Dirección de Desarrollo Digital

Unidad de Científicos de Datos

El futuro es de todos

DNP
Departamento Nacional de Planeación

IDENTIFICACIÓN DE VIVIENDAS SIN SERVICIO DE ENERGÍA ELÉCTRICA BASADO EN ANÁLISIS DE IMÁGENES SATELITALES RGB – FASE 1

Entidad
Departamento Nacional de Planeación
- Dirección de Desarrollo Digital.
- Subdirección de Minas y Energía

Sector	**Lenguaje**	**Fuente de datos**
Planeación	Python	Google Maps, ESRI, información de infraestructura eléctrica UPME

Presentación

Colombia cuenta con una cobertura de infraestructura eléctrica a nivel nacional estimada de 96.96%, siendo 99.72% en el área urbana y un 87.83% en el área rural. Con el fin de lograr el 100% de cobertura, es vital identificar la ubicación geográfica de ese porcentaje de la población, con el objetivo de definir estrategias para brindar el servicio de energía. Uno de los grandes obstáculos que se presentan para lograr la total cobertura, es la falta de información precisa y actualizada sobre la ubicación de las viviendas que no cuentan con servicio de energía. Además, la mayoría de estas, se encuentran en zonas aisladas y de difícil acceso. Lo anterior agrega desafíos para censar y localizar las viviendas de interés.

Por otro lado, las imágenes satelitales se han convertido en una fuente de datos que permite encontrar información relevante en grandes extensiones de terreno y que dan una alternativa para la solución del problema mencionado anteriormente. Para dar solución a esta problemática, desde la Unidad de Cientificos de Datos, se propone una metodología de análisis de imágenes satelitales RGB (obtenidas de Google Maps y ESRI) para detectar construcciones dentro de la imagen. Cada detección hecha, cuenta con información georreferenciada de la ubicación, permitiendo cruzarla con los datos de infraestructura eléctrica (UPME), de tal manera que, se pueda dar un número estimado de viviendas que no cuentan con el servicio de electricidad. Esto dará insumos a las subdirecciones técnicas para definir estrategias y toma de decisiones para desarrollar estrategias de cobertura de cara al futuro.

En la presente sección, se describe de forma detallada la fase 1 del proyecto, que consisten en la detección de viviendas mediante el análisis y procesamiento de imágenes satelitales RGB. Los resultados presentados son pilotos en el municipio de Puerto Leguizamo – Putumayo.

Colombia has an estimated electrical infrastructure coverage of 96.96%, with 99.72% in the urban area and 87.83% in the rural area. To achieve 100% coverage, it is vital to identify the geographic location of that percentage of the population, to define strategies to provide energy service. One of the great obstacles that arise to reach full coverage is the lack of accurate and updated information on the location of homes that do not have energy service. Besides, most of these are in isolated areas and difficult to access. The above includes challenges to census and locate the homes of interest.

On the other hand, satellite images have become a data source that allows finding relevant information in large areas of land and that provide an alternative for solving the problem mentioned above. To solve this, from the

Data Scientists Unit, an RGB satellite image analysis methodology (obtained from Google Maps and ESRI) is proposed to detect constructions within the image. Each detection made, has geo-referenced information, allowing it to be crossed with the electrical infrastructure data so that an estimated number of homes can be provided that do not have electricity service. This will provide inputs to the technical sub-directorates to define strategies and decision-making to provide this service.

In this section, phase 1 of the project is described in detail, which consists of the detection of homes through the analysis and processing of RGB satellite images. The results presented are pilots in the municipality of Puerto Leguizamo – Putumayo.

Objetivo general

Diseñar una metodología para identificar las viviendas en zonas aisladas a través de análisis de imágenes satelitales RBG, y generar información georreferenciada de las detecciones realizadas.

Objetivos específicos

1. Diseñar una metodología de análisis de imágenes satelitales RGB que permita la detección de construcciones dentro de la imagen satelital en regiones aisladas.
2. Entrenar un modelo de aprendizaje de máquina para clasificar regiones (1 región con construcciones, 0 región sin construcciones) dentro de la imagen satelital.
3. Transformar la información de las construcciones detectadas en la imagen a información georreferenciada para que puedan ser cruzadas con información de infraestructura eléctrica.
4. Generar capas de información georreferenciada en formatos shapefile, para que sirvan de insumo en programas especializados como ArcGis o QGIS.

Metodología: Análisis de imágenes satelitales RGB

La metodología se dividió en dos fases: En esta sección se describirá la metodología de la Fase 1, que corresponde a la detección de construcciones y la generación de información georreferenciada de las detecciones realizadas. En la segunda Fase, que no cubre este informe, se cruzará la información de infraestructura eléctrica con la información georreferenciada de cada vivienda detectada para estimar cuales, de ellas, no tienen servicio de infraestructura eléctrica.

La fase de análisis de imágenes satelitales tiene 5 etapas como se ilustra en la Figura 1, las cuales son adquisición, preprocesamiento, caracterización, clasificación y decisión.

Figura 1. Etapas de la metodología de análisis de imágenes satelitales.

1. ***Adquisición:***

El primer paso de la metodología es obtener las imágenes satelitales de Google Maps y ESRI. Para ello es necesario construir la región de análisis mediante un arreglo de baldosas de 256 x 256 pixeles georreferenciados denominadas tiles, dado este hecho, el tamaño mínimo de la imagen de análisis equivale a (1) un tile, como se ilustra en la **¡Error! No se encuentra el origen de la referencia.**. Cada tile se obtiene a un nivel de zoom de 17, por lo que cada pixel en la imagen representa 1.193 metros sobre el ecuador[1].

Figura 2: Los tiles obtenidos por servicios web de Google Maps y ESRI se organizan para formar la imagen satelital RGB para la detección de construcciones.

2. ***Preprocesamiento:***

Cuando se obtiene la imagen, esta presenta en la mayoría de los casos bajos contrastes debido a sombras de nubes cercanas a la región de análisis, por tal motivo es necesario ecualizar los histogramas de color para aumentar el contraste. La ecualización logra una mejor distribución de las intensidades de color sobre todos los canales RGB. En la Figura, se ilustra como cambia la distribución de los histogramas de color después de la ecualización y como esto mejora el contraste de la imagen, dando una mejor representación de los espacios de color.

[1] https://wiki.openstreetmap.org/wiki/Zoom_levels

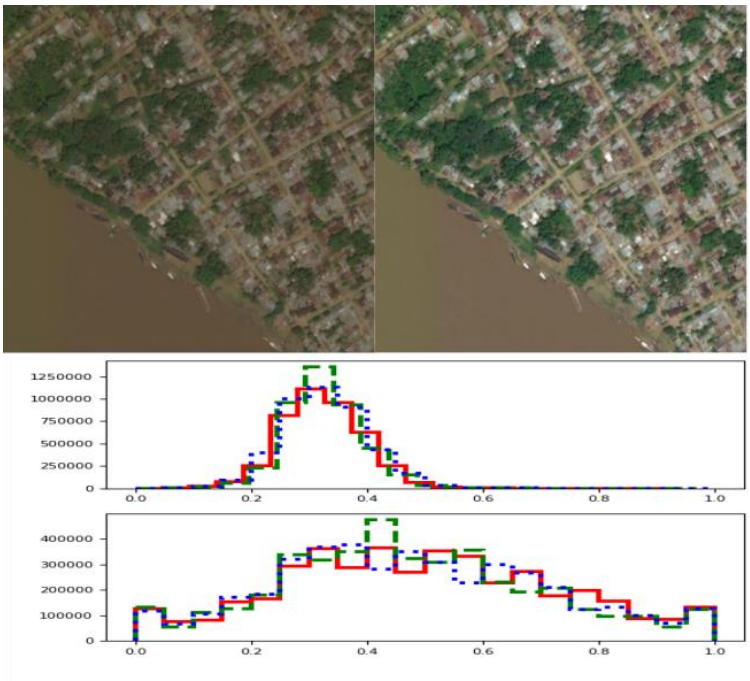

Figura 3. Ecualización de los histogramas de Color. Esta técnica ayuda a mejorar el contraste de la imagen.

3. Caracterización:

En esta etapa, la imagen satelital RGB se divide en una rejilla de pixeles, donde cada cuadricula esta formada por ventanas de pixeles de 16 x 16, como se ilustra en la figura 4.

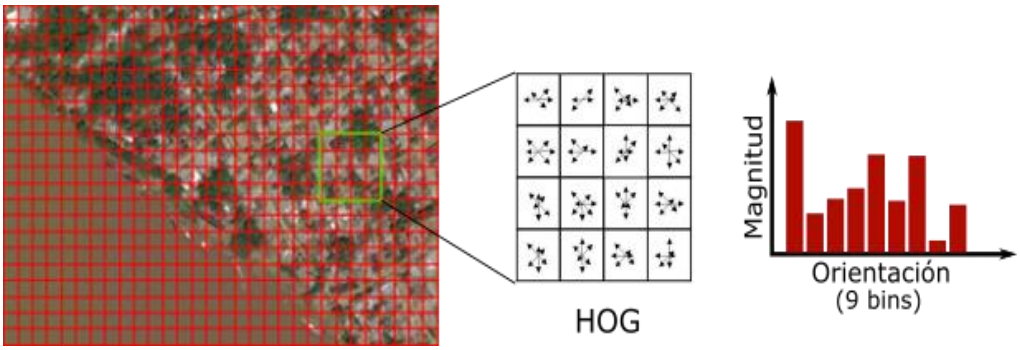

Figura 4. La imagen satelital RGB es dividida en pequeñas cuadriculas para identificar si hay construcciones dentro de ellas, luego cada ventana es representada por descriptores HOG.

Después, cada una de estas ventanas es representada por descriptores HOG (Histogram Orientation Gradients). Estos descriptores codifican la imagen mediante características de bajo nivel, capaces de codificar los bordes y esquinas presentes en la imagen mediante los cambios de magnitud y

orientaciones de gradiente en la imagen. Estos cambios de dirección y magnitud son agrupados en un histograma de n bins (particiones del histograma), como se muestra en la Figura 4, dando una representación numérica por cada ventana de 16x16 pixeles.

4. ***Clasificación:***

   En la etapa de clasificación, existe un paso preliminar que corresponde al etiquetado de cada una de las ventanas del paso anterior. En otras palabras, a cada una de estas regiones se le asigna una etiqueta (1 si la región contiene construcción, 0 en el caso contrario). A partir de esto, se genera un conjunto de datos de entrenamiento, donde cada muestra es una representación numérica de los descriptores HOG encontrados en cada rejilla. Paso siguiente, se entrena un modelo de clasificación capaz de tener una buena discriminación entre construcción y no construcción.

5. ***Decisión:***

   En esta instancia, una nueva imagen se pasa por toda la metodología expuesta en la Figura 1, obviando la parte del etiquetado de regiones, debido a que estas serán estimadas por el clasificador entrenado en el paso anterior. Las regiones positivas (con construcción) se resaltan en la imagen y a partir de la información georreferenciada de la imagen, se hace la estimación para la información georreferenciada de cada región. Luego esta información servirá de insumo para identificar cuáles de ellas no tienen servicio de energía eléctrica al ser cruzadas con los datos de infraestructura eléctrica de la UPME.

Resultados

En esta sección se describen los resultados obtenidos en la etapa de entrenamiento del modelo de clasificación y el desempeño del mejor estimador encontrado a partir de la búsqueda de la combinación de parámetros con respecto a la medida F1, definida como sigue:

$$F_1 = \frac{2TP}{2TP + FP + FN}$$

La matriz de datos de entrenamiento está formada por el número de cuadriculas de 16 x 16 pixeles en la imagen por el número de orientaciones HOG que presenten el mayor desempeño de clasificación. Para el caso específico, se entrenó el modelo con una imagen de 1280 x 2304 pixeles, donde hay 11520 cuadriculas. El número de cuadriculas con etiqueta 1 (con construcción) fue de 3654, lo que corresponde a un 31.7% de las cuadriculas y un 68.3% con etiqueta 0.

Para evitar el desbalance en los datos de entrenamiento, se hizo un submuestreo aleatorio de la clase negativa para obtener el mismo número de muestras que la clase positiva, es decir, 3654 ventanas con etiqueta 0.

Para entrenar el modelo de clasificación se utiliza una metodología de validación cruzada de 5 folds. Se entrenó diferentes clasificadores variando el número de orientaciones de los descriptores HOGs y los hiper-parámetros propios de cada clasificador. En la Figura 5, se muestra el desempeño del estimador con mejor desempeño del conjunto de estimadores probados. A partir de la validación cruzada, se observa que el mejor rendimiento con respecto a la medida F1 (67.6% +/- 0.2) se obtiene con 52 orientaciones, para un clasificador SVM con parámetros C = 0.5 y un kernel RBF con gamma = 0.01.

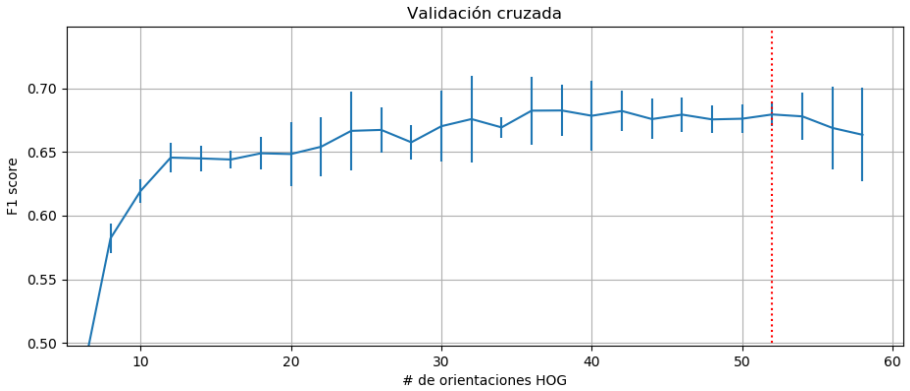

Figura 5. Desempeño del mejor estimador a distintas orientaciones de los descriptores HOG en validación cruzada.

Después de entrenar el clasificador, este se prueba con una nueva imagen y observar de forma visual los resultados de clasificación de viviendas. En la Figura 6, se muestra una imagen de una región aislada de Puerto Leguizamo – Putumayo. Las regiones resaltadas con rojo son las que el clasificador estimo como viviendas. Además, se observa que, aunque existe una cantidad elevada de falsos positivos (estimaciones que no son vivienda), detecto todas las viviendas presentes en la imagen de prueba.

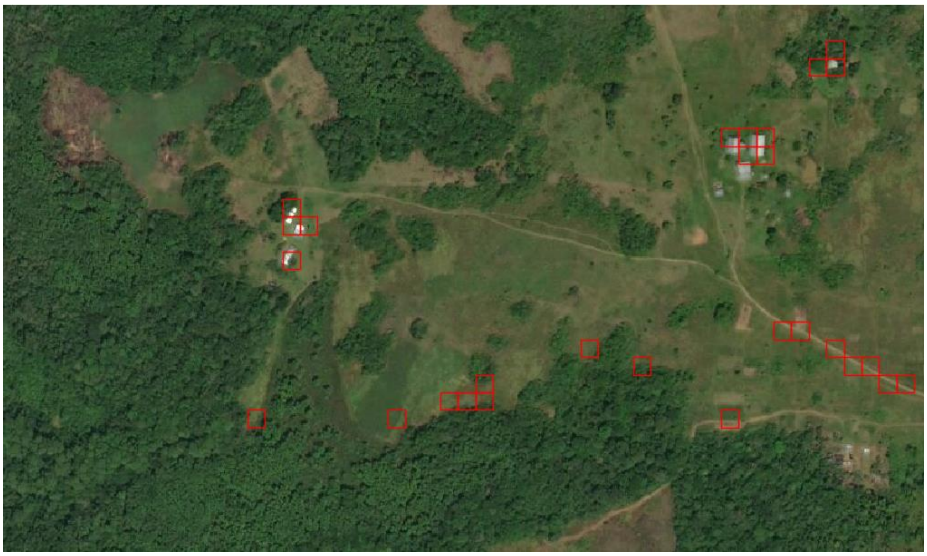

Figura 6. Resultados de clasificación de la detección de viviendas en zonas aisladas. Las regiones en rojo son las que el clasificador estimó como construcción.

Conclusiones y recomendaciones

1. Se desarrolló una metodología de detección de viviendas en zonas aisladas en el municipio de Puerto Leguizamo, la metodología presentó una tasa aceptable de falsos positivos, y una tasa baja de falsos negativos, lo cual con respecto al objetivo de detección de viviendas en zonas aisladas es deseable.

2. A partir de las detecciones que hace el algoritmo, se obtiene información georreferenciada de cada una de las detecciones (viviendas). Esto servirá de insumo a las direcciones técnicas para identificar cuáles de ellas no cuentan con infraestructura eléctrica, y de cara a la posibilidad de la segunda Fase del proyecto, que es realizar esto de forma automática contando con la información de infraestructura eléctrica del país.

3. Los resultados obtenidos en este piloto se pueden expandir al resto del territorio nacional. Para ello es necesario ampliar los datos de entrenamiento con imágenes de otras regiones que contengan la información de etiquetas (1- vivienda, 0- no vivienda).

4. Los resultados obtenidos muestran que las metodologías basadas en análisis de imágenes satelitales pueden ser de gran impacto para el desarrollo de políticas energéticas y planes de electrificación del país.

Socialización

El avance del proyecto se ha socializado con la Subdirección de Minas y Energía de la Dirección de Infraestructura y Desarrollo Sostenible, con la Dirección de Desarrollo Digital y con la Dirección de Desarrollo Urbano.