Em outras palavras, para endereços quase padrão, quando um CEP do DNE encontrou correspondência no CadÚnico, as coordenadas foram copiadas daquela para esta base.

6.3 CadÚnico A: endereços em formato quase padrão

Finalmente, chegamos aos endereços mais difíceis de georreferenciar, aqueles em que a informação sobre o CEP na base do CadÚnico não é consistente, uma vez que CEPs genéricos não representam a localização da família. Essa etapa foi aplicada a 6.525 endereços, 6.521 atrelados a códigos genéricos (6.507 apresentados na Tabela 4, e 14 referentes a endereços atrelados a CEPs de unidades operacionais dos correios), dois vinculados a CEPs inexistentes em quaisquer arquivos do DNE e outros dois geolocalizados fora dos limites de Campinas.

A estratégia para evitar a perda dessas famílias, essenciais para a leitura adequada da territorialidade da vulnerabilidade, consistiu em utilizar as informações de endereço fornecidas na base do CadÚnico, excluindo o CEP, para pesquisa nos três serviços de georreferenciamento. Similar ao georreferenciamento do DNE, foram descartados resultados localizados fora dos limites do município (observe-se que é esperado que nesses requerimentos exista uma maior inconsistência, uma vez que a pesquisa está sendo realizada com dois níveis a menos de informação, CEP e número do domicílio, além de haver a dificuldade imposta por erros de digitação, abreviações etc.). I

sso ocorreu em 791 dos requerimentos de endereços do Bing, 368 do ArcGIS e 154 do Google, sendo atribuídos NA para essas coordenadas. Na sequência, criou-se uma verificação de compatibilidade entre o CEP retornado e aqueles disponíveis no DNE.

Essa etapa dividiu o fluxo em dois. Quando a compatibilidade acontecia, passava-se ao exame da menor distância entre os três pontos encontrados. Assim, os dois pares com menor distância entre si, de forma similar ao que já fora feito com o DNE, foram desempatados por meio do MDC. Caso houvesse apenas dois retornos com CEPs compatíveis, a escolha entre os dois pontos se deu pelo MDC.

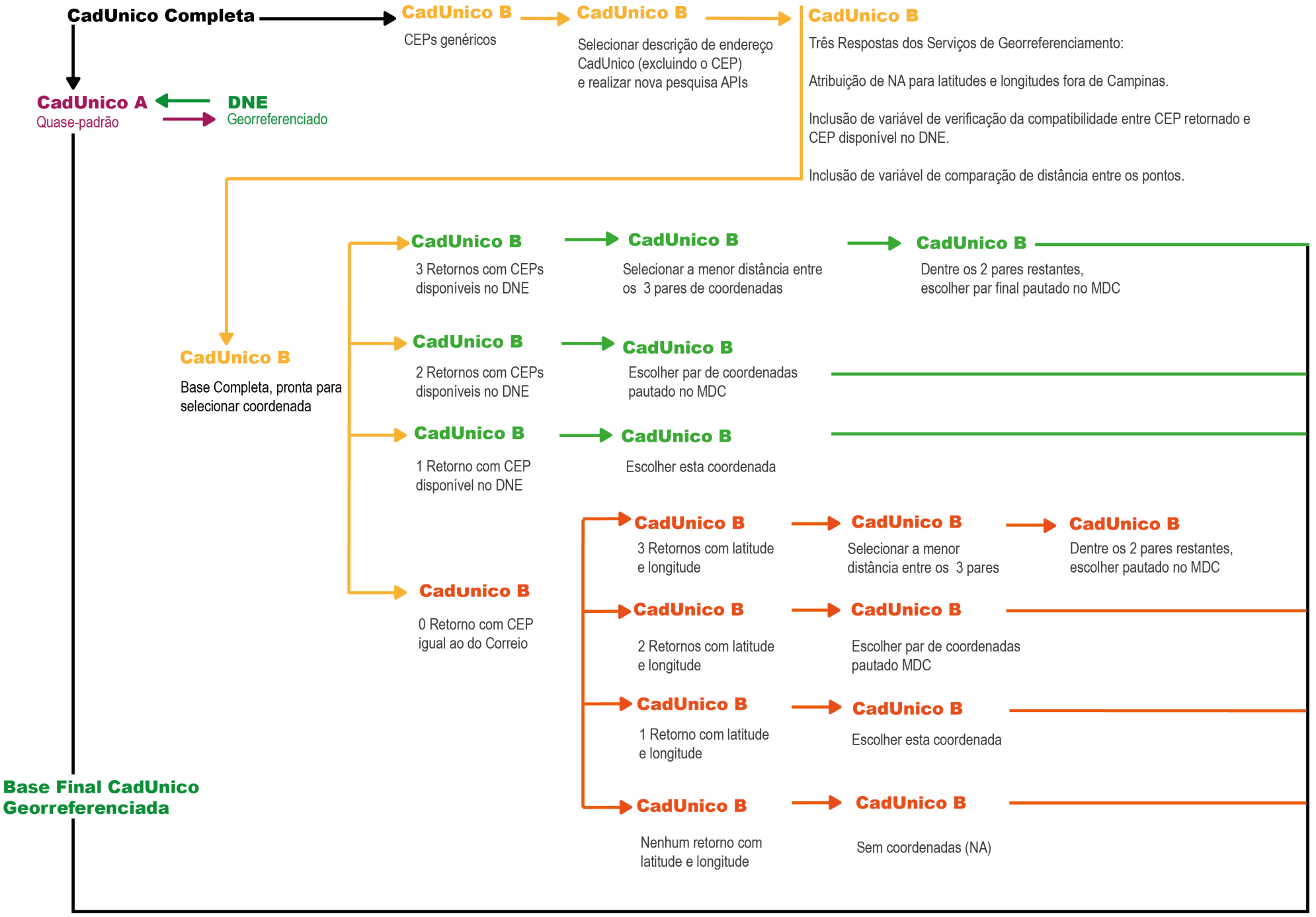
Por fim, no caso de um único retorno com CEP compatível, esse par de coordenadas foi o escolhido. Para a situação em que não houve qualquer compatibilidade, o fluxo seguido foi idêntico ao realizado para o DNE (na ocasião em que não se verificou CEP de retorno igual ao fornecido).

Finalmente, das 71.466 famílias que deveriam ser georreferenciadas, falhou-se com apenas 14, pois referem-se a endereços sem retorno de latitude e longitude em quaisquer serviços.

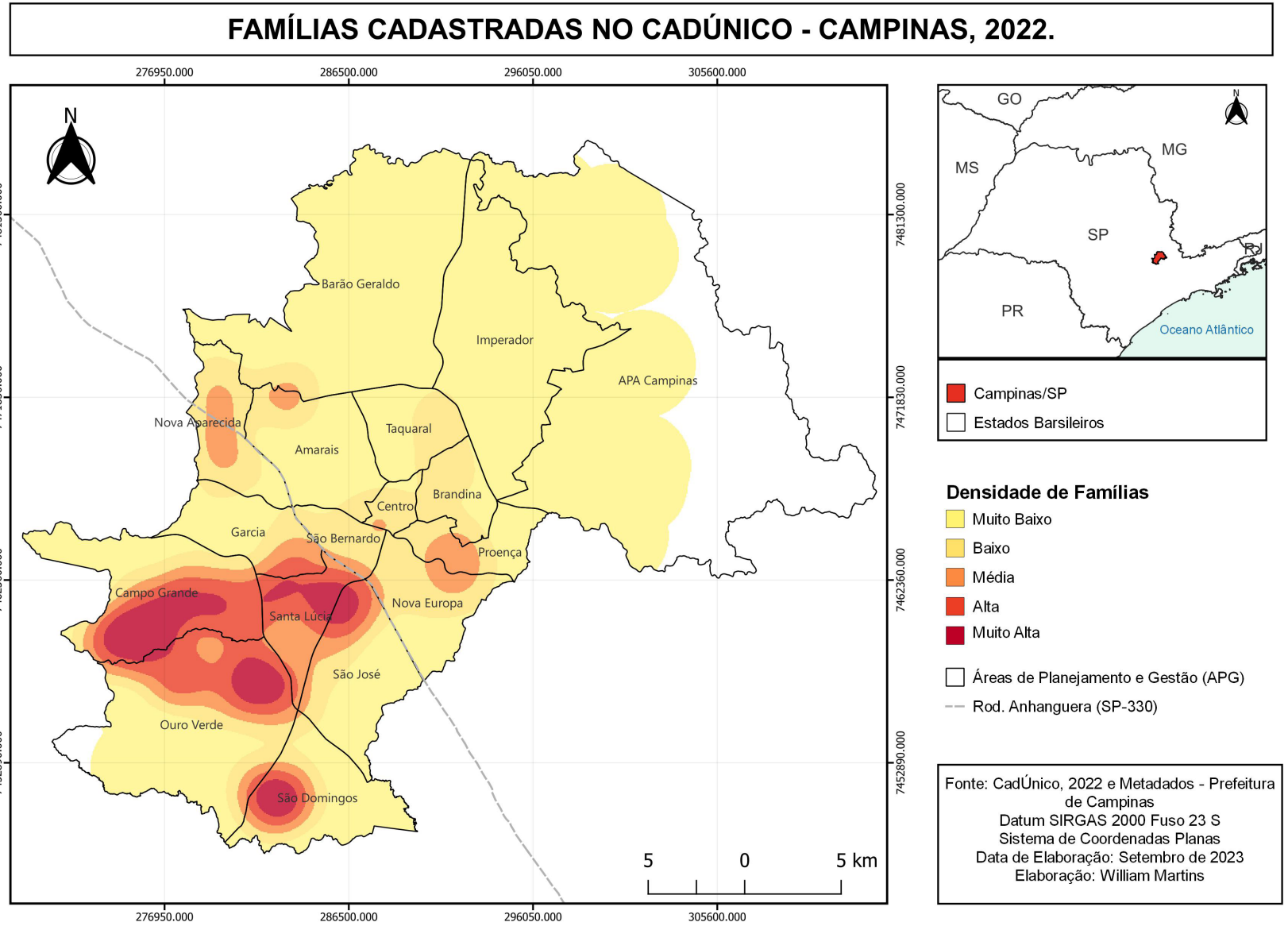
Cumpre dizer que duas famílias teriam sido assim classificadas não fosse a intervenção manual. Trata-se de residências que estão em uma fronteira ambígua do município.

Ao se utilizar os limites municipais divulgados pelo IBGE, elas não são encontradas em Campinas, mas, ao fazermos uso da camada geográfica da prefeitura, observamos que, para esta fonte, fazem parte do território campineiro.

Na sequência, pode-se visualizar uma síntese do processo de georreferenciamento descrito.


CadUnico Completa
CadUnico B
CEPs genéricos
CadUnico B
Selecionar descrição de endereço CadUnico (excluindo o CEP) e realizar nova pesquisa APIs
CadUnico B
Três Respostas dos Serviços de Georreferenciamento:
Atribuição de NA para latitudes e longitudes fora de Campinas.
Inclusão de variável de verificação da compatibilidade entre CEP retornado e CEP disponível no DNE.
Inclusão de variável de comparação de distância entre os pontos.
CadUnico A
Quase-padrão
DNE
Georreferenciado
CadUnico B
Base Completa, pronta para selecionar coordenada
CadUnico B
3 Retornos com CEPs disponíveis no DNE
CadUnico B
Selecionar a menor distância entre os 3 pares de coordenadas
CadUnico B
Dentre os 2 pares restantes, escolher par final pautado no MDC
CadUnico B
2 Retornos com CEPs disponíveis no DNE
CadUnico B
Escolher par de coordenadas pautado no MDC
CadUnico B
1 Retorno com CEP disponível no DNE
CadUnico B
Escolher esta coordenada
Cadunico B
0 Retorno com CEP igual ao do Correio
CadUnico B
3 Retornos com latitude e longitude
CadUnico B
Selecionar a menor distância entre os 3 pares
CadUnico B
Dentre os 2 pares restantes, escolher pautado no MDC
CadUnico B
2 Retornos com latitude e longitude
CadUnico B
Escolher par de coordenadas pautado MDC
CadUnico B
1 Retorno com latitude e longitude
CadUnico B
Escolher esta coordenada
CadUnico B
Nenhum retorno com latitude e longitude
CadUnico B
Sem coordenadas (NA)
Base Final CadUnico Georreferenciada

7• Breves considerações finais



O resultado do que se descreveu pode ser observado no Mapa 1.

O mapa da distribuição das famílias do CadÚnico em Campinas mostra uma segregação a partir da rodovia Anhanguera (SP-330). As porções ao sul e oeste são aquelas com maior densidade de cadastrados.

O confronto desse resultado com outras leituras produzidas do município a partir de dados censitários de 2000 e 2010 tem aderência notável.

Ele reforça uma expressão proposta ainda no primeiro decênio deste século de que a Anhanguera divide Campinas em uma cordilheira da pobreza e outra da riqueza (CUNHA et al. 2005 apud CUNHA e JIMÉNEZ, 2006). Dessa forma, a espacialização apresentada tem o mérito de fornecer subsídios preliminares para indicação das áreas que devem receber atenção prioritária, em que pese a necessidade de qualificar os resultados apresentados com um perfil dessas famílias e áreas, desenvolvimento vindouro sob o índice. .

8• Referências

CAMBON, J. et al. tidygeocoder: An R package for geocoding. Journal of Open Source Software, v. 6, n. 65, p. 3544, 2021. Disponível em: https://doi.org/10.21105/joss.03544. Acesso em: set. 2023.

CUNHA, J. M. P. et al. Expansão metropolitana, mobilidade espacial e segregação nos anos 90: o caso da Região Metropolitana de Campinas. In: ENCONTRO NACIONAL DA ANPUR, 11., 2005, Salvador. Anais... Bahia: ANPUR, 2005.

CUNHA, J. M. P.; JIMÉNEZ, M. A. Segregação e acúmulo de carências: localização da pobreza e condições educacionais na Região Metropolitana de Campinas. In: CUNHA, J. M. P. (org.). Novas metrópoles paulistas: população, vulnerabilidade e segregação. Campinas: Núcleo de Estudos de População-NEPO/Unicamp, 2006.

EICHELBERGER, P. The Importance of Addresses: The Locus of GIS. In: URISA Annual Conference, Atlanta, Georgia, 1993. p. 200-211.

FEITOSA, F. da F. et al. Termo de Execução Descentralizada n. 01/2019 SNH/MDR e Ipea: Pesquisa de núcleos urbanos informais no Brasil. Rio de Janeiro,: Ipea, 2021. 27 p.

QUINTEIROS, M. E. et al. Quality of automatic geocoding tools: a study using addresses from hospital record files in Temuco, Chile. Cadernos de Saúde Pública, v. 38, n. 1, 2022, p. e00288920. Disponível em: https://doi.org/10.1590/0102-311x00288920. Acesso em: set. 2023.

SEADE. Fundação Sistema Estadual de Análise de Dados (org.). Índice Paulista de Vulnerabilidade Social. São Paulo: Governo do Estado, 2013. 18 p.

SECRETARIA NACIONAL DE RENDA DE CIDADANIA (BRASIL). Manual do Entrevistador: cadastro único para programas sociais. 4. ed. [S. L.]: The Union – Smas – Setor de Múltiplas Atividades Sul, 2017. 158 p.

SILVEIRA, I. H. da et al. Utilização do Google Maps para o georreferenciamento de dados do Sistema de Informações sobre Mortalidade no município do Rio de Janeiro, 2010-2012. Epidemiologia e Serviços de Saúde, v. 26, n. 4, nov. 2017, p. 881-86. Disponível em: https://doi.org/10.5123/S1679-49742017000400018. Acesso em: set. 2023.