Dirección de Desarrollo Digital

Unidad de Científicos de Datos

El futuro es de todos

DNP
Departamento Nacional de Planeación

IDENTIFICACIÓN DE VÍAS TERCIARIAS EN IMÁGENES SATELITALES

Entidad

Departamento Nacional de Planeación.

- Dirección de Infraestructura y Energía Sostenible.
- Dirección de Desarrollo Digital.

Ministerio de Transporte.

Sector	**Lenguaje**	**Fuente de datos**
Transporte.	Python.	Instituto Geográfico Agustín Codazzi (IGAC) e Instituto Nacional de Vías (Invías).

Presentación

Este proyecto estudia la viabilidad del uso de algoritmos de aprendizaje de máquina para la identificación y detección de vías terciarias en imágenes satelitales. Para esto, se tuvieron en cuenta dos algoritmos: redes neuronales convolucionales y máquinas de soporte vectorial. Estos algoritmos fueron utilizados en una prueba piloto, utilizando como insumos imágenes satelitales del departamento de Santander, junto con la ubicación demarcada de algunas vías terciarias. Los resultados obtenidos son alentadores; en particular, el modelo obtenido utilizando máquinas de soporte vectorial alcanza una precisión del 86.5% en datos de validación. Esta cifra permite concluir que, con la suficiente cantidad, variedad y calidad de imágenes y vías demarcadas de diferentes regiones del país, es posible entrenar un modelo que permita encontrar las vías terciarias en todo el territorio nacional Colombia con un desempeño aceptable.

This project studies the feasibility of using machine learning algorithms to identify tertiary roadways in satellite imagery. Two methods are tested: convolutional neural networks and support vector machines. These algorithms were used in a pilot test, over a sample of images covering the Santander Department, together with the demarcated location of some tertiary roads in the area. The results of the pilot test are encouraging: the support vector machine model was able to achieve a precision of 86.5% over an independent test set. This suggests that, given enough varied images of high quality and a good sample of roads already tracked, it is possible to train a model to locate tertiary roads over the whole area of the country with acceptable performance levels.

Objetivo general

Establecer un intervalo de precisión que permita determinar si es factible la inversión en un equipo dedicado a obtener el primer archivo cartográfico con la malla vial terciaria de Colombia.

Objetivos específicos

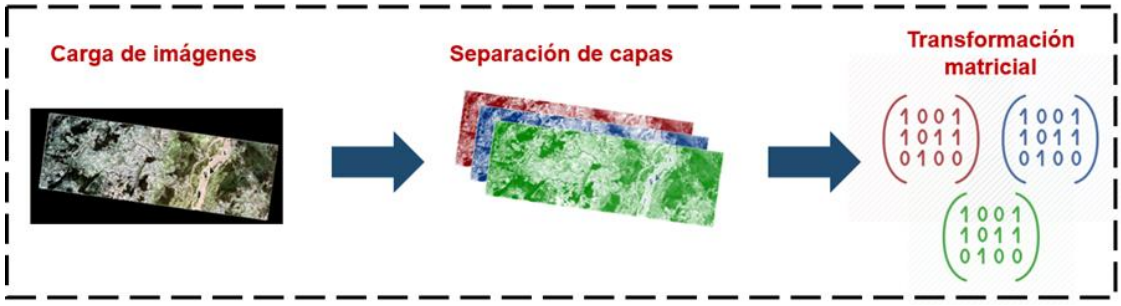
- Entender los insumos suministrados (imágenes satelitales [*raster*], *shape files*) y acceder a ellos para obtener la información que contienen.
- Procesar la información disponible, para obtener un conjunto de datos etiquetados que permita entrenar y validar los modelos de aprendizaje de máquina.
- Entrenar modelos de aprendizaje de máquina, siguiendo las dos metodologías propuestas (redes neuronales convolucionales y máquinas de soporte vectorial).
- Validar el desempeño de los modelos desarrollados sobre un conjunto de datos de prueba.

Metodología

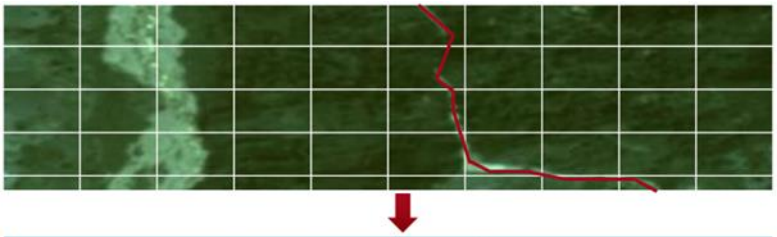
La metodología para el desarrollo del proyecto se dividió en dos pasos principales: el preprocesamiento de los datos y el entrenamiento de los modelos. A continuación, se describen estos dos pasos.

Preprocesamiento

Las imágenes satelitales que fueron utilizadas como insumo estaban en formato *raster*. Cada imagen fue cargada a través de Python y separada en 4 matrices, en las cuales cada elemento de la matriz corresponde a un pixel de la imagen. Las primeras 3 matrices corresponden a la capa de rojos, verdes y azules, más una cuarta matriz que tiene información sobre la transparencia de la imagen.



A las imágenes se les agrega una quinta capa/matriz, que proviene de los *shape files* y corresponde a la ubicación de las vías terciarias. Esta matriz contiene un 1 si en dicho pixel hay vía, y 0 en caso contrario. Utilizando la información de esta capa, las imágenes satelitales son divididas en ventanas más pequeñas, y cada ventana es etiquetada de acuerdo con si contiene o no vías terciarias.



Identificación de vías *									
No	No	No	No	No	Si	No	No	No	No
No	No	No	No	No	Si	No	No	No	No
No	No	No	No	No	Si	No	No	No	No
No	No	No	No	No	No	Si	Si	Si	No

Entrenamiento de modelos

Una vez se obtienen las ventanas etiquetadas, estas son utilizadas como insumo para entrenar los modelos. Se tuvieron en consideración dos algoritmos para realizar el modelo de aprendizaje supervisado: Redes Neuronales Convolucionales (CNN por sus siglas en inglés) y Máquinas de soporte vectorial (SVM). La metodología para desarrollar los modelos fue la siguiente:

1. Lectura y preprocesamiento de los datos (descrito en la sección anterior).
2. Procesamiento adicional de los datos: En el caso del modelo SVM, se hicieron dos pasos adicionales para procesar los datos, antes de entrenar el modelo.
   a. El primero de estos pasos fue el de balancear la cantidad de datos (ventanas) que había por cada categoría. Originalmente las ventanas que contenían vías eran menos de un 1% de los datos totales. Para remediar esto, se usaron técnicas de muestreo y sobre muestreo para balancear ambas categorías y llegar a un conjunto de datos que sea aproximadamente 40% vía y 60% no vía.
   b. El segundo paso fue utilizar técnicas de reducción de dimensionalidad, en particular el análisis de componentes principales, para reducir el número de variables que representan cada ventana de la imagen, sin perder demasiada información.

3. Una vez se tiene el conjunto de datos listo, este se dividió en grupos de entrenamiento (para desarrollar el modelo) y validación (para probar el desempeño del modelo en datos distintos a los de entrenamiento).
4. Se entrenó el modelo sobre los datos de entrenamiento.
5. Se probó el desempeño del modelo sobre los datos de validación. Este es un proceso iterativo en el cual se van variando parámetros del modelo en la etapa de entrenamiento, y se observa qué combinación de parámetros logra un mejor desempeño (definido como precisión en la clasificación) sobre los datos de validación.

Resultados

Redes Neuronales Convolucionales

A continuación, se muestra el desempeño que obtuvo el modelo de CNN sobre los datos de validación. Se presenta la precisión obtenida sobre estos datos, así como la matriz de confusión, que señala la cantidad de registros que fueron predichos correcta e incorrectamente.

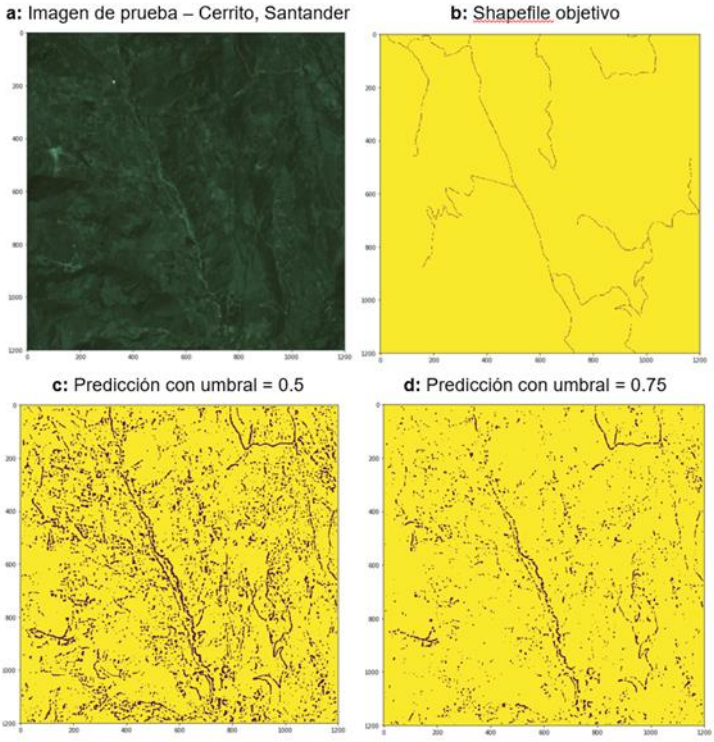
	Datos de validación		
Precisión	**54%**		
Matriz de confusión	**Real / Predicción**	**No vía**	**Vía**
	No vía	116	74
	Vía	47	28

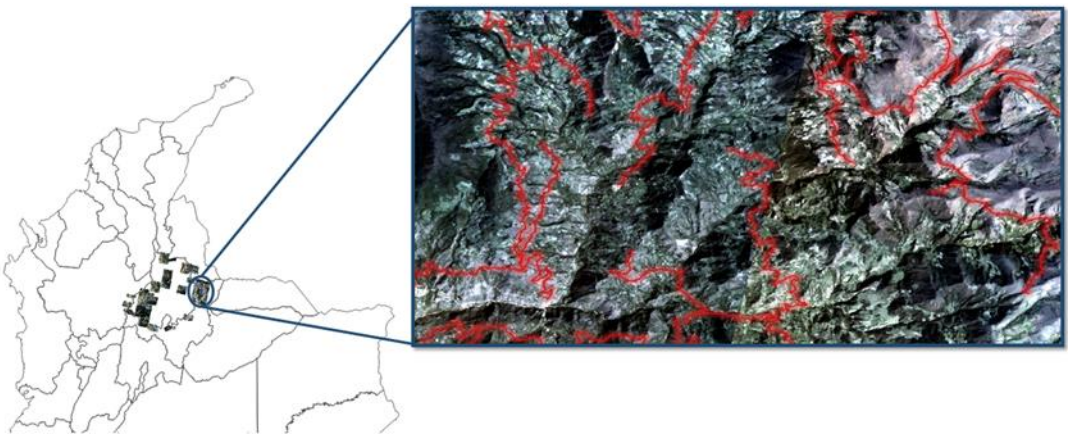
Máquinas de Soporte Vectorial

De forma similar, la Tabla 2 muestra el desempeño obtenido por el modelo SVM, tanto en los datos de entrenamiento como de validación. Se puede observar que el desempeño en ambos grupos de datos es muy similar, lo que muestra que no hubo sobreajuste en el entrenamiento.

	Datos de entrenamiento			**Datos de validación**		
Precisión	**86,70%**			**86.5%**		
Matriz de confusión	**Real / Predicción**	**No vía**	**Vía**	**Real / Predicción**	**No vía**	**Vía**
	No vía	34747	1382	**No vía**	11537	454
	Vía	6779	18445	**Vía**	2301	6159

El modelo SVM obtenido fue utilizado para clasificar una de las imágenes satelitales del departamento de Santander no utilizadas en el entrenamiento. La idea era comparar el *shape file* (ubicación de las vías terciarias) disponible para la imagen, con el que produce el modelo. Los resultados de esta prueba se pueden observar en la siguiente gráfica. Las dos imágenes de abajo (c y d) corresponden a las predicciones del modelo, variando el umbral de probabilidad con el cual se determina si una ventana en particular contiene vía o no.





Conclusiones

1. El uso de modelos de aprendizaje de máquina es una buena opción para abordar este problema en particular.
2. Los datos de desempeño obtenidos permiten considerar esta opción como una alternativa viable y de bajo costo para identificar la red de vías terciarias del país.
3. Adicionalmente, los resultados obtenidos pueden ser mejorados al incrementar el número y la calidad de imágenes satelitales.
4. También es importante contar con imágenes de distintas regiones de Colombia, para poder desarrollar un modelo más robusto.

Socialización

Los resultados de la prueba piloto fueron presentados internamente a la Dirección de Desarrollo Digital y la Dirección de Infraestructura y Energía Sostenible. Externamente al Ministerio de Transporte y al Instituto Geográfico Agustín Codazzi, como a entidades interesadas. De igual, los resultados se presentaron al Banco Interamericano de Desarrollo y al Banco Mundial como apoyo a la solicitud de recursos multilaterales para el escalamiento del proyecto a nivel nacional.