**Fecha:** 14 de marzo de 2022

**Versión:** 0.0\_2022\_03\_14

**Título:** Estimación de áreas de deslizamientos de tierra basado en modelos de redes neuronales convolucionales y herramientas de Google Earth Engine/Sentinel-Hub sobre imágenes ópticas satelitales.

**IDENTIFICANDO EL PROBLEMA DE INVESTIFACIÓN**

El problema de investigación nace de la necesidad de mejorar la información respecto a los fenómenos de remoción en masa en Colombia. Actualmente el servicio geológico colombiano (SGC) es la entidad gubernamental responsable del monitoreo de los deslizamientos de tierras en el país, para esta finalidad cuanta con el “Sistema de Información de Movimientos en Masa” (SIMMA), este inventario cuenta con información puntual georreferenciada de los deslizamientos en el territorio nacional, pero no cuenta con capas de las superficies afectadas en cada deslizamiento.

Estas labores de reconocimiento se ven supeditadas a estudios regionales de amenaza, vulnerabilidad y riesgos geológicos, estudios cuyas fuentes de información e insumos no siempre son de fácil acceso para fines académicos. Adicionalmente, al tratarse de labores de detección mediante métodos tradicionales de fotointerpretación, suelen ser llevadas a cabo manualmente por personal experto, factor que implica largos periodos de tiempo y costos.

Se ve entonces la necesidad de encontrar medios que permitan estimar las zonas de afectación por deslizamientos, qué no dependa directamente de las labores manuales de personal capacitado, ni que implique grandes costos. Es entonces que la geomática y en especial las técnicas de percepción remota brindan soluciones, que permiten dar una solución parcial pero económicamente viable a la identificación de áreas afectadas por movimientos de remoción en masa en Colombia.

Desde la rama del conocimiento de la teledetección, podemos aprovechar nuevas tecnologías y herramientas, que se encuentran a fácil disposición: En primer lugar aprovechar los avances en computación visual con técnicas como las redes neuronales convolucionales (CNN), algoritmos que han demostrado una importante capacidad de detección de elementos en imágenes (como puede ser la existencia de células cancerígenas en una radiografía e incluso la detección de coberturas de suelos en fotografías satelitales), que mediante lenguajes de programación como Python (como por ejemplo TensorFlow), facilita a investigadores y docentes la automatización de procesamiento de datos. En segundo lugar, plataformas como Google Earth Engine o Sentinel-Hub, que facilitan la búsqueda y aprovechamiento de imágenes satelitales, sirviendo como repositorios para los principales programas espaciales de percepción remota terrestre (por ejemplo, sentinel o landsat), estas plataformas cuentan con interfaces de programación de aplicaciones (API) disponibles en Python.

Finalmente, y ante la posibilidad de integración de estas tecnologías, podemos plantearnos la pregunta, ¿cuál es el modelo más adecuado y eficiente para integrar

CNN y las API de Google Earth Engine o Sentinel-Hub, con el objetivo de estimar las áreas de deslizamientos de tierra, mediante el análisis de imágenes satelitales ópticas?

**PREGUNTAS DE INVESTIGACIÓN**

¿Cuál sería el mejor modelo de red neuronal convolucional que permita aprovechar las bases de datos de imágenes satelitales ópticas existentes en Google Earth Engine/Sentinel Hub para la estimación de áreas de deslizamientos de tierra?

Preguntas especificas:

* ¿Qué estructura y características debe tener la base de datos de entrenamiento y testeo para el modelo de red neuronal convolucional?
* ¿Cómo construir la interconexión entre tecnologías API Google Earth Engine/Sentinel Hub y Framework TensorFlow, que permitan la extracción, adecuación, compilado y análisis de las imágenes ópticas satelitales?
* ¿Cuál es el modelo de red neuronal convolucional más adecuado para la extracción de información en las imágenes ópticas, revisando tipos de convolución, función de activación de cada neurona, función de optimización y diseño de las capas?
* ¿Cuáles son las características espectrales, tipo de deslizamiento o condiciones geomorfológicas que permiten una mejor estimación de los deslizamientos de tierra para el modelo de red neuronal planteado?

**PLANOS PARA LA CONSTRUCCIÓN DEL CAPITULO 1 / PROPUESTA DE TESIS**

1. **Resumen del proyecto**

En esta sección, debemos enfocarnos en describir las tecnologías y herramientas utilizadas para la detección de deslizamientos, haciendo un breve énfasis en la descripción de la metodología utilizada: lenguaje de programación y librerías. Adicional a dar un acercamiento a la zona de estudio y bases de datos utilizados, resultados esperados y aportes al conocimiento de la Geomática.

1. **Justificación**

En este apartado, enfocarse en justificar el problema de investigación: Primeramente, un contexto general de las problemáticas asociadas a los movimientos de remoción en masa, el aumento de estos procesos por efectos del cambio climático, la importancia de los estudios relacionados a estos procesos geomorfológicos y la situación actual del país en términos de información disponible a la fecha.

En un segundo párrafo comenzare a explicar, como desde el campo de la geomática han surgido nuevas técnicas asociadas a la clasificación de superficies, mediante imágenes ópticas multiespectrales, Y como se han abordado problemas de clasificación con algoritmos de aprendizaje de máquina.

Finalmente, explorar las posibilidades de usar una técnica de aprendizaje profundo (redes neuronales convolucionales) en asociación con repositorios de imágenes satelitales como Google Earth Engine o Sentinel Hub, para ofrecer un método de estimación de áreas de deslizamientos. En esta justificación quiero ir explicando cómo en estudios recientes del año 2020 y 2021, los investigadores que han trabajado con GEE se han enfocado sus análisis al uso de técnicas de ML tradicional como RF, SVM o algoritmos bayesianos. Y los que han usado CNN para detección de deslizamientos, no se han apoyado en el uso de las APIs.

(Nota: El detalle de que no existan muchos estudios previos donde conecten estas dos tecnologías, me plantea dos grandes dudas: 1. ¿Sera que hay muchos problemas en la unificación de esas tecnologías, como para que nadie los haya utilizado?, 2. ¿Será computacionalmente viable una integración de las APIs y TensorFlow?)

1. **Planteamiento del problema**

En el planteamiento del problema debe enfocarse en las características del inventario de deslizamientos en Colombia, cuales son las fuentes de estos inventarios y facilidades o dificultades en su obtención. Dado que en capítulos previos se busco justificar la dificultad en la obtención de áreas/polígonos de movimientos de remoción en masa, debo expresar en una tabla las fuentes disponibles de las entidades gubernamentales y cuál es su proporción entre las existentes en el SIMMA y las áreas que han sido detectadas y digitalizadas.

También debo describir los procedimientos de identificación de deslizamientos aplicados actualmente en el país, buscando todas las metodologías o guías existentes. Esto sería para justificar posteriormente el aporte que la detección con CNN pueda aportar al inventariado de deslizamientos actual (importante hacer énfasis en el aporte y no en el reemplazo, los modelos de aprendizaje de máquina no implican un reemplazo de los modelos tradicionales, si no un apoyo o complemento para estos mismos).

En el tercer párrafo, nos enfocaríamos en buscar referencias que ilustren el aporte del aprendizaje profundo en la detección de objetos en imágenes satelitales, tanto coberturas del terreno, como cuerpos de agua y otros. Y como estas tecnologías aportan una primer acercamiento económico y eficiente en tiempo a la recolección y/o recuperación de información de superficies afectadas.

En el cuarto párrafo, se deberá entonces hablar del alcance de estos métodos, las posibilidades que estos tienen para la resolución de la necesidad identificada, así como características de las herramientas como pueden ser, la dependencia de la resolución de las imágenes multiespectrales, perdida de información por presencia de nubes y la necesidad de grandes conjuntos de datos ya categorizados para su entrenamiento.

En el último párrafo, será construir argumentalmente el problema de investigación, aclarando desde que rama del conocimiento se le quiere dar solución y una breve introducción a los objetivos de investigación.

1. **Marco teórico y estado del arte**

Desde el marco teórico, será importante aclarar 4 componentes:

* Primero: Movimientos de remoción en masa y técnicas de detección.
* Segundo: Percepción remota
* Tercero Plataformas GEE/Sentinel Hub
* Cuarto: Redes neuronales Convolucionales
  1. **Movimientos de remoción en masa:**

Se dará una definición de los movimientos de remoción en masa, incluyendo también una clasificación los referentes bibliográficos. Una vez esa introducción de la clasificación, hablaremos un poco más a fondo de las metodologías aplicadas en el territorio colombiano para la detección de deslizamientos, para pasar a técnicas de detección más relacionadas con la percepción remota y el aprendizaje de máquina, para finalmente describir el “Sistema de Información de Movimientos en Masa” (SIMMA) y la información disponible a nivel público.

* 1. **Percepción remota**

Hablaremos de los principales fundamentos de la teledetección, describiendo los instrumentos, tipos de sensores, características y principales constelaciones de satélites con los cuales contamos actualmente.

* 1. **Plataformas GEE/Sentinel Hub**

En este apartado nos enfocaremos en hablar de la creación y evolución de las plataformas Google Earth Engine y Sentinel Hub, su evolución, las herramientas de las cuales disponen, funcionalidades y especialmente ejemplos puntuales de su aplicación sobre detección de coberturas y deslizamientos de tierras.

* 1. **Redes neuronales convolucionales (CNN)**

En este párrafo nos enfocaremos en explicar de manera concisa y resumida, los fundamentos detrás de las CNN, enfocando en explicar algunas características como pueden ser las funciones de activación, algoritmos de optimización, construcción de capas y las convoluciones. También incluir ejemplos muy específicos de como las CNN tienen el potencial de extraer características de las imágenes como pueden ser bordes, contornos, formas, entre otras. Al final de esta sección se mencionarán ejemplos de aplicaciones en campos como la detección de coberturas como deslizamientos de tierras.

1. **Objetivos**
   1. **Objetivo general**

--- Sin comentarios ---

* 1. **Objetivos específicos**

--- Sin comentarios ---

1. **Metodología propuesta**

Debemos buscar entre la bibliografía una metodología de trabajo aplicable a CNN, actualmente había considerado la metodología CRISP-DM, que fue el estándar utilizado durante los diplomados de ciencia de datos de la universidad nacional, sin embargo, no se descartan otros estándares utilizados por investigadores.

La metodología CRISP-DM cuenta con las siguientes Fases:

**Fase I. Definición de necesidades del usuario:** Esta fase inicial se enfoca en la comprensión de los objetivos de proyecto. Definido el objetivo se establece el tipo de tratamiento y minería de los datos, así como la definición de un plan preliminar diseñado para alcanzar los objetivos.

**Fase II. Estudio y comprensión de los datos:** La fase comienza por la recolección de los datos. Lo cual incluye la definición de la base de datos y cual tipo es más conveniente para el problema (ya sea relacional o no relacional). También contiene la familiarización de los datos, identificar los problemas de calidad y describir características preliminares sobre los datos.

**Fase III. Adecuación y preparación de los datos (selección de características):** La preparación de datos cubre todas las actividades necesarias para construir el conjunto final de datos. Esto incluye la selección, adecuación de las imágenes, transformación y la limpieza de datos para el algoritmo de aprendizaje.

**Fase IV. Modelado:** En esta fase, se seleccionan y aplican los modelos pertinentes al problema, en este segmento se deben calibrar parámetros para obtener resultados óptimos. Será en este segmento que se buscaran las distintas combinaciones de capas, activadores, optimización y convoluciones, para conseguir mejores resultados del modelo. Nos enfocaremos en recuperar de la bibliografía los mejores modelos tanto de cobertura de suelos, como de aplicaciones en deslizamientos. Y estudiar a fondo las implicaciones de cada variable seleccionada en cada modelo.

**Fase V. Evaluación (obtención de resultados):** Para esta etapa se han construido uno o varios modelos, en la fase V se procederán a evaluar los resultados mediante métricas de desempeño del modelo (usualmente métricas de precisión, pero a posterior tendremos que buscar más métricas que puedan darnos mejores interpretaciones).

**Fase VI. Despliegue:** Una vez probado el modelo, parte del planteamiento del proyecto de investigación es dar una solución parcial a la necesidad de crear una base de datos de superficies de movimientos de remoción en masa. Por lo cual procedemos a estimar las áreas de asociadas a deslizamientos del inventario SIMMA, como producto final del trabajo de investigación. (el producto puede ser reconsiderado con la evolución del documento).

Todo el procesamiento y análisis de los datos será efectuado mediante el lenguaje de programación Python, aprovechando librerías como geopandas, Sklearn, TensorFlow, Rasterio y las API de Google Earth Engine y/o Sentinel Hub. El uso de una u otra API esta abierto a discusión dependiendo cual de las dos ofrece mayores facilidades para la manipulación de los repositorios de imágenes.

* 1. **Materiales:**

Nos apoyaremos en 4 componentes principales:

* **Bases de datos de entrenamiento:** Se plantea la utilización de insumos de estudios para la zonificación de deslizamientos que disponga el SGC, IGAC o CAR. En caso de no disponerse datos de deslizamientos de las entidades anteriormente nombradas, se tiene por plan opcional buscar Bases de datos de otros países de los cuales pueda disponerse información gratuitamente. En caso de disponerse varias fuentes de información, se buscará adecuar lo mayor cantidad posible de muestras para una mayor cantidad de datos de entrenamiento.  
    
  En las muestras de entrenamiento, se espera contar con un agrupado de las imágenes ópticas y una capa tipo vector o ráster, con la clasificación de zonas con deslizamientos. En este apartado se explicaran su origen, características, resolución y ubicación.
* **Inventario de deslizamientos SIMMA:**

El inventario de deslizamientos, al ser una capa de datos vectoriales tipo Punto, nos servirá como indicador espacial de las imágenes ópticas sobre la cual debemos predecir las áreas de deslizamientos, una vez entrenado el modelo de la red neuronal. Es decir: una vez entrenado el modelo CNN, buscaremos las imágenes ópticas asociadas a un evento de deslizamiento (un punto en el inventario SIMMA), sobre ellas aplicaremos el modelo con finalidad de que se detecten las zonas con movimientos de remoción en masa.

Nos servirá entonces, como indicadores de donde debemos buscar los deslizamientos, para su estimación. En este segmento explicaremos las características del inventario y como va a ser su uso durante el proyecto.

* **Repositorios de constelaciones API GEE o API Sentinel Hub:**

Es sobre este repositorio que buscaremos los puntos de predicción, si bien el entrenamiento se hará sobre la base de datos de entrenamiento, las predicciones se construirán sobre los datos recuperados de las APIs. En este párrafo adicionalmente se comentará las características de las constelaciones disponibles en cada API: resoluciones temporales, espaciales, espectrales, entre otros.

* 1. **Métodos**

Hablaremos a fondo de los distintos modelos de CNN recuperados de la bibliografía, y sobre los cuales evaluaremos para nuestra aplicación.

También podemos considerar construir nuestro propio modelo, desviándose parcialmente de lo recuperado en los artículos y enfocándose en buscar las variables optimas (convoluciones, función de activación, algoritmo optimización, distribución de capas, entre otros).

Enfoque pendiente de seleccionar y profundizar.

1. **Resultados Esperados**

Basándome en investigaciones similares, podemos plantear un porcentaje de áreas estimadas respecto al total de deslizamientos inventariados (teniendo en cuenta, que un gran porcentaje de los deslizamientos reportados no debe tener una superficie mayor a lo detectable por los sensores remotos).

Dentro de los resultados esperados, también se comprobará el desempeño del modelo, recuperando así métricas de precisión. Podemos probar distintos tipos de modelos de CNN y determinar así cual tiene mejor resultado para los datos evaluados.

1. **Cronograma de Actividades**

--- Pendiente de completar---