RACEF
Revista de Administração, Contabilidade e Economia da **Fundace**

FUNDACE

Ribeirão Preto, Novembro de 2023
Edição: v. 14, n.4 (2023)

DESAFIOS PARA A EXPANSÃO DO SANEAMENTO EM ÁREAS RURAIS: CARACTERIZAÇÃO DE RURALIDADES POR MEIO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA

CHALLENGES FOR THE EXPANSION OF WATER SUPPLY AND SANITATION IN RURAL AREAS: CHARACTERIZATION OF RURALITIES THROUGH

DOI: HTTP://DX.DOI.ORG/10.13059/RACEF.V14I4.1125

Diego Augustus Senna
augustus.senna@yahoo.com.br
Universidade Federal de Minas Gerais

Luiz Philippe Pereira
luizphilippep@gmail.com
Universidade Federal de Minas Gerais

Sonaly Rezende
srezende@desa.ufmg.br
Universidade Federal de Minas Gerais

Data de envio do artigo: 02 de Junho de 2023.
Data de aceite: 19 de Setembro de 2023.

Resumo: As distintas formas de ocupação do território rural brasileiro produzem demandas específicas que devem ser consideradas em políticas públicas. O cenário configurado diante da atualização do Marco Legal do Saneamento, e as incertezas quanto ao atendimento das áreas rurais, reforçam a importância de se reafirmar os modos de vida associados às múltiplas ruralidades. O Programa Nacional de Saneamento Rural (PNSR) definiu uma tipologia de ruralidades para orientar a criação de matrizes tecnológicas para o saneamento. Essa tipologia resultou da agregação de setores censitários do IBGE, vinculando ruralidades específicas a diferenças no atendimento. Neste artigo, algoritmos de clusterização, redes neurais e Random Forests foram aplicados para delinear as ruralidades. Os resultados apontam para notáveis diferenças no alcance do saneamento entre as ruralidades do PNSR, revelando-se aderência dessa divisão às condições encontradas. Este trabalho reforça a importância da análise de dados para a compreensão das desigualdades rurais, auxiliando em tomadas de decisão.

Palavras-chave: Ruralidades; Saneamento Rural; Big Data; Mineração de dados; Aprendizado de Máquina; Clusterização; Redes Neurais; *Random Forests*.

Abstract: *The distinct forms of occupation in the Brazilian rural territory produce specific demands that must be considered in public policies. The configured scenario given the update of the Sanitation Legal Framework, and the uncertainties regarding service in rural areas, reinforce the importance of reaffirming ways of life associated with multiple ruralities. The National Rural Sanitation Program (PNSR) defined a typology of ruralities to guide the creation of technological matrices for water supply and sanitation. This typology resulted from the aggregation of IBGE's census sectors, linking specific ruralities to differences in service. In this article, clustering, neural networks, and Random Forests algorithms were applied to delineate ruralities. The results point to notable differences in the reach of water supply and sanitation among the ruralities of the PNSR, revealing adherence of this division to the conditions encountered. This work reinforces the importance of data analysis for understanding rural inequalities, helping in decision-making.*

Keywords*: Ruralities; Rural Sanitation; Big Data; Data Mining; Machine Learning; Clustering,;Neural Networks; Random Forests.*

1 INTRODUÇÃO

Com a atualização do Marco Legal do Saneamento pela Lei 14.026/2020 (BRASIL, 2020), a prestação de serviços públicos de saneamento por meio de autorização a usuários organizados em cooperativas ou associações, destinadas a atender às demandas rurais, foi desautorizada. Criaram-se empecilhos à gestão multiescalar, caracterizada pelo Programa Nacional de Saneamento Rural (PNSR) como ações de natureza intersetorial, estruturais e estruturantes, em que "cada setor da sociedade, do usuário ao Poder Público Federal, detém responsabilidades sobre ações e políticas desenvolvidas" (BRASIL, 2019; p. 116).

Os municípios têm autonomia para prover a descentralização dos serviços de saneamento e sua integração em todo o território e o PNSR trouxe à tona a multidimensionalidade do rural, importante para o fortalecimento de políticas públicas que considerem a diversidade inerente ao território brasileiro, valorizando os diversos atores considerados necessários para a construção de soluções perenes e sustentáveis. A qualificação de distintos contextos rurais, e respectivas demandas, revelou-se importante com o reconhecimento de que o alcance de soluções técnicas e de gestão é pautado em ruralidades (BRASIL, 2019).

O processo de construção do PNSR foi desenvolvido a partir dos marcos teórico-conceituais e metodológicos que orbitam a ruralidade, tema incorporado às reflexões de naturezas antropológica, socioambiental e demográfica. Tais estudos mostraram formas de se reconhecer vínculos entre distintos

lugares rurais, em suas múltiplas funções e significados, e as demandas de saneamento de seus habitantes. O aprofundamento teórico alcançado e a possibilidade de diálogo e reflexão coletiva moldaram a delimitação de "rural" assumida pelo PNSR.

Deste modo, foi possível definir conceitualmente os termos "rural" e "ruralidades" e criar estratégias de agrupamento de áreas com características homogêneas em termos de distribuição espacial de domicílios, considerando que, para o delineamento de soluções sanitárias – coletivas ou individuais –, é necessário o atendimento a premissas relacionadas ao adensamento populacional e às distâncias a centros urbanos consolidados. A interpretação de dada ruralidade qualifica um modo de vida e revela os distintos contextos. Assim, a categorização de ruralidades desenvolvida e apresentada pelo PNSR adota recortes que favorecem a operacionalização das demandas, permitindo novos agrupamentos de domicílios para os quais são definidos blocos de soluções. Com isso, a política de saneamento pode adequar-se à lógica das demandas, respeitando-as e fomentando meios de reduzir o déficit persistente nas áreas rurais.

Este trabalho abrange as perspectivas do Big Data e da mineração de dados (Data Mining – DM), sustentadas por mecanismos de aprendizado de máquina, visando à avaliação da metodologia adotada pelo PNSR para criar categorias de ruralidades. Big Data é a ciência responsável por estudar técnicas voltadas à administração e ao tratamento de grandes quantidades de dados. DM é um processo de exploração de dados, frequentemente associado ao contexto do Big Data. Aprendizado de máquina compreende sistemas que podem modificar automaticamente seu comportamento tendo como base a sua própria experiência – ou seja, aprendem a partir dos dados. A metodologia consiste em aplicar algoritmos às informações do Censo Demográfico de 2010 (IBGE, 2011), referência do PNSR, para identificar padrões relevantes de soluções de saneamento e analisar se os agrupamentos estabelecidos apresentam aderência aos respectivos padrões.

Objetiva-se explorar a existência de padrões entre soluções de saneamento – adequadas e precárias – e tipologias de localidades rurais, considerando a unidade de análise de dados estabelecida pelo IBGE, o setor censitário, de forma a determinar ocorrências frequentes, ressaltar diferenças no atendimento entre ambientes urbanos e rurais e indicar tendências de priorização de investimentos. Duas hipóteses foram delineadas: 1) Os dados reforçam que existem distintas ruralidades, sendo a categorização importante para a indicação de soluções de saneamento; 2) Há aderência entre as soluções de saneamento presentes em cada tipo de ruralidade e a divisão por agrupamentos proposta no PNSR.

2 FUNDAMENTAÇÃO TEÓRICA

2.1. As ruralidades do PNSR: aspectos teórico-metodológicos

Há acúmulo de elementos teóricos de natureza histórica que têm conduzido pesquisas a um debate que aporta na discussão do rural e das ruralidades. É perceptível a crescente convergência em torno de conceitos dinâmicos, pautados em escalas que abrangem a complexidade dos fenômenos intrínsecos ao rural e às distintas configurações de ruralidade, dada a diversidade existente no país, sobretudo a partir das desigualdades estruturais historicamente constituídas.

Certo é que a visão reproduzida e assimilada por parcela significativa da sociedade sobre o rural, vinculada à ideia simplista de um território não urbano, tem sido cada vez mais criticada, gerando debates intensos, como os que tiveram lugar no processo de construção do PNSR. O diálogo estabelecido entre pesquisadores de distintas áreas – Antropologia, Geografia, Demografia e Engenharia Sanitária – e lideranças do Grupo da Terra, resultou na concepção conceitual e na operacionalização de um "rural" para o saneamento no Brasil.

O estudo que subsidiou a formulação do PNSR envolveu reflexões sobre o Brasil rural e suas múltiplas faces, revelando o quão distante

o país se encontrava da garantia do direito humano ao saneamento. Nesse contexto, destaca-se a participação do Grupo da Terra, composto por representantes dos movimentos sociais do campo, da floresta e das águas, que protagonizaram a elaboração da Política Nacional de Saúde Integral das Populações do Campo, da Floresta e das Águas (PNSIPCFA), lançada em 2011, pelo Ministério da Saúde (BRASIL, 2011).

O Grupo da Terra, durante o processo de construção do PNSR, mostrou grande resistência ao termo "rural", interpretado segundo a perspectiva hegemônica do capitalismo e considerado incapaz de traduzir a diversidade cultural e socioambiental dos sujeitos que produzem a vida no campo. Essa denominação, quando analisada à luz do saneamento básico, é revestida de relação desigual entre as populações urbanas e as ditas "rurais", sobretudo em função da atuação do poder público direcionada às primeiras e apoiada no princípio da economia de escala.

A partir dos marcos teóricos da Antropologia, o PNSR propôs interpretação do rural e das ruralidades à luz das demandas de saneamento. Galizoni (2021) enuncia em seu texto de subsídio ao PNSR a existência de estudiosos do conceito de rural que reconhecem sua natureza histórica repleta de significados. Propõe que o rural para o saneamento seja compreendido a partir dos sujeitos que nele habitam, segundo condições objetivas, subjetivas e simbólicas (sua distribuição no território e as dinâmicas econômica, de pluriatividade, mobilidade e sociabilidade que permeiam seu modo de vida), bem como as peculiaridades do território e da gestão dos recursos existentes.

Até a década de 1930, a sociedade agrária era dominante e se sobrepunha à urbana. No contexto de transição política que levou ao Estado desenvolvimentista de Getúlio Vargas, o historiador Sérgio Buarque de Holanda (1999) evidenciou a questão rural na perspectiva da metodologia dos opostos, visão destacada por Schwarcz (2008, p. 85) que descreve na tradição rural a "repulsa ao trabalho regular e a atividades utilitárias", associada à "pouca organização formal" então vigente no país. Galizoni (2021) reforça essa ideia a partir das obras de Euclides da Cunha, em referência ao isolamento e distanciamento de Canudos do resto das povoações brasileiras, e de Monteiro Lobato, que revelou um rural arcaico e atrasado, habitado por população que representavam um obstáculo ao plano de modernização do país.

Essa visão do Brasil rural emoldurou os preceitos da chamada Revolução Verde, justificando a importância de se modernizar os meios de produção agrícola, mesmo que isso resultasse em emigração em massa, além do agravamento da concentração de terras, em viés de valorização da monocultura e de estabelecimento da agroindústria. Galizoni (2021) ressalta os impactos dessa nova cultura sobre o saneamento básico, pela escassez hídrica relacionada ao uso intensivo das fontes, de um lado, e pela contaminação associada aos agrotóxicos e outras práticas poluidoras, de outro. Apesar desses impactos negativos, alardeou-se a importância da modernização como veículo de acesso a bens e serviços, em processo reconhecido por Graziano da Silva (1981) como precursor do "novo rural", representado pela adaptação de práticas cotidianas presentes no campo à realidade urbana. Tais mudanças trouxeram novas formas de distribuição espacial da população e novas ocupações, graças ao excedente de mão de obra existente nas áreas rurais.

Galizoni (2021) aporta também reflexões sobre a quebra de paradigma que revelou críticas às contradições exploradas na discussão urbano-rural. Na década final do século XX, em vários contextos, se revelava a integração entre os "opostos", apontando para uma nova dinâmica em que passaram a coexistir a urbanização do rural e a ruralização do urbano. A autora alude ao movimento social rural como vanguardista na forma de reivindicar direitos sociais, citando o MST, os povos da floresta e as populações tradicionais, grupos que se tornaram referência para os movimentos urbanos). A facilidade de se estabelecer comunicação, aspecto que na visão de Harvey (1989) acentuou a volatilidade de valores e práticas, também promoveu

trocas entre moradores de áreas urbanas e rurais, ampliando as bases da modernização e promovendo a assimilação das novas tecnologias.

Por sua vez, Lefebvre (1999) traduz campo e cidade como construções sociais produzidas a partir de relações sociais distintas, a despeito de serem impulsionadas por dinâmica produtiva única, pautada no capitalismo. Assim, a crescente pluriatividade entre agricultores e a ocupação não agrícola dos territórios ganharam espaço, sobretudo com o advento da agroindústria e dos loteamentos de elevado padrão socioeconômico, que reúnem a tranquilidade da vida no campo e a manutenção do acesso a bens e serviços disponíveis nas cidades.

O entendimento das ruralidades advém de debate recente no Brasil, tomando forma no início do século XXI, ganhando adeptos que o repercutem, com destaque ao Grupo da Terra. A visão das ruralidades apresenta maior capacidade de romper com as bases hegemônicas que respaldam as definições oficiais de delimitação dos territórios brasileiros, segundo critérios administrativos e tributários que ainda persistem, revelando a pouca aderência ao que se passa nesses lugares e ao sentido real da vida que ali se sustenta. Cada vez mais está presente a crítica à forma inapropriada de delineamento dos lugares rurais e seus habitantes frente a duas tendências contrastantes, representadas pelo processo de urbanização e pela resistência de grupos sociais fiéis à própria cultura.

As ruralidades, na visão de Abramovay (2000), ecoam os valores, significados e funções dos territórios e sua relação com o urbano, revelando interdependência entre o urbano e o rural. Ao refletir sobre o conceito adotado nos Estados Unidos para a definição de áreas rurais, o autor destaca que o rural não é oposto ao urbano, mas um espaço de integração às cidades, a partir de relações que se intensificam cada vez mais e promovem um continuum rural-urbano, perspectiva que se assenta na premissa de que os modos de vida e de organização social e cultural não estão condicionados a uma vinculação espacial.

Essa perspectiva parece ser o ponto de partida de Kageyama (2006), que reforça a multidimensionalidade do rural segundo suas funções, ocupações, características físicas e socioculturais que, a partir de bases normativas, teóricas e metodológicas, permitem a composição de indicadores de desenvolvimento rural com o objetivo de "obter medidas passíveis de comparação entre regiões e ao longo do tempo para captar de forma mais adequada a evolução do fenômeno" (KAGEYAMA, 2006, p. 31).

Laschefski (2021) situou a nova perspectiva para o rural, no âmbito do saneamento, no curso das trajetórias de um grupo numeroso estabelecido no território rural, o de camponeses e agricultores familiares, vis a vis grupos dominantes que se dedicam à agricultura patronal e extensiva. A concepção de Laschefski (2021) abrange três elementos principais: o primeiro situa o campesinato e a agricultura familiar como atividades exercidas por sujeitos de luta e engajamento no processo que desencadeou a transformação da sociedade, resultando em estratégias voltadas para a sua integração ao mundo capitalista. O segundo elemento retoma a superação da abordagem dicotômica descrita por Holanda (1999), com seus elementos antagônicos: campo versus cidade, urbano versus rural, interdependência versus continuum. O terceiro elemento repercute a hierarquização da exploração dos recursos nos territórios rurais, por meio da monoculturização agropecuária e florestal, da mineração, da produção hidrelétrica e da subordinação do espaço ao valor de troca para a produção de mercadorias específicas, forçando as populações do campo a conviverem com o risco ambiental crescente. Tais categorias foram adotadas como referencial para o estudo Da delimitação territorial do "rural" para um método de localização de grupos alvo do PNSR (LASCHEFSKI, 2021).

Laschefski (2021) discorre sobre as políticas setoriais contemporâneas, revelando campos que se opõem na vertente de uma política de desenvolvimento a cargo do Ministério do Desenvolvimento Agrário – voltada para a agricultura familiar e seus interesses –,

e de uma política que reforça a agroindústria ou agronegócio, no âmbito do Ministério da Agricultura, Pecuária e Abastecimento. Agrega a essa visão um panorama no qual sujeitos sociais estão inseridos em contextos socioambientais dinâmicos (campo moderno versus tradicional, cidade formal versus informal), apresentando situações econômicas desiguais. Argumenta, portanto, sobre a necessidade de adequação da gestão do território – e do saneamento – segundo suas particularidades. Na prática, as questões de saneamento estão vinculadas às regras de mercado, que são antagônicas aos princípios de direitos humanos à água e ao esgotamento sanitário e, deste modo, Laschefski (2021) enfatiza a necessidade de se realizar deslocamento conceitual visando à delimitação do novo território rural, contemplando grupos vulneráveis no contexto do saneamento básico.

Importante mencionar o estudo que buscou revelar ruralidades intrínsecas ao território, visando ao planejamento do setor agrícola, intitulado Concepções da ruralidade contemporânea: as singularidades brasileiras (MIRANDA e SILVA, 2013), fomentado pelo Instituto Interamericano de Cooperação para a Agricultura – IICA. Nele adotou-se a base municipal como unidade sujeita à reclassificação, a partir de variáveis como o tamanho da população, os percentuais de população rural e de valor agregado da produção agropecuária, a menor distância entre as localidades e a sede mais próxima, e a região de influência das cidades de economia mais dinâmica.

O estudo de Wanderley e Favareto (2013) também fomentou a criação de uma tipologia com seis categorias de rural, sendo três delas calcadas em padrões de regionalização que articulam práticas agrícolas (familiar ou patronal) ao entorno socioeconômico, na perspectiva da absorção de mão de obra (desde um contexto de dinamismo até a estagnação); uma que reflete a situação de produtividade extensiva com baixa absorção de mão de obra; e duas que denotam fragilidade, tanto do entorno socioeconômico quanto dos ecossistemas, que desfavorecem o desenvolvimento da agricultura e chegam a impedir a geração de trabalho não agrícola.

Na vertente internacional, são referências a Organização para a Cooperação e Desenvolvimento Econômico (OCDE), que adotou critérios demográficos – tamanho da população e densidade demográfica (< 150 hab/km2) – para a caracterização de territórios rurais; e a União Europeia (UE), que estabeleceu como áreas rurais aquelas que não satisfazem ao critério de densidade demográfica urbana (adotando-se quadrante de 1 km2, com densidade populacional > 300 hab/km2 e clusters de quadrantes com tamanho superior a 5 mil habitantes, agrupados segundo respectivas proximidades entre si).

Pautando-se nos aportes de Galizoni (2021), Freitas (2021) e Laschefski (2021) e na visão de um território de grande extensão física e heterogeneidade socioambiental, Rigotti e Hadad (2021) discorreram sobre estudos que se dedicam a operacionalizar novas categorias de rural, trazendo importantes contribuições teórico-metodológicas, em perspectiva demográfica, ao PNSR.

O estudo PNSR – Delimitação das áreas rurais brasileiras, de Rigotti e Hadad (2021), subsidiou a definição do rural considerando demandas de saneamento. Partindo de premissas embasadas por critérios pertinentes, relativos a estudos preexistentes, mas que, para se adequarem às necessidades do PNSR, requeriam ajustes conceituais, os autores apoiaram-se em critérios de densidade demográfica e de similaridades entre lugares próximos, em termos de adensamento populacional. Para a operacionalização do conceito utilizaram os setores censitários do IBGE, do Censo Demográfico de 2010.

O Censo Demográfico envolve um conjunto limitado de informações – dentre as quais as de saneamento básico – para o universo da população, e dados amostrais para um conjunto mais amplo de variáveis. Ao identificar a população rural, o IBGE designa setores censitários que a representam (bem como à população urbana), estabelecendo limites territoriais para o urbano e classificando-o segundo três categorias (1 - Área urbanizada de cidade ou vila; 2 - Área não-urbanizada

de cidade ou vila; e 3 - Área urbana isolada). As categorias rurais são 5 e variam de acordo com a proximidade do urbano, o adensamento populacional e o isolamento (4 - Aglomerado rural de extensão urbana; 5 - Aglomerado rural isolado – povoado; 6 - Aglomerado rural isolado – núcleo; 7 - Aglomerado rural isolado – outros aglomerados; e 8 - Zona rural, exclusive aglomerado rural).

A classificação do IBGE apresenta certo grau de abstração em sua metodologia, fundamentada em dados pré-definidos (renda, atendimento por serviços públicos, presença de infraestrutura, tipos de imóveis etc.). Os setores censitários são unidades de registro de dados que contêm, pelo menos, 400 domicílios, pautados na capacidade de atuação de um recenseador no período de coleta. Os setores urbanos, nessa configuração de aporte de dados, apresentam áreas menores e números de domicílios maiores, tendo em vista que áreas rurais são menos densas que as urbanas. Por abrangerem o universo dos dados, essas unidades são compatíveis com análises que resultam em rearranjos de domicílios e novas composições urbano-rurais.

Rigotti e Hadad (2021), na reclassificação de setores censitários, mantiveram a caracterização original dos códigos 4 a 8. As mudanças foram realizadas nos códigos 1 a 3, oficialmente urbanos. Em relação aos dois últimos (setores de código 2: de ocupação predominantemente rural e de código 3: de ocupação urbana, embora afastados, por áreas rurais, de áreas urbanas mais consolidadas), os autores consideraram os argumentos de Valadares (2014), que enfatiza não haver razões para tratá-los como urbanos, a não ser pela lei municipal que estabeleceu essa divisão. Setores 2 e 3 são, portanto, considerados rurais no PNSR.

Em relação aos setores de código 1, adotou-se divisão com base em critérios de ocupação e inserção no ambiente. Setores com baixa densidade demográfica e, simultaneamente, contiguidade a pelo menos outro setor de mesma característica, foram reclassificados como rurais (denominados 1b). Os demais setores foram mantidos como urbanos (denominados 1a). A obtenção de um parâmetro de densidade demográfica para a identificação de potenciais setores rurais fundamentou-se na análise dos setores rurais mais adensados, os de códigos 4 e 5. Após a realização de estatísticas descritivas, adotou-se densidade igual a 605 hab/km2, referente à média dos dois conjuntos de setores, relativa ao 1º quartil. Deste modo, a reclassificação de setores censitários do IBGE, de acordo com a concepção de rural do PNSR, resultou em 39.914.415 pessoas residindo em áreas rurais (21,03% do total), valor que supera substancialmente o definido pelo IBGE, no Censo Demográfico de 2010: 29,54 milhões de habitantes (15,57% do total).

A classificação adotada pelo PNSR resultou em quatro categorias de ruralidades, em 2010. A primeira agregou setores censitários urbanos de baixa densidade demográfica a setores rurais de extensão urbana, criando o grupo intitulado Aglomerações próximas do urbano. Esta categoria é composta por 9.945.562 habitantes residindo em 2.957.204 domicílios. As outras três categorias criadas foram: Aglomerações mais adensadas isoladas, caracterizadas por setores censitários originalmente definidos como urbanos pelo IBGE, agregando 1.291.422 habitantes residentes em 381.233 domicílios; Aglomerações menos adensadas isoladas, que reúnem 4.558.856 habitantes em 1.210.558 domicílios, dispostos no território em aglomerações de maior dispersão e menor adensamento; e Sem aglomerações, com domicílios relativamente próximos de aglomerações ou isolados, correspondendo a 24.118.575 habitantes residentes em 6.643.101 domicílios.

2.2. Metodologias digitais para a exploração de dados

O mundo presencia intensa e volátil transformação digital impulsionada pela Indústria 4.0. Esta representa não apenas uma revolução, mas profunda ruptura. Lenz, Wuest e Westkämper (2018) apontaram que bases de dados se tornam, cada vez mais, recurso de imenso valor, tendo em vista que podem revelar

informações imprescindíveis que não poderiam ser obtidas de outras fontes. Neste contexto, conceitos como Big Data (BD) e mineração de dados (DM) se destacam.

O termo Big Data é aplicado na literatura de variadas formas. Pode descrever grandes bases de dados, por vezes tão extensas e específicas, que dependem de equipamentos e técnicas especiais para visualização e exploração. Ultimamente, BD tem recebido abrangência muito maior, estabelecendo-se como ciência independente, responsável por ações de coleta, armazenamento, filtragem, transformação e análise de grandes quantidades de dados. Laney (2001), ao desenvolver o conceito tridimensional do BD, admitiu que dados são gerados em grande volume, em formato muitas vezes variável e com requisitos de velocidade específicos. Esse conceito foi expandido por outros autores, incorporando variáveis como veracidade dos dados, relacionada à precisão e confiabilidade da informação, ou valor agregado, ao considerar o gasto de recursos para analisar determinado conteúdo, conforme apresentado por Ducange, Fazzolari e Marcelloni (2020).

O BD pode auxiliar tomadas de decisões, uma vez que tende a fornecer visão mais ampla e precisa de determinadas situações, reduzindo o tempo de implementação e aprimorando a eficácia de soluções. Witten et al. (2016) demonstraram que, para gerar valor agregado a partir dos dados, é necessário extrair conhecimento realmente útil e aplicável. De outra forma, o mero armazenamento de informações não teria sentido. Neste contexto, Chamikara et al. (2020) destacaram a importância do DM, levando em consideração que pode revelar relacionamentos não previstos entre variáveis, fornecendo direcionamentos críticos aos detentores dos dados. Minerar dados consiste em aplicar algoritmos visando à extração de padrões e conhecimento.

Os métodos de DM podem ser divididos em duas categorias: os não-supervisionados, como visualização, clustering, detecção de outliers ou redução de dimensões, voltados principalmente ao estudo inicial de dados complexos para buscar padrões e relações, e os supervisionados, como árvores de decisão, redes neurais artificiais (RNAs), método do vizinho mais próximo (k-NN) ou máquinas de suporte vetorial, utilizados quando algumas variáveis são conhecidas e quando regras ou funções que compõem um modelo são identificáveis, possibilitando observações prévias (LIEBER et al., 2013). As duas abordagens são trabalhadas neste artigo.

BD e DM são importantes para qualquer tipo de atividade de caráter gerencial, envolvendo não apenas empresas, mas também políticas públicas. Por meio da produção e da utilização de dados provenientes de diversas fontes, pode-se, por exemplo, realizar a análise de programas sociais, como forma de avaliar seu sucesso e aprimorar sua gestão, além de promover a transparência (VICTORINO et al., 2017). É possível, também, melhor delinear as demandas por infraestrutura ao se identificar e acompanhar as reais necessidades dos habitantes de determinada região, que tendem a se alterar ao longo do tempo (WEY; HUANG, 2018).

3 MATERIAIS E MÉTODOS

A metodologia Knowledge Discovery in Databases (KDD), desenvolvida por Fayyad, Shapiro-Piatetsky e Smyth (1996), é empregada neste trabalho. É frequentemente utilizada em análises de DM. Pode ser definida em cinco etapas sequenciais:

1. Seleção: escolha de variáveis com base no domínio dos dados e no conhecimento desejado;
2. Pré-processamento: tratamento de registros em branco, ruído e outliers;
3. Transformação: redução de dimensões ou transformação em variáveis mais representativas das características desejadas;
4. Data Mining: processamento de dados;
5. Interpretação e avaliação: análise de resultados e consolidação.

Nas três primeiras etapas, utilizou-se o

software Microsoft Excel. Nas demais, o software Orange. Os métodos de DM foram condicionados pela natureza da análise e por limitações do programa: o grande número de registros inviabilizou o cálculo de matrizes de distâncias e a aplicação de métodos de processamento mais intensivo, como a clusterização hierárquica. Foram aplicadas as seguintes técnicas:

• Clusterização por algoritmo K-Means: o agrupamento não-supervisionado de registros em clusters pode indicar tipologias de setores associadas a condições de saneamento semelhantes, dentro de um mesmo grupo. Consequentemente, trata-se de boa metodologia para a exploração inicial dos dados. Adotou-se número fixo de cinco clusters, tendo em vista que o PNSR define cinco agrupamentos de setores e a clusterização poderia indicar aderência a essa divisão (KANUNGO et al, 2002).

• Rede neural artificial (RNA): a RNA pode, a partir da base de dados, gerar modelo capaz de inferir a tipologia de um setor a partir de suas condições de saneamento, ou vice-versa. Diferenças no acesso podem ser evidenciadas caso padrões consistentes sejam encontrados (SCHMIDHUBER, 2014).

• *Random Forests*: ao contrário de árvores unitárias, que utilizam o ganho de informação, a metodologia é aleatória. Random Forests são compostas por conjuntos de árvores simples que, por votação, atribuem valores a determinada característica. Podem ser utilizadas para predição, assim como as RNAs (BREIMAN, 2001).

A base de dados analisada, proveniente da Amostra do Censo Demográfico de 2010 do IBGE (IBGE, 2011) e utilizada durante a concepção do PNSR, contém dados de 310.120 setores censitários, dispostos em linhas, para os quatro componentes do saneamento: abastecimento de água (AA), esgotamento sanitário (ES), manejo de resíduos sólidos (MRS) e manejo de águas pluviais (MAP). O resumo das variáveis, dispostas em colunas, está apresentado no Quadro 1.

Quadro 1: Composição da base de dados analisada

Coluna	Descrição
Código do Setor	Código do setor censitário, conforme o IBGE
Ruralidade do PNSR	Índice de rural definido no PNSR
Índice IBGE Rural	Índice de rural do IBGE
DPPs	Nº de domicílios particulares permanentes (DPPs) no setor censitário
Moradores	Nº de moradores no setor censitário
AA	Nº de domicílios com solução primária: 1) Rede de abastecimento; 2) Poço ou nascente; 3) Cisterna; 4) Outras.
ES	Nº de domicílios com solução primária: 1) Rede de coleta; 2) Fossa séptica; 3) Fossa rudimentar; 4) Vala; 5) Rio, lago ou mar; 6) Outra; 7) Sem banheiro.
MRS	Nº de domicílios com solução primária: 1) Coletado diretamente; 2) Coletado indiretamente; 3) Queimado na propriedade; 4) Enterrado na propriedade; 5) Jogado em terreno ou na rua; 6) Jogado em rio, lago ou mar; 7) Outra.
MAP	Nº de moradores que residem em logradouros: 1) Com pavimentação; 2) Sem pavimentação; 3) Com “bocas de lobo”; 4) Sem “bocas de lobo”.

Fonte: Autores (2022).

Conforme a metodologia KDD, iniciou-se pela seleção de dados. O código de cada setor censitário e o índice de rural do IBGE foram desconsiderados, uma vez que não são relevantes para esta análise. As colunas referentes ao MAP foram excluídas, tendo em vista que a base do IBGE não abrange todos os setores para este componente.

Em seguida, realizou-se o pré-processamento: 6.945 registros em branco foram excluídos, sendo a análise realizada para os 303.175 restantes. Na etapa de transformação, buscou-se obter variáveis mais representativas da presença de serviços de saneamento adequados. As soluções foram, então, classificadas em adequadas ou inadequadas, para cada componente do saneamento, conforme exposto no Quadro 2 (próxima página).

Quadro 2: Classificação de soluções de saneamento

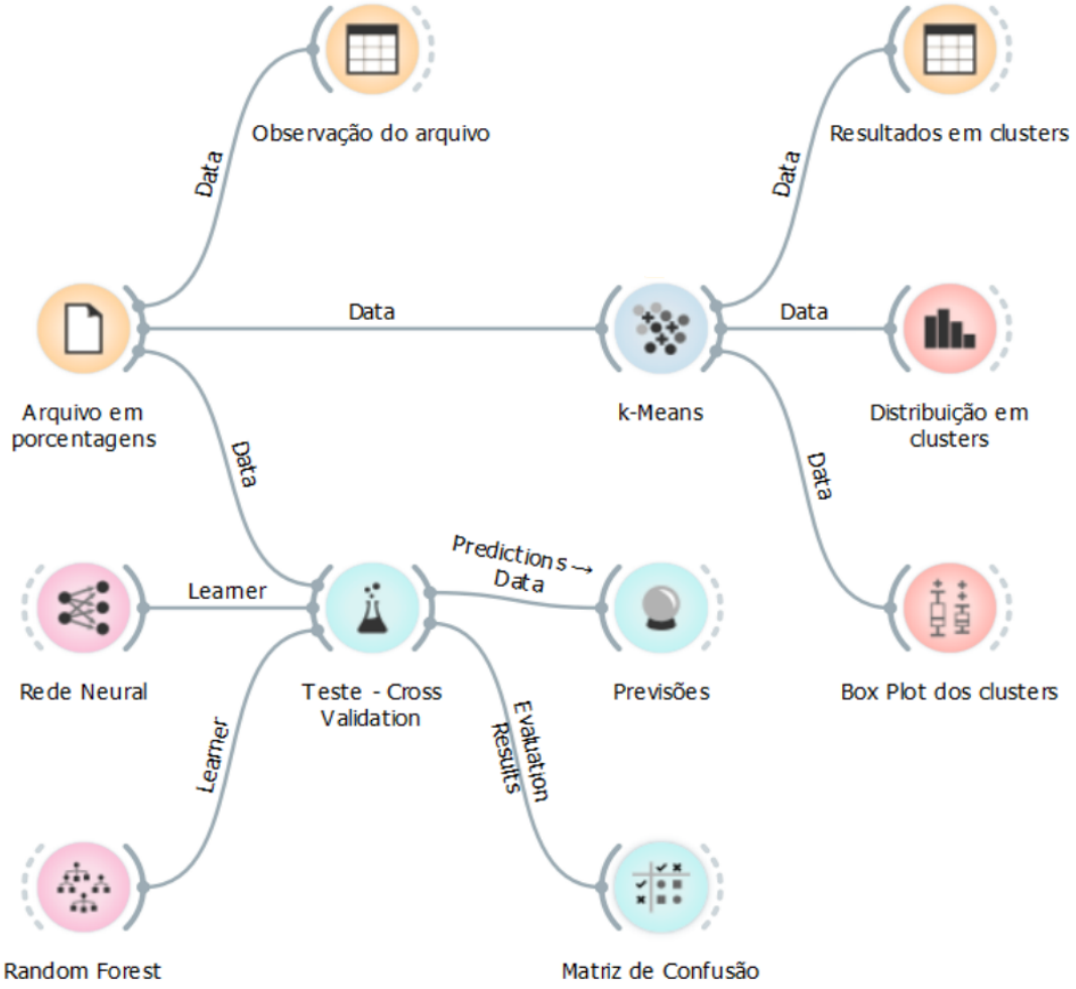
Componente	Categoria	Possibilidades
AA	Adequado	Rede, poço ou nascente
	Inadequado	Cisterna (unicamente), outra
ES	Adequado	Rede ou fossa séptica
	Inadequado	Fossa rudimentar, vala, rio, lago ou mar, outro, sem banheiro
MRS	Adequado	Coleta direta ou coleta indireta
	Inadequado	Queimado na propriedade, enterrado na propriedade, jogado em terreno ou rua, jogado em rio, lago ou mar ou outro

Fonte: Autores (2022).

Para cada setor censitário da amostra, os domicílios que utilizam soluções adequadas foram somados e o resultado foi dividido pelo número total de domicílios para determinar a porcentagem de adequação de cada componente. A base de dados ficou com quatro colunas: Ruralidades do PNSR e porcentagem de atendimento adequado para AA, ES e MRS.

Para a realização da etapa de DM, a variável categórica Ruralidades do PNSR foi selecionada como variável-alvo em todos os casos. As três variáveis restantes, numéricas, foram selecionadas como características dos setores censitários. A Figura 1 apresenta o diagrama representativo da metodologia, elaborado no software Orange.

Figura 1: Diagrama da metodologia no software Orange



Fonte: Autores (2022).

Os resultados foram comparados com os agrupamentos do PNSR, conforme apresentado no Quadro 3 (próxima página).

Quadro 3: Agrupamento de setores censitários, conforme o PNSR

Ambiente	Setores	Descrição
Urbano	1a	Regiões urbanas
Rural	1b, 2 e 4	Aglomerações próximas do urbano
	3	Aglomerações mais adensadas e isoladas
	5, 6 e 7	Aglomerações menos adensadas e isoladas
	8	Sem aglomerações, domicílios próximos de aglomerações ou isolados

Fonte: Autores (2022).

4. RESULTADOS E DISCUSSÕES

4.1. Clusterização: K-Means

O algoritmo K-Means foi executado com número de clusters K = 5 e limite de iterações igual a 300. Na Tabela 1 são apresentadas as porcentagens típicas de soluções adequadas nos setores que formaram cada cluster, para cada componente do saneamento. Os dados foram obtidos por Box Plot, considerando o 1º quartil, a mediana, o 3º quartil e a média.

Tabela 1: Resultados de Box Plot da clusterização

Comp.	Cluster	1º Quartil (%)	Mediana (%)	3º Quartil (%)	Média (%)
AA	C1	0,9966	1,0000	1,0000	0,9877 ± 0,0634
	C2	0,7273	0,8648	0,9572	0,8323 ± 0,1414
	C3	0,9419	0,9889	1,0000	0,9276 ± 0,1600
	C4	0,0222	0,1429	0,3269	0,1828 ± 0,1660
	C5	0,9360	0,9889	1,0000	0,9173 ± 0,1755
ES	C1	0,9667	0,9956	1,0000	0,9711 ± 0,0498
	C2	0	0,0169	0,1150	0,1236 ± 0,2283
	C3	0,0127	0,0464	0,1412	0,0876 ± 0,0953
	C4	0	0	0,0323	0,0706 ± 0,1733
	C5	0,4816	0,6069	0,7083	0,5948 ± 0,1357
MRS	C1	1,0000	1,0000	1,0000	0,9889 ± 0,0580
	C2	0	0,0164	0,1533	0,0960 ± 0,1396
	C3	0,8984	0,9835	1,0000	0,9227 ± 0,1195
	C4	0	0	0,0269	0,0769 ± 0,1890
	C5	0,9292	0,9903	1,0000	0,9280 ± 0,1342

Fonte: Autores (2022).

A porcentagem de setores censitários presentes em cada cluster, considerando as Ruralidades do PNSR, está apresentada na Tabela 2 (próxima página).

Tabela 2: Porcentagem de setores censitários em cada cluster

Ruralidades do PNSR	C1	C2	C3	C4	C5	Total
1a	71,0%	0,7%	17,2%	0,3%	10,9%	100,0%
1b	36,1%	7,3%	34,7%	2,9%	19,1%	100,0%
2	36,9%	9,6%	30,3%	2,6%	20,6%	100,0%
3	40,3%	4,0%	37,0%	3,2%	15,5%	100,0%
4	35,4%	8,0%	35,6%	3,6%	17,4%	100,0%
5	6,9%	35,1%	31,2%	19,1%	7,7%	100,0%
6	33,8%	13,5%	30,6%	13,5%	8,6%	100,0%
7	5,9%	38,8%	13,4%	38,6%	3,3%	100,0%
8	4,0%	52,0%	8,4%	30,6%	5,0%	100,0%

Fonte: Autores (2022).

Os resultados da Tabela 1 apontam a existência de realidades de saneamento muito distintas entre os cinco clusters, reforçando a hipótese 1. Há situações de quase universalização de soluções para os três componentes, bem como casos de desenvolvimento assimétrico entre eles, ou mesmo notável falta de infraestrutura em todos. É evidente, portanto, a existência de múltiplas ruralidades, que devem ser compreendidas com base na heterogeneidade de ocupações e nas múltiplas formas de se produzir a vida no campo, como destacado por Galizoni (2021).

A análise da Tabela 2 permite determinar a similaridade entre tipos de setores agrupados em uma mesma categoria no PNSR. Espera-se que a distribuição de ocorrências de tipos de setores de mesma categoria entre os clusters seja semelhante.

Os setores 1b, 2 e 4 pertencem ao grupo de aglomerações próximas do urbano, conforme apresentado no Quadro 3. Os dados da Tabela 2 indicam fortes semelhanças na distribuição de frequências entre todos os clusters para essas três categorias de setores. Evidencia-se, assim, que contemplam situações de saneamento e ruralidade similares. Há, portanto, forte aderência ao agrupamento realizado no PNSR, concordando com a hipótese 2.

Os setores 5, 6 e 7 pertencem ao grupo de aglomerações menos adensadas e isoladas. Entre os setores 5 e 7, há distribuição semelhante de casos nos clusters C1 e C2 e considerável variação nos demais clusters. Por outro lado, existe, entre os setores 5 e 6, distribuição semelhante nos clusters C3 e C5 e notável diferença nos demais. As frequências entre os setores 6 e 7 são muito distintas. Conclui-se, portanto, que há evidência de semelhanças nas situações de saneamento entre esses três tipos de setores, embora exista também heterogeneidade. Essa observação, contudo, não refuta a hipótese 2: conforme apresentado por Galizoni (2021), o rural e o urbano não são entes isolados e constituem um mosaico. Os setores 1b, 2 e 4 são fortemente influenciados por zonas urbanas próximas e apresentam menor variação de soluções de saneamento. Os setores 5, 6 e 7, por outro lado, apresentam menor adensamento e maior isolamento, o que condiciona maior variabilidade de soluções. A natureza desses setores é mais volátil, dificultando a modelagem.

Outra abordagem de análise é a comparação das características de cada cluster, na Tabela 1, com a distribuição de casos, na Tabela 2. Na Tabela 1, o cluster C1 representa locais próximos da universalização para os três componentes do saneamento. Observando-se a Tabela 2, é natural que locais urbanos (1a) apareçam em destaque nesse cluster, tendo em vista que a economia mais dinâmica e os maiores níveis de adensamento condicionam maior volume de investimentos no setor de saneamento. Essa consideração corrobora o estudo de Laschefski (2021), que correlaciona investimentos com as dinâmicas de mercado. Setores 1b, 2 e 4 apresentam proporção também considerável no cluster C1, em torno de 40%, tendo em vista que sua proximidade com áreas urbanas facilita a expansão dos serviços que já operam nas cidades. Seria ideal, contudo, que a proporção fosse maior, indicando maior expansão de serviços para esses locais. Os setores 3 e 6 correspondem ao ambiente rural isolado, mas o maior adensamento favorece a adoção de soluções de saneamento coletivas, o que explica a porcentagem significativa de casos nesse cluster.

O cluster C2 corresponde a locais com baixa presença de serviços de ES e MRS, mas considerável atendimento por AA, que tem sido historicamente priorizado em políticas públicas, o que explica a grande presença de locais rurais com menores aglomerações (setores 5, 6, 7 e 8) nesse cluster, uma vez que menor volume de investimentos tende a ser direcionado a esses locais e os demais componentes (ES e MRS) acabam ficando em segundo plano, algo agravado pela frequente associação desses componentes a soluções coletivas, mais onerosas.

O cluster C3 retrata locais com bom nível de atendimento por AA e MRS, mas pouco atendimento por ES. Os setores dos tipos 1a a 6 apresentam proporção considerável de casos nesse cluster, o que indica que o menor alcance do componente ES é um problema generalizado. Considerando que a implementação de sistemas de coleta e tratamento de ES tende a ser onerosa, este componente acaba sendo menos favorecido, o que poderia explicar a situação, mesmo em ambientes mais adensados, onde há forte presença de redes de água. A menor presença de setores 7 e 8 nesse cluster justifica-se pelo fato dos outros componentes do saneamento – AA e MRS – também apresentarem baixo alcance nesses locais, o que os direciona para os clusters C2 e C4.

O cluster C4 corresponde às piores situações, onde os três componentes apresentam adequação de atendimento muito baixa. Fica evidente a distância entre ambientes urbanizados (1a) ou rurais próximos a locais urbanos (setores 1b a 4), que quase não aparecem nesse cluster, e rurais mais isolados e menos aglomerados (setores 5 a 8), que são representados em grande proporção.

Por fim, o cluster C5 indica regiões com bom atendimento por AA e MRS, enquanto o ES apresenta indicadores de cobertura médios: corresponde a setores urbanizados ou rurais próximos do urbano que estão investindo em ES, mas ainda não em volume suficiente para se aproximar da universalização. É reforçada, portanto, a tendência de menor favorecimento do componente de ES.

4.2. Predição: RNA e Random Forests

O algoritmo de RNA foi configurado para utilizar 20 neurônios nas camadas intermediárias, com ativação ReLu, solução por método Adam e limite de 200 iterações, com treinamento replicável. Na sequência, o número de neurônios foi alterado para 40, buscando-se determinar se haveria variação na precisão. O algoritmo de Random Forests foi definido para 10 árvores, parando a divisão para subconjuntos com 5 itens ou menos. Posteriormente, realizou-se a classificação também para 20 árvores. A validação cruzada foi executada para os dois cenários, visando realizar comparações. Não houve alteração significativa ao modificar os parâmetros e os resultados de validação, nos dois métodos, foram semelhantes (Tabela 3).

Tabela 3: Resultados da validação cruzada para RNA e *Random Forests*

Modelo	AUC	CA	F1	Precisão	Recall
RNA - 20 neurônios	0,445	0,859	0,811	0,783	0,859
RNA - 40 neurônios	0,448	0,859	0,811	0,778	0,859
Random Forest - 10 árvores	-0,018	0,846	0,811	0,785	0,846
Random Forest - 20 árvores	0,017	0,849	0,812	0,786	0,849

Fonte: Autores (2022).

A matriz de confusão da RNA, para 40 neurônios, encontra-se na Tabela 4 (próxima página).

Tabela 4: Matriz de confusão para RNA com 40 neurônios

		Previsto									
		1a	1b	2	3	4	5	6	7	8	Soma
Atual	1a	206.618	0	0	0	0	174	0	0	4.145	210.937
	1b	10.970	0	0	0	0	62	0	0	2.183	13.215
	2	4.139	0	0	0	0	13	0	0	1.088	5.240
	3	2.601	0	0	0	0	25	0	0	306	2.932
	4	1.193	0	0	0	0	11	0	0	283	1.487
	5	3.642	0	0	0	0	284	0	0	5.169	9.095
	6	151	0	0	0	0	5	0	0	66	222
	7	238	0	0	0	0	15	0	0	960	1.213
	8	5.157	0	0	0	0	106	0	0	53.571	58.834
	Soma	234.709	0	0	0	0	695	0	0	67.771	303.175

Fonte: Autores (2022).

A matriz de confusão do *Random Forests*, considerando a configuração de 20 árvores, está apresentada na Tabela 5.

Tabela 5: Matriz de confusão para *Random Forests* com 20 árvores

		Previsto									
		1a	1b	2	3	4	5	6	7	8	Soma
Atual	1a	204.188	1.293	292	181	58	622	5	13	4.285	210.937
	1b	10.224	366	93	56	7	251	0	7	2.211	13.215
	2	3.768	183	74	23	4	75	1	5	1.107	5.240
	3	2.391	102	24	40	3	48	0	3	321	2.932
	4	1.164	20	6	6	0	31	0	1	259	1.487
	5	3.281	200	32	20	4	733	5	32	4.788	9.095
	6	126	8	0	5	0	14	0	0	69	222
	7	211	20	2	8	2	67	1	12	890	1.213
	8	5.019	467	140	39	8	1.224	1	71	51.865	58.834
	Soma	230.372	2659	663	378	86	3.065	13	144	65.795	303.175

Fonte: Autores (2022).

Para os dois algoritmos há elevado percentual de acertos nos setores 1a e 8, o que indica que as diferenças de adequação do saneamento entre o ambiente urbano e o rural sem adensamento são claramente amplas, a ponto de serem modeladas por predição com considerável precisão. O índice de acertos para os setores intermediários (1b a 7) foi baixo, indicando a existência de características mais variáveis e difíceis de modelar por métodos de predição. Esse resultado é esperado, considerando-se a distribuição de casos entre múltiplos clusters encontrada anteriormente para cada setor. Cabe ressaltar, que setores 1a e 8 existem em quantidade muito maior, facilitando o aprendizado de máquina e aumentando a precisão, enquanto os demais, menos comuns, estão associados a menores bases de treinamento.

5 CONCLUSÕES

Diferenças no alcance dos serviços de saneamento no Brasil são claras, demonstrando as múltiplas faces do ambiente rural, conforme apresentado no PNSR. A desigualdade é mais evidente ao comparar casos extremos, como aglomerações próximas de urbanizações e domicílios isolados. O maior adensamento, que facilita a adoção de soluções coletivas, não justifica tamanha discrepância, tendo em vista que soluções individuais poderiam ser adotadas em locais remotos. Infere-se,

portanto, que o problema está provavelmente associado à desigualdade no direcionamento de investimentos, atrelada ao desenvolvimento econômico de cada região, conforme apontado por Laschefski (2021).

Comprova-se que a análise de dados envolvendo Big Data e mineração de dados pode ser utilizada como mecanismo balizador e validador de políticas públicas para o saneamento, identificando múltiplas ruralidades e auxiliando na definição de soluções para cada local, sendo destacadas as áreas que dependem da priorização de investimentos. Podem ser gerados indicadores de desenvolvimento rural, importantes para políticas públicas, como ressaltado por Kageyama (2006).

Por um lado, a clusterização – não supervisionada –, ao buscar padrões sem critérios pré-definidos, possibilitou identificar diferentes ruralidades que podem ser agrupadas em categorias, conforme proposto no PNSR. Por outro, a imprecisão dos métodos supervisionados, baseados em treinamento prévio, demonstrou que o equacionamento preditivo seria difícil, tendo em vista a variabilidade de soluções de saneamento e a menor quantidade de amostras para setores intermediários (códigos 1b a 7). Essa constatação apenas reforça a complexidade da temática. Outros métodos podem ser aplicados para revelar informações, e variáveis adicionais, como a renda média familiar, podem ser incluídas em futuras pesquisas, fornecendo retrato mais abrangente das diferenças socioeconômicas do País.

O PNSR foi criado para modificar o quadro de priorização de investimentos em áreas urbanas e fortalecer as bases de apoio ao atendimento às áreas rurais do País, revertendo a dispersão de competências entre os diferentes Ministérios. Como é papel do Governo Federal coordenar o PNSR, é importante que haja contraposição ao atual ambiente regulatório, para que sejam assegurados recursos humanos e econômicos à materialização de medidas estruturais e estruturantes. Nesse sentido, é necessária a real interpretação das ruralidades para que seja possível alcançar o desenvolvimento tecnológico sustentável, por meio de instrumentos perenes de gestão, educação e participação social.

Agradecimentos

Os autores agradecem à Fundação Nacional de Saúde – Funasa, à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES, e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq, pelo apoio financeiro.

REFERÊNCIAS

ABRAMOVAY, R. **Do setor ao território: funções e medidas da ruralidade no desenvolvimento contemporâneo**. Rio de Janeiro: Instituto de Pesquisa Econômica Aplicada (IPEA), 2000.

BRASIL. Ministério da Saúde. **Portaria nº 2.866, de 2 de dezembro de 2011**. Brasília, 2011. Recuperado em 6 de setembro, 2022, de https://bvsms.saude.gov.br/bvs/saudelegis/gm/2011/prt2866_02_12_2011.html.

BRASIL. Ministério da Saúde. Fundação Nacional de Saúde. **Programa Nacional de Saneamento Rural – PNSR**. Brasília: Fundação Nacional de Saúde (Funasa), 2019. Disponível em: http://www.funasa.gov.br/documents/20182/38564/MNL_PNSR_2019.pdf/08d94216-fb09-468e-ac98-afb4ed0483eb. Acesso: 9 de setembro, 2022.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, p. 5-32, 2001. DOI: https://doi.org/10.1023/A:1010933404324.

CHAMIKARA, M. A. P. et al. Efficient privacy preservation of big data for accurate data mining. **Information Sciences**, v. 527, p. 420-443, 2020. DOI: https://doi.org/10.1016/j.ins.2019.05.053.

DUCANGE, P.; FAZZOLARI, M.; MARCELLONI, F. An overview of recent distributed algorithms for learning fuzzy models in Big Data classification. **Journal of Big Data**, v. 7, n. 19, 29 p, 2020. DOI: https://doi.org/10.1186/s40537-020-00298-6.

FAYYAD, U.; SHAPIRO-PIATETSKY, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, v. 17, n. 3, p. 37-54, 1996. DOI: https://doi.org/10.1609/aimag.v17i3.1230.

FREITAS, E. S. M. Reflexões sobre o conceito de rural e ruralidade para o contexto do Programa Nacional de Saneamento Rural. *In*: REZENDE, S. (Org.). **Aspectos conceituais da ruralidade no Brasil e interfaces com o saneamento básico**. Série Subsídios ao PNSR, vol. 1. Brasília: Fundação Nacional de Saúde (Funasa), p. 101-125, 2021. Disponível em: https://repositorio.funasa.gov.br/handle/123456789/670. Acesso: 6 de setembro, 2022.

GALIZONI, F. M. Rural e Ruralidades: Reflexões para o Programa Nacional de Saneamento Rural. *In*: REZENDE, S. (Org.), **Aspectos conceituais da ruralidade no Brasil e interfaces com o saneamento básico**. Série Subsídios ao PNSR, vol. 1. Brasília: Fundação Nacional de Saúde (Funasa), p. 9-22, 2021. Disponível em: https://repositorio.funasa.gov.br/handle/123456789/670. Acesso: 6 de setembro, 2022.

GRAZIANO DA SILVA, J. **A modernização dolorosa**. Rio de Janeiro: Zahar Editores, 1981.

HARVEY, D. **Condição pós-moderna**. São Paulo: Loyola, 1989.

HOLANDA, S. B. **Raízes do Brasil**. Rio de Janeiro: José Olympio, 1999 (trabalho original publicado em 1936).

IBGE. **Censo Demográfico 2010**. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística, 2011. Disponível em: https://www.ibge.gov.br/estatisticas/sociais/saude/9662-censo-demografico-2010.html?=&t=publicacoes. Acesso: 6 de setembro, 2022.

KAGERMANN, H. et al. (Eds.). **Industrie 4.0 in a Global Context: Strategies for Cooperating with International Partners (acatech STUDY)**. Munique: Herbert Utz Verlag, 2016. Disponível em: https://en.acatech.de/publication/industrie-4-0-in-a-global-context-strategies-for-cooperating-with-

international-partners/. Acesso: 14 de setembro, 2022.

KAGEYAMA, A. **Desenvolvimento rural no Rio Grande do Sul**. *In*: SCHNEIDER, S. (Org.), *A diversidade da agricultura familiar*. Porto Alegre: UFRGS Editora, 2006.

LANEY, D. *3D* **Data Management: Controlling Data Volume, Velocity and Variety**. Stamford: META Group Res Note, 2001.

LEFBVRE, H. **A revolução urbana**. Belo Horizonte: UFMG, 1999.

LENZ, J.; WUEST, T.; WESTKÄMPER, E. Holistic approach to machine tool data analytics. **Journal of Manufacturing Systems**, v. 48, p. 180-191, 2018. DOI: https://doi.org/10.1016/j.jmsy.2018.03.003.

LIEBER, D. et al. Quality Prediction in Interlinked Manufacturing Processes based on Supervised & Unsupervised Machine Learning. **Procedia CIRP**, v. 7, p. 193-198, 2013. DOI: https://doi.org/10.1016/j.procir.2013.05.033.

MIRANDA, C.; SILVA, H. (Orgs.). Concepções da ruralidade contemporânea: as singularidades brasileiras. **Série Desenvolvimento Rural Sustentável**, vol. 21. Brasília: Instituto Interamericano de Cooperação para a Agricultura (IICA), 2013. Disponível em: http://repiica.iica.int/DOCS/B3226P/B3226P.PDF. Acesso: 6 de setembro, 2022.

RIGOTTI, J. I. R.; HADAD, R. PNSR – Delimitação das áreas rurais brasileiras. *In*: REZENDE, S. (Org.), **Aspectos conceituais da ruralidade no Brasil e interfaces com o saneamento básico**. Série Subsídios ao PNSR, vol. 1. Brasília: Fundação Nacional de Saúde (Funasa), p. 77-100, 2021. Disponível em: https://repositorio.funasa.gov.br/handle/123456789/670. Acesso: 6 de setembro, 2022.

RUBINGER, S. D. **Desvendando o conceito de saneamento no Brasil: uma análise da percepção da população e do discurso técnico contemporâneo**. Dissertação (Mestrado em Saneamento, Meio Ambiente e Recursos Hídricos). Escola de Engenharia, Universidade Federal de Minas Gerais. Belo Horizonte, 2008.

SCHMIDHUBER, J. Deep learning in neural networks: An overview. **Neural Networks**, v. 61, p. 85-117, 2015. DOI: https://doi.org/ 10.1109/TPAMI.2002.1017616.

SCHWARCZ, L. M. Sérgio Buarque de Holanda e essa tal de "Cordialidade". *Ide (São Paulo)*, v. 31, n. 146, p. 83-89, 2008.

VALADARES, A. A. **O Gigante invisível: território e população rural para além das convenções oficiais**. Brasília: Instituto de Pesquisa Econômica Aplicada (IPEA), 2014. Disponível em: http://repositorio.ipea.gov.br/bitstream/11058/2866/1/TD_1942.pdf. Acesso: 6 de setembro, 2022.

VICTORINO, M. de C. et al. Uma proposta de ecossistema de Big Data para a análise de dados abertos governamentais conectados. **Informação e Sociedade**, v. 27, n. 1, p. 225-242, 2017. Disponível em: https://periodicos.ufpb.br/ojs/index.php/ies/article/view/29299. Acesso: 6 de setembro, 2022.

WANDERLEY, M. N. B.; FAVARETO, A. A singularidade do rural brasileiro: as implicações para as tipologias territoriais e a elaboração de políticas públicas. *In*: MIRANDA, C. S.; SILVA, H. (Orgs.). **Concepções da ruralidade contemporânea: as singularidades brasileiras**. Série Desenvolvimento Rural Sustentável, vol. 21. Brasília: Instituto Interamericano de Cooperação para a Agricultura (IICA), 2013. Disponível em: http://repiica.iica.int/DOCS/B3226P/B3226P.PDF. Acesso: 6 de setembro, 2022.

WEY, W.-M.; HUANG, J.-Y. Urban sustainable transportation planning strategies for livable

city's quality of life. **Habitat International**, v. 82, p. 9-27, 2018. DOI: https://doi.org/10.1016/j.habitatint.2018.10.002.

WITTEN, I. H. et al. **Data Mining: Practical Machine Learning Tools and Techniques** (4ª ed.). Morgan Kaufmann, 2017.