

Tema:

Estimativa de Indicadores Socioeconômicos a partir de Imagens de Satélite na área NEXUS

Introdução

O Brasil é o quinto maior país do mundo e a área Nexus (Figura 1), contempla as bacias dos rios São Francisco e Parnaíba, o equivalente a aproximadamente 40% do território nacional, representando uma rica diversidade de recursos naturais. Por conta disso, há a motivação de aplicar políticas de preservação e desenvolvimento sustentável para proteger toda essa diversidade, as quais são auxiliadas pelo monitoramento da região através de indicadores ambientais e socioeconômicos.

Entretanto, esta tarefa se torna difícil dada a baixa frequência de aquisição de dados, já que o censo é conduzido apenas a cada 10 anos. Além disso, há uma grande disponibilidade de imagens de satélite capturadas com uma alta frequência por diversas fontes.

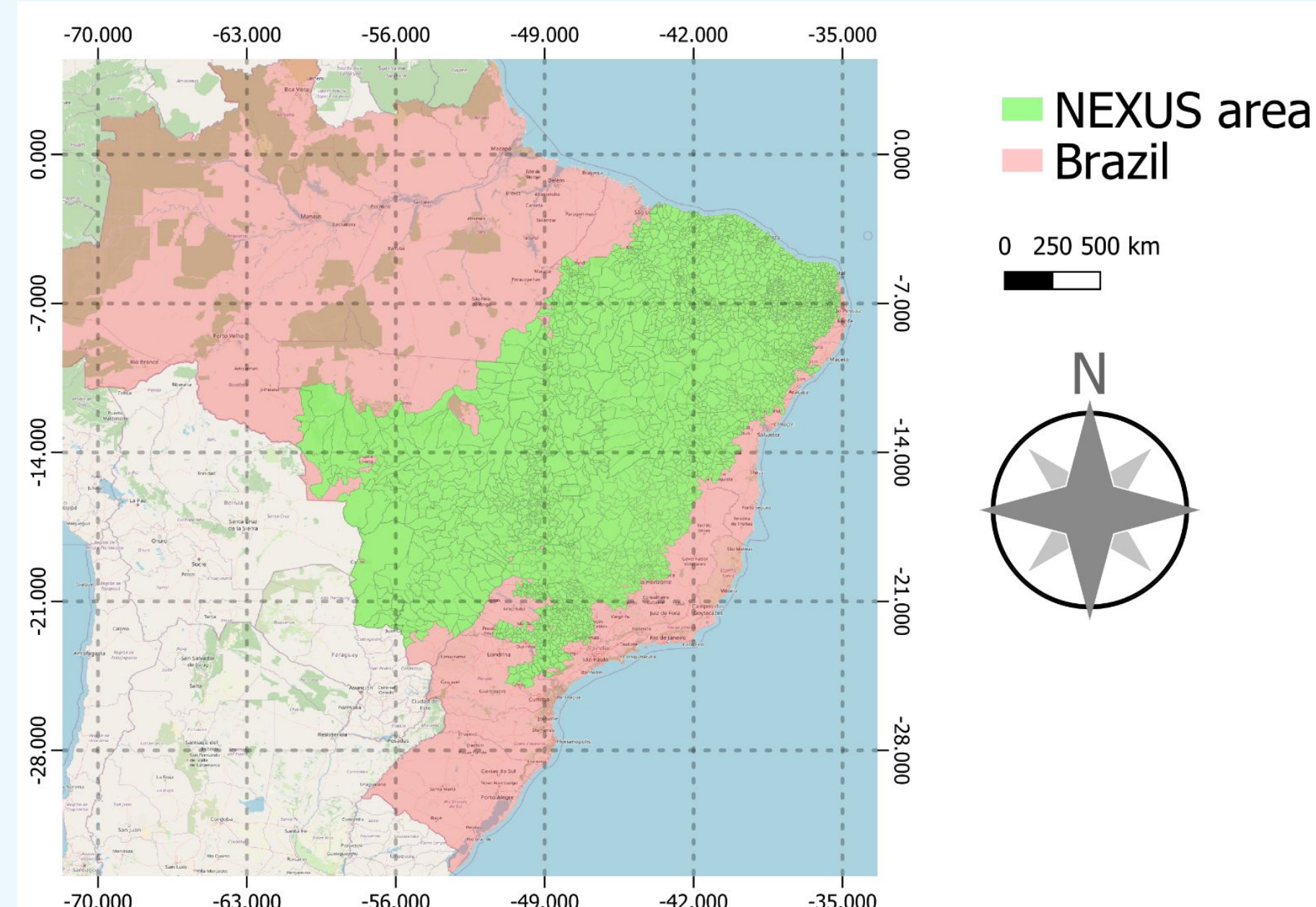


Figura 1 - Mapa da área NEXUS

Metodologia

1. Indicadores socioeconômicos

Obtenção e filtragem de indicadores na granularidade de setores censitários. Os 3 indicadores considerados como relevantes para estudar o desenvolvimento socioeconômico de comunidades são longevidade, renda e alfabetização.

2. Regiões a serem capturadas por imagens de satélite

A seleção das regiões das imagens de satélite exige um critério rigoroso com condições que visam a garantir um modelo não enviesado e redução do custo de armazenamento: (a) cada imagem deve representar apenas um tipo de setores: urbano ou rural, (b) estes dois tipos devem estar disponíveis na base de dados em uma proporção que retrate a realidade, (c) uma determinada região só deve estar presente em apenas uma imagem selecionada.

3. Imagens de satélite

A coleta das imagens se dá por meio da API do *Google Earth Engine*. As principais informações a serem definidas são as coordenadas do centro das imagens (*clusters*), a escala (30 metros/pixel), a resolução (224 x 224 pixels) e o ano.

4. Adaptação do algoritmo de referência para a Área NEXUS

O artigo de referência, Yeh. et al (2020), estima índices de riqueza em diversos países Africanos. A estrutura da rede neural convolucional (Figura 2) utilizada apresenta entrada dupla: imagens de satélite diurnas e imagens noturnas ("*nightlights*"). O modelo conta com dois backbones de Resnet 18 pré-treinados para a extração de features das imagens. A camada *fully connected* combina as informações extraídas das entradas para uma regressão de Ridge, a qual avalia a capacidade do modelo de explicar os indicadores estimados por meio de seu coeficiente de determinação.

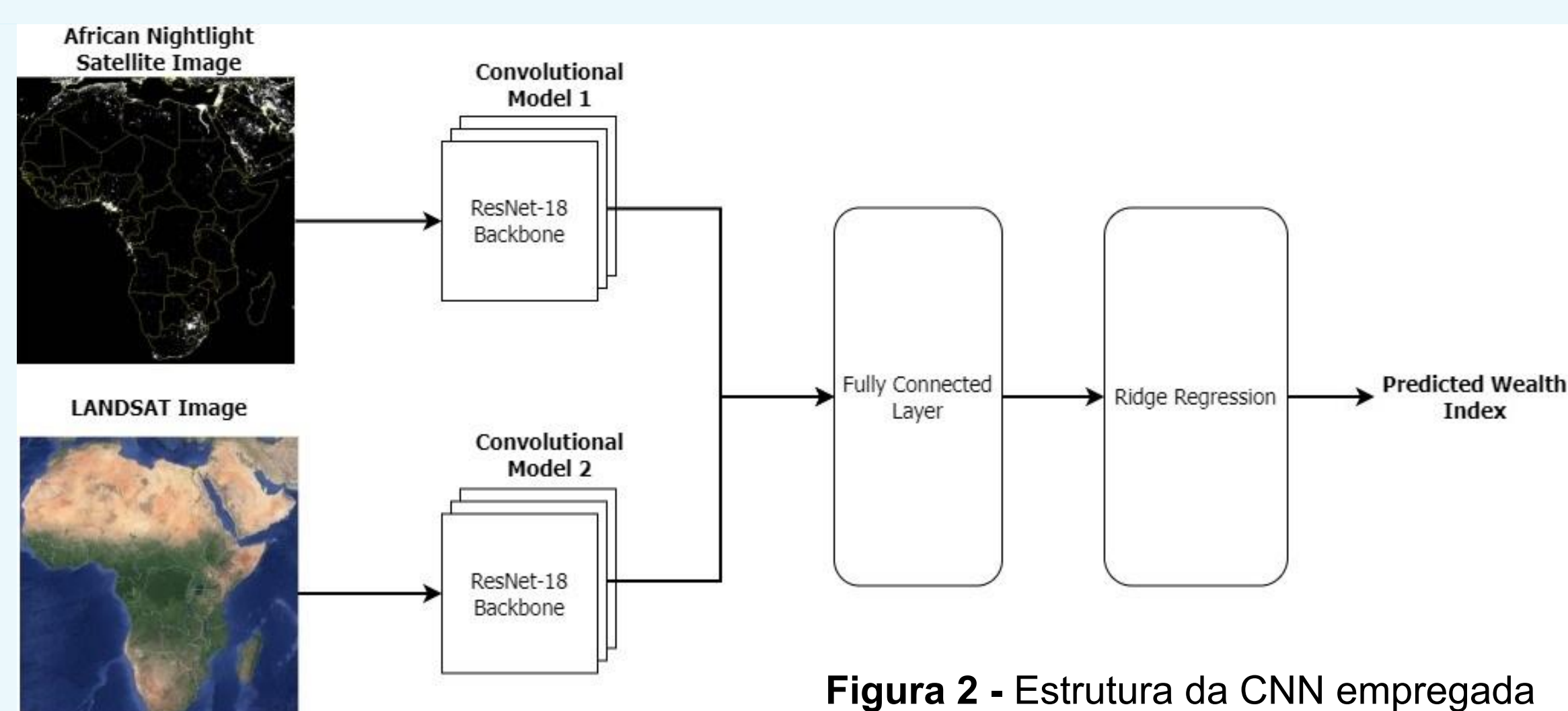


Figura 2 - Estrutura da CNN empregada

5. Análise dos resultados obtidos pela Rede Neural Convolucional (CNN)

Com os indicadores estimados, é possível fazer análises como por exemplo a correlação entre eles, a sua distribuição ao longo dos setores censitários da área NEXUS e as possíveis diferenças entre setores urbanos e rurais.

6. Desenvolvimento da plataforma de visualização

A plataforma permitirá que o usuário visualize estimativas dos indicadores socioeconômicos de uma região arbitrária, fornecendo seu centróide. A arquitetura desse sistema é dada de forma distribuída utilizando serviços da AWS, como: *API Gateway* para estabelecer a comunicação entre a interface da plataforma e os serviços na nuvem, *StepFunctions* para definir o fluxo de trabalho da aplicação, serviços de computação serverless para processamento de requisições e *DynamoDB* para armazenar e consultar os dados.

Resultados

Este trabalho contribui através da criação de uma nova base de dados, originada da mescla de uma base de mais de 100 indicadores socioeconômicos fornecida pelo INPE, e do conjunto de informações administrativas, geográficas e geométricas dos setores censitários brasileiros, fornecido pelo IBGE. Os setores censitários dentro da área NEXUS são agrupados em diversos grupos urbanos ou rurais, os quais passam por um processo de clusterização (Figura 3) para gerar as amostras da base de dados.

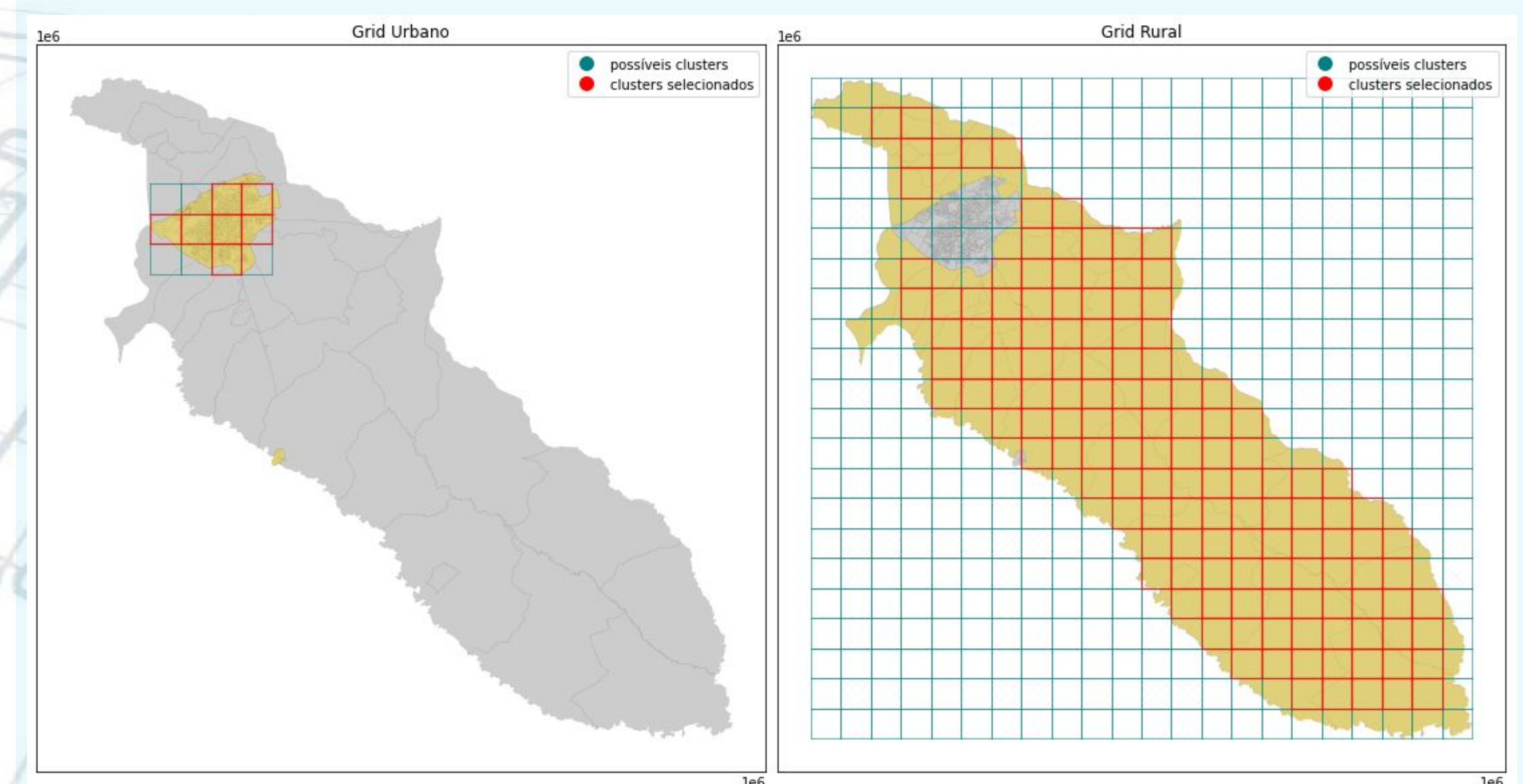


Figura 3 - Exemplo de clusterização do banco de dados

Três critérios são avaliados: longevidade, renda e alfabetização, uma vez que estes são os que melhor se aproximam das métricas de IDH. Para cada setor censitário, a longevidade é calculada a partir de sua expectativa de vida, enquanto que a renda é obtida por uma relação entre a renda per capita e o PIB do município e a alfabetização leva em conta a proporção de pessoas alfabetizadas antes e depois de 14 anos de idade.

Com esta base de dados, foram treinados modelos preditivos de base (*BaseModels*) para validá-la e também avaliar qual dos três indicadores têm maior correlação com as características das imagens. Os melhores resultados, apresentados na Tabela 01, foram para o indicador renda. Além disso, foram treinados dois modelos de *Deep Learning*, um para imagens multiespectrais (MS) e outro para *NightLights* (NL). Com estes resultados, conclui-se que NLS, no caso do Brasil, não são dados representativos para os indicadores, diferente de MS que apresenta um ótimo coeficiente de determinação (r^2). Tendo em vista os resultados obtidos, entende-se que a adaptação da metodologia de referência proposta por este trabalho é adequada para a estimativa de indicadores socioeconômicos no cenário brasileiro.

Experimentos com o indicador Renda				
BaseModel	r2	R2	mse	rank
Ridge MS+NL hist	0.32390	0.32389	0.01687	0.63732
Ridge MS hist	0.31866	0.31862	0.01700	0.63291
Ridge RGB+NL hist	0.22525	0.22525	0.01933	0.53157
Ridge RGB hist	0.21615	0.21615	0.01956	0.52053
Ridge NL hist	0.01265	0.01263	0.02464	0.06251
KNN NL center scalar	0.09975	0.08367	0.02474	0.05500
KNN NL hist	0.00933	0.00832	0.02474	0.07183
Ridge NL mean scalar	0.00893	0.00893	0.02473	-0.00815
Ridge NL center scalar	0.00888	0.00888	0.02473	0.00092
KNN NL mean scalar	0.00673	0.00532	0.02482	0.05620
Deep Learning				
Resnet_18 MS validação	0.45161	0.45132	0.01369	0.72836
Resnet_18 NL validação	0.00292	-0.0175	0.02450	0.02872