IDENTIDADES TERRITORIAIS
MATO GROSSO

Nota Técnica

Análise de clusters territoriais dos municípios de Mato Grosso

Identidades Territoriais
Outubro de 2026

SEBRAE

1 Apresentação e objetivo

Esta nota técnica documenta a construção dos clusters territoriais dos 142 municípios de Mato Grosso. Complementa o relatório de inteligência dos perfis municipais do estudo Identidades Territoriais e apresenta as fontes, os tratamentos dos indicadores, a configuração da modelagem e os critérios de avaliação e interpretação dos resultados.

O estudo identifica grupos de municípios com características semelhantes a partir de informações econômicas, sociais, demográficas, educacionais e de infraestrutura. Os resultados fornecem insumos à análise territorial, à formulação de políticas públicas e ao planejamento empresarial. A unidade de análise é o município; a caracterização das empresas e das pessoas exige informações específicas sobre esses públicos.

A solução final adotada foi o método K-means com quatro clusters, aplicado aos dez componentes retidos pela Análise de Componentes Principais (PCA). A PCA sintetiza as informações dos indicadores selecionados em um conjunto menor de dimensões, utilizado como base para a classificação dos municípios. A definição dessa solução foi precedida pela exploração dos dados e pela comparação de alternativas de modelagem, conforme os procedimentos apresentados nesta nota.

O documento registra as decisões aplicadas no processamento, distinguindo as recomendações automáticas dos tratamentos executados. Apresenta os diagnósticos e as limitações da solução final para permitir a compreensão dos resultados e de suas condições de uso.

Configuração da solução documentada

A matriz reúne 142 municípios e 29 indicadores selecionados entre 34 medidas preparadas. O K-means foi ajustado com quatro grupos sobre os dez componentes principais, que representam aproximadamente 82,43% da variância da matriz de entrada. Os agrupamentos descrevem perfis territoriais e não constituem um ranking de desempenho.

2 Base de dados e indicadores

2.1 Abrangência territorial e unidade de análise

A base consolidada abrange os 142 municípios de Mato Grosso. Cada registro reúne os indicadores utilizados na caracterização de um município.

2.2 Fontes e períodos de referência

As fontes públicas e institucionais foram organizadas em oito temas: perfil territorial, pujança econômica, vulnerabilidade social, vocação produtiva, conectividade, inovação e capital, educação e infraestrutura.

Foram preparados 34 indicadores. Esse número corresponde às medidas organizadas para análise; uma mesma fonte pode fornecer mais de um indicador.

A base combina referências de 2010 a 2026, conforme a disponibilidade das fontes, e indicadores de crescimento populacional de 2021–2024 e 2024–2025. A classificação reúne períodos distintos, condição que deve orientar sua interpretação.

O Anexo A reúne os 34 indicadores preparados, com temas, fontes e períodos de referência, antes dos tratamentos e das exclusões.

2.3 Organização da base e composição dos indicadores

A consolidação utilizou o código municipal para reunir as fontes e preservar a identificação dos territórios nos resultados.

A base inclui percentuais, taxas, valores por habitante, contagens e índices, definidos no Anexo A. A seção 3 descreve os tratamentos e as mudanças de representação anteriores à modelagem.

Após a preparação e a seleção, cinco indicadores foram excluídos. Os 29 restantes compuseram a matriz utilizada na PCA, conforme a distribuição por tema a seguir.

Quadro 1 Composição dos indicadores por tema

Tema	Preparados	Excluídos	Selecionados
Perfil territorial	5	0	5
Pujança econômica	8	0	8
Vulnerabilidade social	4	1	3
Vocação produtiva	7	2	5
Conectividade	1	1	0
Inovação e capital	1	1	0
Educação	4	0	4
Infraestrutura	4	0	4
Total	34	5	29

Fonte: registros da etapa de seleção de variáveis do projeto. Os critérios e os motivos das exclusões são apresentados no tópico seguinte.

3 Preparação dos dados

A preparação compreendeu diagnóstico, tratamento de ausentes e extremos, transformação das distribuições e seleção das variáveis. As decisões executadas foram registradas para acompanhar as alterações anteriores à modelagem.

O Anexo B registra, para cada indicador, as ausências originais, seu preenchimento, a transformação e a representação final. Inclui as variáveis posteriormente excluídas e documenta as exceções de tratamento descritas nesta seção.

3.1 Diagnóstico e tratamento de valores ausentes

A análise exploratória examinou disponibilidade, frequência de zeros, distribuição dos valores e observações extremas. As sugestões geradas foram aplicadas conforme as decisões registradas no projeto.

Na imputação por mediana, os ausentes foram substituídos pelo valor central das observações disponíveis do próprio indicador. O procedimento completa a informação necessária ao processamento, mas não recupera o valor efetivamente observado no município.

Em empresas de base tecnológica e atendimento por esgoto, os ausentes foram substituídos por zero antes da conversão em indicadores de presença. Nesses casos, o zero final pode corresponder a um registro original igual a zero ou a uma ausência de informação convertida no tratamento.

O Índice de Ecossistema de Inovação permaneceu com 105 registros ausentes, equivalentes a aproximadamente 73,94% dos municípios, e foi excluído na etapa de seleção.

3.2 Tratamento de valores extremos e transformação das distribuições

A rotina aplicou winsorização entre os percentis 1 e 99 aos 34 indicadores preparados. Valores abaixo do primeiro percentil ou acima do percentil 99 foram substituídos pelos respectivos limites. O procedimento preserva as observações municipais e reduz a influência dos extremos.

As transformações buscaram reduzir a assimetria: Box-Cox foi aplicada a valores estritamente positivos; Log1p calcula o logaritmo de um mais o valor observado e permite zeros; Yeo-Johnson admite positivos, zeros e negativos. Nos demais casos, a distribuição foi mantida sem transformação.

No fluxo geral, a imputação antecedeu a winsorização, seguida pela transformação. Nas duas variáveis com substituição de ausentes por zero, foi preservada a regra específica do código. O Anexo B registra as decisões por indicador.

3.3 Conversão em indicadores de presença

Quatro indicadores foram convertidos em variáveis binárias: matrículas em formação inicial e continuada (FIC), matrículas em educação profissional técnica, empresas de base tecnológica e atendimento por esgoto.

A conversão atribuiu 1 aos valores positivos após o tratamento de ausentes e 0 aos demais. O código preserva essa condição antes das transformações. Essas variáveis distinguem a presença registrada, sem representar quantidade ou intensidade.

Em empresas de base tecnológica e esgoto, essa leitura também depende da substituição de ausentes por zero, descrita na seção 3.1.

3.4 Seleção dos indicadores

A seleção sobre a base tratada previu exclusão por mais de 70% de ausentes; indicador constante ou com variância menor ou igual a 10^{-12}; ou redundância, definida por correlação absoluta de Pearson maior ou igual a 0,80 com outro indicador já selecionado.

Para reduzir redundâncias, a rotina priorizou menor quantidade de ausentes e maior quantidade de valores distintos, com desempate pela ordem original das variáveis. Foram excluídos cinco indicadores.

Quadro 2 Indicadores excluídos e motivos registrados

Indicador excluído	Motivo	Indicador mantido relacionado	Correlação absoluta
Beneficiários do Bolsa Família	Redundância	População em situação de pobreza	0,837
Participação da administração pública no PIB	Redundância	PIB per capita	0,983
Produção agrícola per capita	Redundância	Participação da agropecuária no PIB	0,836
Cobertura 4G/5G	Redundância	Taxa de urbanização	0,910
Índice de Ecossistema de Inovação	73,94% de ausentes	Não se aplica	Não se aplica

Fonte: registros da seleção de variáveis. As correlações referem-se aos indicadores tratados. As exclusões reduzem redundâncias na matriz e não tornam essas medidas irrelevantes para a compreensão dos municípios.

3.5 Padronização e composição da matriz final

Após a seleção, os 25 indicadores contínuos foram padronizados com StandardScaler, subtraindo a média e dividindo pelo desvio padrão populacional de cada variável. O procedimento reduz diferenças de escala entre medidas originalmente expressas em unidades distintas.

Os quatro indicadores de presença permaneceram na escala de 0 e 1, sem aplicação do StandardScaler. Os nomes e códigos dos municípios foram preservados para identificação, sem integrar as variáveis da modelagem.

A matriz final reuniu 142 municípios e 29 indicadores, constituindo a entrada para a Análise de Componentes Principais.

4 Redução de dimensões e clusterização

4.1 Análise de Componentes Principais

A PCA foi ajustada diretamente à matriz selecionada, com os indicadores contínuos já padronizados e os quatro indicadores de presença preservados em 0/1. Não foi aplicada uma segunda padronização antes desse ajuste. A rotina utilizou decomposição completa (svd_solver="full") e centralização das variáveis.

Os componentes principais são combinações lineares dos indicadores e são ordenados pela variância que representam. A solução final utilizou os escores de PC1 a PC10, que concentram aproximadamente 82,43% da variância total da matriz de entrada. Esse percentual descreve a informação representada nessa matriz e não constitui uma medida de acerto da classificação.

Quadro 3 Variância representada pelos componentes principais

Componente	Variância explicada	Variância acumulada
PC1	33,84%	33,84%
PC2	14,76%	48,60%
PC3	6,96%	55,57%
PC4	5,33%	60,90%
PC5	4,96%	65,86%
PC6	4,08%	69,94%
PC7	3,47%	73,41%
PC8	3,32%	76,73%
PC9	2,96%	79,70%
PC10	2,74%	82,43%

Fonte: cálculo sobre a mesma matriz selecionada utilizada pela PCA da entrega final, sem segunda padronização. As diferenças residuais decorrem do arredondamento.

4.2 Alternativas exploradas e solução adotada

O processamento contemplou a comparação de K-means, K-medoids e misturas gaussianas, com e sem redução por PCA. O HDBSCAN também foi explorado como diagnóstico de densidade, considerando os municípios classificados como ruído. As rotinas de comparação produziram métricas para apoiar a avaliação das alternativas.

A solução aprovada pelo projeto foi K-means com quatro grupos sobre dez componentes principais. Esta nota documenta essa configuração. As métricas ajudam a avaliar a organização dos agrupamentos, mas não estabelecem, isoladamente, uma classificação universalmente superior ou um número natural e definitivo de perfis municipais.

4.3 Especificação do K-means

O K-means atribui os municípios a grupos procurando reduzir a soma das distâncias quadráticas entre os escores de cada município e o centro de seu grupo. O ajuste considera simultaneamente os dez componentes selecionados. A configuração da entrega é apresentada a seguir.

Quadro 4 Parâmetros do K-means da entrega final

Parâmetro	Valor
Número de clusters	4
Dimensões de entrada	PC1 a PC10
Inicialização	k-means++
Número de inicializações	50
Semente aleatória	7
Algoritmo	lloyd
Máximo de iterações	300
Tolerância	0,0001

Os rótulos produzidos pelo algoritmo foram acrescidos de uma unidade para apresentação como Clusters 1 a 4. As denominações foram definidas pela equipe a partir da interpretação dos perfis e não representam uma ordem de desempenho.

Quadro 5 Denominações e tamanho dos clusters

Cluster	Denominação	Municípios
1	Territórios Rurais de Vulnerabilidade Socioeconômica	62
2	Polos Econômicos, Educacionais e Urbanos	14
3	Territórios de Base Agroeconômica	41
4	Territórios Urbanos, Serviços e Industriais	25

Fonte: configuração e classificação da entrega final. Os quatro grupos totalizam 142 municípios.

5 Avaliação da solução final

5.1 Separação e compactação dos grupos

A solução foi examinada com métricas calculadas no espaço dos dez componentes utilizados no K-means. Cada medida descreve um aspecto da configuração e deve ser interpretada em conjunto com os perfis e os diagnósticos municipais.

Quadro 6 Métricas da solução final

Métrica	Resultado	Interpretação
Silhouette médio	0,188619	Compara a proximidade ao próprio grupo com a proximidade a outros grupos. Valores maiores indicam maior separação.
Calinski-Harabasz	37,536947	Relaciona dispersão entre grupos e dispersão interna. Valores maiores favorecem a comparação entre configurações compatíveis.
Davies-Bouldin	1,500907	Avalia a relação entre dispersão interna e distância entre grupos. Valores menores indicam melhor separação relativa.
Inércia	1644,043397	Soma das distâncias quadráticas aos centros. Depende da escala, das dimensões e do número de grupos.

Fonte: métricas da entrega final.

O silhouette médio de aproximadamente 0,19 indica separação limitada entre os perfis. Há sobreposição de características, de modo que os grupos devem ser lidos como sínteses de uma realidade territorial contínua. A inércia não deve ser usada isoladamente para escolher o número de clusters, pois tende a diminuir com a inclusão de grupos.

5.2 Municípios de fronteira

O diagnóstico municipal utiliza duas medidas: o silhouette individual e a margem relativa entre o centro do próprio grupo e o centro alternativo mais próximo. A margem é calculada por (distância ao centro alternativo − distância ao próprio centro) / distância ao centro alternativo, considerando distâncias euclidianas nos dez componentes.

Um município foi sinalizado como fronteira quando apresentou silhouette individual menor ou igual a 0,10 ou margem relativa menor ou igual a 0,10. O termo identifica maior proximidade entre perfis na classificação e não corresponde a uma fronteira geográfica.

Quadro 7 Municípios sinalizados como fronteira

Cluster	Municípios de fronteira
1	22
2	0
3	16
4	4
Total	42

Fonte: diagnósticos municipais da solução final.

A sinalização não altera a alocação original e não representa probabilidade de pertencimento. Ela orienta uma leitura mais cuidadosa dos casos com características compartilhadas entre grupos.

5.3 Estabilidade da classificação

Foram realizados dois experimentos mantendo os escores da PCA fixos. No primeiro, o K-means foi ajustado com vinte sementes, de 0 a 19, preservando cinquenta inicializações por ajuste. No segundo, foram realizados cinquenta ajustes em subamostras sem reposição de 114 municípios, correspondentes a aproximadamente 80% da base, com sorteio controlado pela semente 2026. Cada modelo ajustado na subamostra foi utilizado para classificar todos os 142 municípios.

Os resultados foram comparados à classificação de referência por meio do Índice de Rand Ajustado (ARI) e da concordância após alinhamento dos rótulos por máxima sobreposição. O ARI mede a semelhança entre partições com ajuste para concordâncias ao acaso; valor 1 representa igualdade entre as classificações. A concordância corresponde à proporção de municípios que reproduzem o grupo de referência após esse alinhamento.

Quadro 8 Estabilidade em relação à classificação de referência

Experimento	Repetições	ARI médio	Concordância média
Variação das sementes	20	0,9645	98,27%
Subamostras de aproximadamente 80%	50	0,7875	88,63%

Fonte: resumo de estabilidade da entrega final.

A classificação mostrou maior estabilidade à mudança de semente do que à alteração da amostra. Os experimentos avaliam o K-means condicionado à PCA fixa; não medem a estabilidade de todo o processamento, desde as fontes até os componentes. Concordância não é probabilidade de pertencimento ou garantia de permanência em atualizações futuras.

6 Critérios de caracterização e interpretação

6.1 Construção dos perfis e denominações

A caracterização relaciona os agrupamentos aos indicadores que os compõem. As denominações aprovadas sintetizam combinações recorrentes no grupo e não implicam que todos os municípios apresentem a mesma intensidade de cada característica.

Os componentes podem ser interpretados pelas contribuições dos indicadores e por suas relações com cada dimensão. Essa leitura deve utilizar os componentes da PCA da entrega final, pois uma PCA obtida com outra padronização pode produzir escores e cargas diferentes. O sinal de um componente não indica qualidade ou desempenho e pode ser invertido sem alterar sua informação geométrica.

6.2 Indicadores característicos e unidades de apresentação

O gráfico de indicadores característicos utiliza as médias da base selecionada para a modelagem. Os 25 indicadores contínuos estão padronizados, enquanto os quatro indicadores de presença permanecem em 0/1. Nos contínuos, as médias expressam desvios da média geral; nos binários, correspondem à proporção de registros positivos do grupo. A leitura dos binários permanece própria dessa escala, mesmo quando a saída denomina a coluna ou o eixo como média padronizada.

O gráfico destaca médias com valor absoluto maior ou igual a 0,50. Nos contínuos, o corte corresponde a meio desvio padrão em relação à média geral. Nos binários, seleciona proporções iguais ou superiores a 50%. Trata-se de um critério descritivo de destaque, sem significado de teste de hipótese ou de importância causal.

As fichas complementares utilizam outro cálculo: padronizam novamente os indicadores, inclusive os binários, para descrever os perfis. Esse processamento descritivo não altera a PCA nem o K-means. Os valores dessas fichas e os do gráfico de indicadores característicos não devem ser tratados como equivalentes.

As barras descrevem indicadores após os tratamentos. Para indicadores transformados, uma diferença em desvios da média não equivale diretamente a uma diferença em reais, habitantes ou pontos percentuais. As tabelas de indicadores em unidades originais utilizam a base consolidada anterior ao tratamento e apresentam mediana, quartis, extremos e ausências. As duas apresentações respondem a perguntas distintas e não devem ser confundidas.

Nos indicadores de presença, a média da variável 0/1 corresponde à proporção de registros positivos do grupo. Para esgoto e empresas de base tecnológica, essa leitura depende da decisão de converter ausentes em zero. As escalas dos painéis do gráfico podem variar, exigindo atenção aos valores dos eixos ao comparar barras entre clusters.

6.3 Leitura dos valores padronizados e de seus sinais

Nos quadros quantitativos do relatório de inteligência, foi utilizado o campo Valor Padronizado da base descritiva recebida, sem nova padronização na elaboração dos quadros. A média de cada cluster é a média aritmética dos escores de seus municípios, com peso igual por município. O desvio padrão populacional descreve a dispersão dos escores dentro do grupo.

Nessa base descritiva, zero representa a média de referência. Valores positivos indicam posição acima dessa média e valores negativos, abaixo. A interpretação depende da natureza do indicador. O sinal é mantido como recebido e não é convertido em uma nota geral de desempenho.

Em pobreza, participação das classes D e E e índice de Gini, valores mais altos expressam maior presença relativa dessas condições. Estar abaixo da média em pobreza ou desigualdade pode representar condição mais favorável nessa dimensão. Em emprego, remuneração ou acesso a serviços, valores maiores podem ter outra interpretação, conforme a definição e os tratamentos do indicador.

Participações dos setores no PIB, urbanização, rebanho, área plantada e dependência etária caracterizam estruturas produtivas ou demográficas. Seus sinais não estabelecem, por si só, uma condição desejável ou indesejável. Os escores dos indicadores binários padronizados para descrição também não equivalem a quantidades originais ou percentuais de cobertura.

6.4 Representantes e visualizações territoriais

Os municípios representativos foram identificados pela menor distância euclidiana ao centro do próprio cluster nos dez componentes. A entrega registra cinco representantes por grupo. Essa proximidade descreve a posição no espaço da modelagem e não estabelece prioridade de atendimento ou liderança territorial.

O mapa apresenta a distribuição geográfica das alocações. A proximidade espacial não foi utilizada como restrição para formar os grupos: municípios do mesmo cluster podem estar distantes entre si. O gráfico de dispersão em PC1 e PC2 mostra uma projeção em duas dimensões de uma classificação que utiliza dez, de modo que a sobreposição visual nessa figura não resume toda a separação considerada pelo algoritmo.

7 Limitações e condições de uso

A classificação depende das fontes, dos períodos de referência, dos tratamentos e dos indicadores selecionados. Como os dados combinam diferentes anos, os agrupamentos não representam uma fotografia de todas as dimensões em um único período. A atualização das bases pode alterar perfis e alocações.

A imputação por mediana reduz as lacunas necessárias à modelagem, mas introduz valores estimados. Nos indicadores com maior quantidade de imputações, essa decisão deve receber atenção adicional. A conversão de ausentes em zero para esgoto e empresas de base tecnológica impede distinguir esses casos de registros originalmente iguais a zero na variável final. A representação binária, por sua vez, deixa de expressar a intensidade da característica.

A winsorização limita a influência de valores extremos, e as transformações alteram a escala dos indicadores. As exclusões por correlação reduzem redundâncias, mas também delimitam as dimensões representadas. A PCA retém aproximadamente 82,43% da variância da matriz e, portanto, não preserva toda a informação disponível. Como os indicadores binários não receberam a padronização aplicada aos contínuos antes da PCA, sua escala também participa da configuração final.

O K-means atribui cada município a um único grupo, embora existam características compartilhadas e municípios de fronteira. O silhouette médio recomenda cautela na leitura de separações rígidas. Os testes de estabilidade realizados não abrangem novas coletas, outras decisões de tratamento ou o reajuste da PCA em cada repetição.

A unidade de análise é o município. Os resultados não permitem atribuir as características médias territoriais a cada empresa ou pessoa, nem demonstram relações de causa e efeito. As denominações dos clusters não constituem um ranking de desenvolvimento, potencial empresarial ou prioridade de atuação.

O uso dos resultados deve combinar a classificação com o conhecimento das equipes, as particularidades municipais e outras informações pertinentes. Para reproduzir a entrega, devem ser preservadas as versões das bases, as decisões de tratamento, a ordem das variáveis, a PCA sem segunda padronização e os parâmetros do K-means. A documentação dos arquivos permite acompanhar essas escolhas e orientar futuras atualizações.

A interpretação dos quadros deve identificar qual representação está em uso: unidades originais, matriz da modelagem ou base descritiva padronizada. Essa identificação evita confundir proporções de presença, escores padronizados e medidas observadas nos municípios.

Anexo A Fontes e períodos dos indicadores

Relação dos 34 indicadores inicialmente preparados. Os cinco indicadores excluídos permanecem neste anexo para documentar o levantamento original.

Quadro 9 Indicadores preparados e respectivas fontes

Tema	Indicador	Fonte	Referência
Perfil territorial	Densidade demográfica	IBGE	2025
Perfil territorial	Taxa de urbanização	IPCMaps	2024
Perfil territorial	Grau de dependência	IPCMaps	2024
Perfil territorial	Crescimento populacional	IBGE	2021–2024
Perfil territorial	Crescimento populacional	IBGE	2024–2025
Pujança econômica	PIB per capita	IBGE	2023
Pujança econômica	Densidade empresarial	RFB	2025
Pujança econômica	Saldo de empregos formais	CAGED	2025
Pujança econômica	Massa salarial per capita	RAIS	2025
Pujança econômica	Taxa de sobrevivência empresarial em três anos	RFB	2025
Pujança econômica	Exportação FOB em relação ao PIB	MDIC/SECEX	2023
Pujança econômica	Crédito per capita	Banco Central	2025
Pujança econômica	Fluxo de comércio per capita	MDIC/SECEX	2023
Vulnerabilidade social	Percentual da população em situação de pobreza	CadÚnico	2025
Vulnerabilidade social	Participação das classes D e E	IPCMaps	2024
Vulnerabilidade social	Percentual de beneficiários do Bolsa Família	MDS	2025
Vulnerabilidade social	Índice de Gini	IBGE	2010
Vocação produtiva	Participação da agropecuária no PIB	IBGE	2023
Vocação produtiva	Participação da indústria no PIB	IBGE	2023

Tema	Indicador	Fonte	Referência
Vocação produtiva	Participação dos serviços no PIB	IBGE	2023
Vocação produtiva	Participação da administração pública no PIB	IBGE	2023
Vocação produtiva	Valor da produção agrícola per capita	IBGE/PAM	2024
Vocação produtiva	Rebanho bovino per capita	IBGE/PPM	2024
Vocação produtiva	Percentual de área plantada	IBGE/PAM	2024
Conectividade	Percentual de moradores cobertos por 4G/5G	ANATEL	2026
Inovação e capital	Índice de Ecossistema de Inovação	Sebrae	2026
Educação	Matrículas em formação inicial e continuada, FIC	INEP	2024
Educação	Matrículas em educação profissional técnica	INEP	2024
Educação	Matrículas no ensino superior	INEP	2024
Educação	Total de matrículas das modalidades consideradas	INEP	2024
Infraestrutura	Quantidade de empresas de base tecnológica	Inova Simples	2026
Infraestrutura	Percentual da população atendida com água	SINISA	2023
Infraestrutura	Percentual da população atendida com esgoto	SINISA	2023
Infraestrutura	Percentual da população atendida com coleta de lixo	SINISA	2023

Fonte: documentação das bases e referências validadas pelo responsável pelo projeto. O quadro registra as informações originalmente levantadas, antes dos tratamentos e das exclusões.

Anexo B Registro dos tratamentos por indicador

O quadro reproduz os tratamentos documentados para os 34 indicadores preparados, antes das exclusões. A representação contínua precede a padronização. Os quatro indicadores de presença permanecem em 0/1 na matriz da modelagem.

Quadro 10 Tratamentos e representação final por indicador

Indicador	Ausentes antes	Tratamento de ausentes	Transformação	Representação
Matrículas em formação inicial e continuada, FIC	1	Mediana	Log1p	Presença 0/1
Matrículas em educação profissional técnica	1	Mediana	Log1p	Presença 0/1
Fluxo de comércio per capita	1	Mediana	Log1p	Contínua
Total de matrículas das modalidades consideradas	0	Sem ausentes	Log1p	Contínua
Saldo de empregos formais	0	Sem ausentes	Yeo-Johnson	Contínua
Crescimento populacional	0	Sem ausentes	Yeo-Johnson	Contínua
Matrículas no ensino superior	38	Mediana	Log1p	Contínua
Crescimento populacional	0	Sem ausentes	Nenhuma	Contínua
Densidade demográfica	1	Mediana	Box-Cox	Contínua
Taxa de urbanização	1	Mediana	Nenhuma	Contínua
Grau de dependência	1	Mediana	Nenhuma	Contínua
PIB per capita	1	Mediana	Box-Cox	Contínua
Densidade empresarial	1	Mediana	Box-Cox	Contínua
Massa salarial per capita	0	Sem ausentes	Box-Cox	Contínua
Taxa de sobrevivência empresarial em três anos	1	Mediana	Nenhuma	Contínua
Exportação FOB em relação ao PIB	49	Mediana	Box-Cox	Contínua
Crédito per capita	65	Mediana	Nenhuma	Contínua
Percentual da população em situação de pobreza	0	Sem ausentes	Nenhuma	Contínua
Participação das classes D e E	1	Mediana	Nenhuma	Contínua
Percentual de beneficiários do Bolsa Família	0	Sem ausentes	Box-Cox	Contínua
Índice de Gini	1	Mediana	Box-Cox	Contínua
Participação da agropecuária no PIB	1	Mediana	Nenhuma	Contínua

Indicador	Ausentes antes	Tratamento de ausentes	Transformação	Representação
Participação da indústria no PIB	1	Mediana	Box-Cox	Contínua
Participação dos serviços no PIB	1	Mediana	Box-Cox	Contínua
Participação da administração pública no PIB	1	Mediana	Box-Cox	Contínua
Valor da produção agrícola per capita	1	Mediana	Box-Cox	Contínua
Rebanho bovino per capita	1	Mediana	Box-Cox	Contínua
Percentual de área plantada	1	Mediana	Box-Cox	Contínua
Percentual de moradores cobertos por 4G/5G	1	Mediana	Yeo-Johnson	Contínua
Índice de Ecossistema de Inovação	105	Preservados	Box-Cox	Contínua
Quantidade de empresas de base tecnológica	117	Zero	Box-Cox	Presença 0/1
Percentual da população atendida com água	20	Mediana	Yeo-Johnson	Contínua
Percentual da população atendida com esgoto	102	Zero	Nenhuma	Presença 0/1
Percentual da população atendida com coleta de lixo	17	Mediana	Yeo-Johnson	Contínua

Fonte: log de tratamentos aplicados. A representação contínua precede a padronização; os indicadores excluídos também aparecem neste registro porque passaram pela etapa de tratamento. A imputação de matrículas FIC e técnicas utilizou mediana igual a zero.

A mediana foi utilizada nos casos com ausentes, exceto nas duas decisões de substituição por zero e no Índice de Ecossistema de Inovação, cujas ausências foram preservadas. A representação final dos quatro indicadores binários preserva a presença identificada após o tratamento de ausentes.

As duas séries de crescimento populacional aparecem com o mesmo rótulo no registro original de tratamentos. O Anexo A distingue seus períodos, 2021–2024 e 2024–2025. Este quadro preserva o registro fornecido sem atribuir uma transformação específica a um intervalo que não foi identificado nele.

SEBRAE