Aplicaciones de los embeddings satelitales de Google AlphaEarth Foundations en el análisis geoespacial: clasificación de uso del suelo, modelado de distribución de especies y diseño de muestreo

José Ramón Martínez Batlle

Universidad Autónoma de Santo Domingo (UASD)

Actualizado: 2025-11-19

¿Qué son los embeddings de Google AlphaEarth Foundations (AEF)?

- Los embeddings de **Google AlphaEarth Foundations (AEF)** condensan información espacio–temporal derivada de series multiespectrales completas (p.ej., Sentinel, Landsat, mapas de cobertura, LiDAR, bioclima).
- Cada píxel es representado por un **vector numérico de alta dimensión** que captura características espectrales, temporales y espaciales.
- Su principal fortaleza es que **permiten realizar tareas de clasificación con muy pocos datos de entrenamiento**, evitando el entrenamiento de modelos de deep learning desde cero.

AlphaEarth Foundations: An embedding field model for accurate and efficient global mapping from sparse label data

Christopher F. Brown[*,1], Michal R. Kazmierski[*,1], Valerie J. Pasquarella[*,2], William J. Rucklidge[2], Masha Samsikova[1], Chenhui Zhang[1], Evan Shelhamer[1], Estefania Lahera[2], Olivia Wiles[1], Simon Ilyushchenko[2], Noel Gorelick[2], Lihui Lydia Zhang[1], Sophia Alj[1], Emily Schechter[2], Sean Askay[2], Oliver Guinan[2], Rebecca Moore[2], Alexis Boukouvalas[1] and Pushmeet Kohli[1]
[*]Equal contributions, [1]Google DeepMind, [2]Google

Unprecedented volumes of Earth observation data are continually collected around the world, but high-quality labels remain scarce given the effort required to make physical measurements and observations. This has led to considerable investment in bespoke modeling efforts translating sparse labels into maps. Here we introduce AlphaEarth Foundations, an embedding field model yielding a highly general, geospatial representation that assimilates spatial, temporal, and measurement contexts across multiple sources, enabling accurate and efficient production of maps and monitoring systems from local to global scales. The embeddings generated by AlphaEarth Foundations are the only to consistently outperform all previous featurization approaches tested on a diverse set of mapping evaluations without re-training. We will release a dataset of global, annual, analysis-ready embedding field layers from 2017 through 2024.

Introduction

Management of global food supplies, public health, and disaster response all start from maps that geographically anchor questions like "which forests pose an unacceptable wildfire risk?" or "where are soybeans grown?". The launch of the first Landsat satellite in 1972 marked the dawn of an era where spaceborne monitoring could serve the interests of global environmental policy-making and provide critical insights into our changing planet (Cohen and Goward, 2004). Over the following decades Earth observation (EO) data became widely available, and streams from both historic and modern EO instruments are now routinely used to create maps that answer questions about the past, present, and future of Earth's ecosystems and climate (Wulder et al., 2022). Nonetheless, advancements in deriving planetary-scale insights from petabytes of satellite imagery and other environmental datasets remain hamstrung by the relative scarcity of ground-based measurements and annotations, and a new problem: the overwhelming volume of geospatial data (Tuia et al., 2024). In this work, we introduce a foundational geospatial embedding model that solves fundamental challenges in the institution of mapping through the generation of a universal feature space. The features produced by our model consistently achieve top performance in all application domains tested when compared to other general and even domain specific approaches (Figure 1A). This marks a shift from the previous state-of-the-art for which no single approach was dominant.

From sparse labels to maps

High-quality maps depend on high-quality labeled data, yet when working at global scales, a balance must be struck between measurement precision and spatial coverage. Many global mapping efforts focus on individual ecosystems like forests (Hansen et al., 2013), water (Pekel et al., 2016), tidal wetlands (Murray et al., 2022a) or other broad legends, e.g., (Brown et al., 2022; Zanaga et al., 2022). This simplifies the label collection process, allowing trained interpreters to collect larger volumes at scale at the expense of descriptive power for certain use cases. In the cases where high-quality annotations and/or field

AlphaEarth Foundations: An embedding field model for accurate and efficient global mapping from sparse label data

Christopher F. Brown[*,1], Michal R. Kazmierski[*,1], Valerie J. Pasquarella[*,2], William J. Rucklidge[2], Masha Samsikova[1], Chenhui Zhang[1], Evan Shelhamer[1], Estefania Lahera[2], Olivia Wiles[1], Simon Ilyushchenko[2], Noel Gorelick[2], Lihui Lydia Zhang[1], Sophia Alj[1], Emily Schechter[2], Sean Askay[2], Oliver Guinan[2], Rebecca Moore[2], Alexis Boukouvalas[1] and Pushmeet Kohli[1]
[*]Equal contributions, [1]Google DeepMind, [2]Google

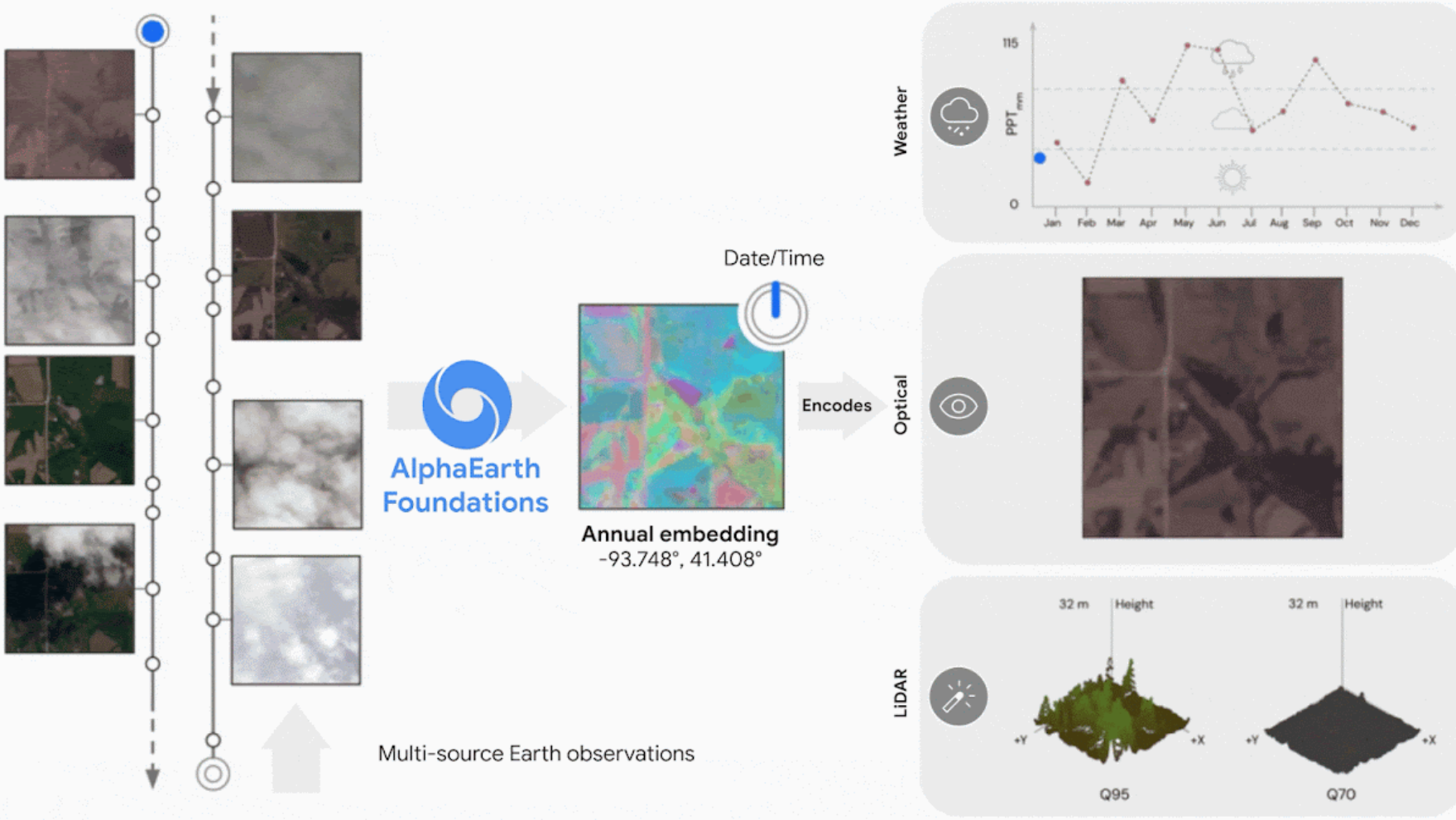
Unprecedented volumes of Earth observation data are continually collected around the world, but high-quality labels remain scarce given the effort required to make physical measurements and observations. This has led to considerable investment in bespoke modeling efforts translating sparse labels into maps. Here we introduce AlphaEarth Foundations, an embedding field model yielding a highly general, geospatial representation that assimilates spatial, temporal, and measurement contexts across multiple sources, enabling accurate and efficient production of maps and monitoring systems from local to global scales. The embeddings generated by AlphaEarth Foundations are the only to consistently outperform all previous featurization approaches tested on a diverse set of mapping evaluations without re-training. We will release a dataset of global, annual, analysis-ready embedding field layers from 2017 through 2024.

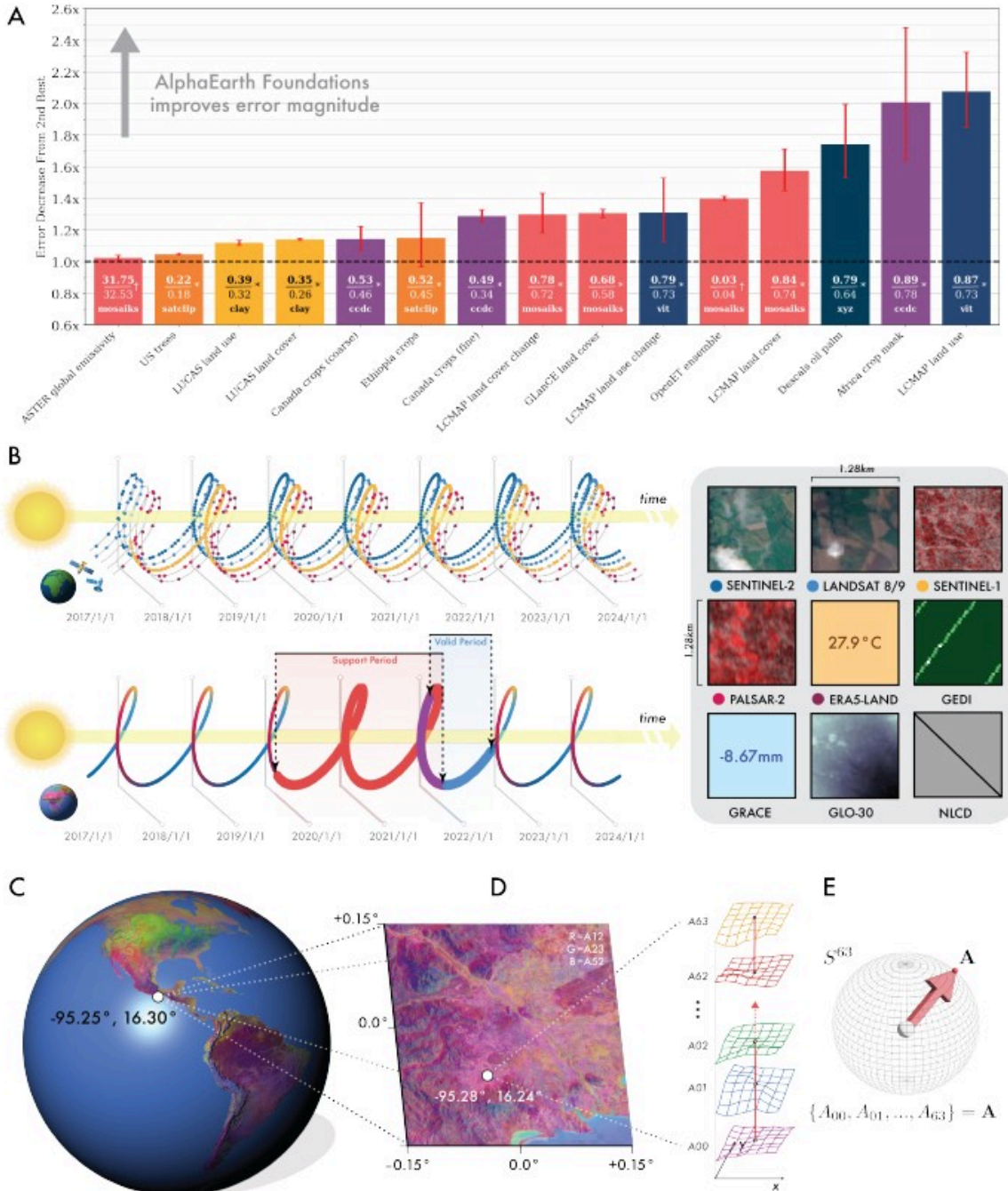
Introduction

Management of global food supplies, public health, and disaster response all start from maps that geographically anchor questions like "which forests pose an unacceptable wildfire risk?" or "where are soybeans grown?". The launch of the first Landsat satellite in 1972 marked the dawn of an era where spaceborne monitoring could serve the interests of global environmental policy-making and provide critical insights into our changing planet (Cohen and Goward, 2004). Over the following decades Earth observation (EO) data became widely available, and streams from both historic and modern EO instruments are now routinely used to create maps that answer questions about the past, present, and future of Earth's ecosystems and climate (Wulder et al., 2022). Nonetheless, advancements in deriving planetary-scale insights from petabytes of satellite imagery and other environmental datasets remain hamstrung by the relative scarcity of ground-based measurements and annotations, and a new problem: the overwhelming volume of geospatial data (Tuia et al., 2024). In this work, we introduce a foundational geospatial embedding model that solves fundamental challenges in the institution of mapping through the generation of a universal feature space. The features produced by our model consistently achieve top performance in all application domains tested when compared to other general and even domain specific approaches (Figure 1A). This marks a shift from the previous state-of-the-art for which no single approach was dominant.

From sparse labels to maps

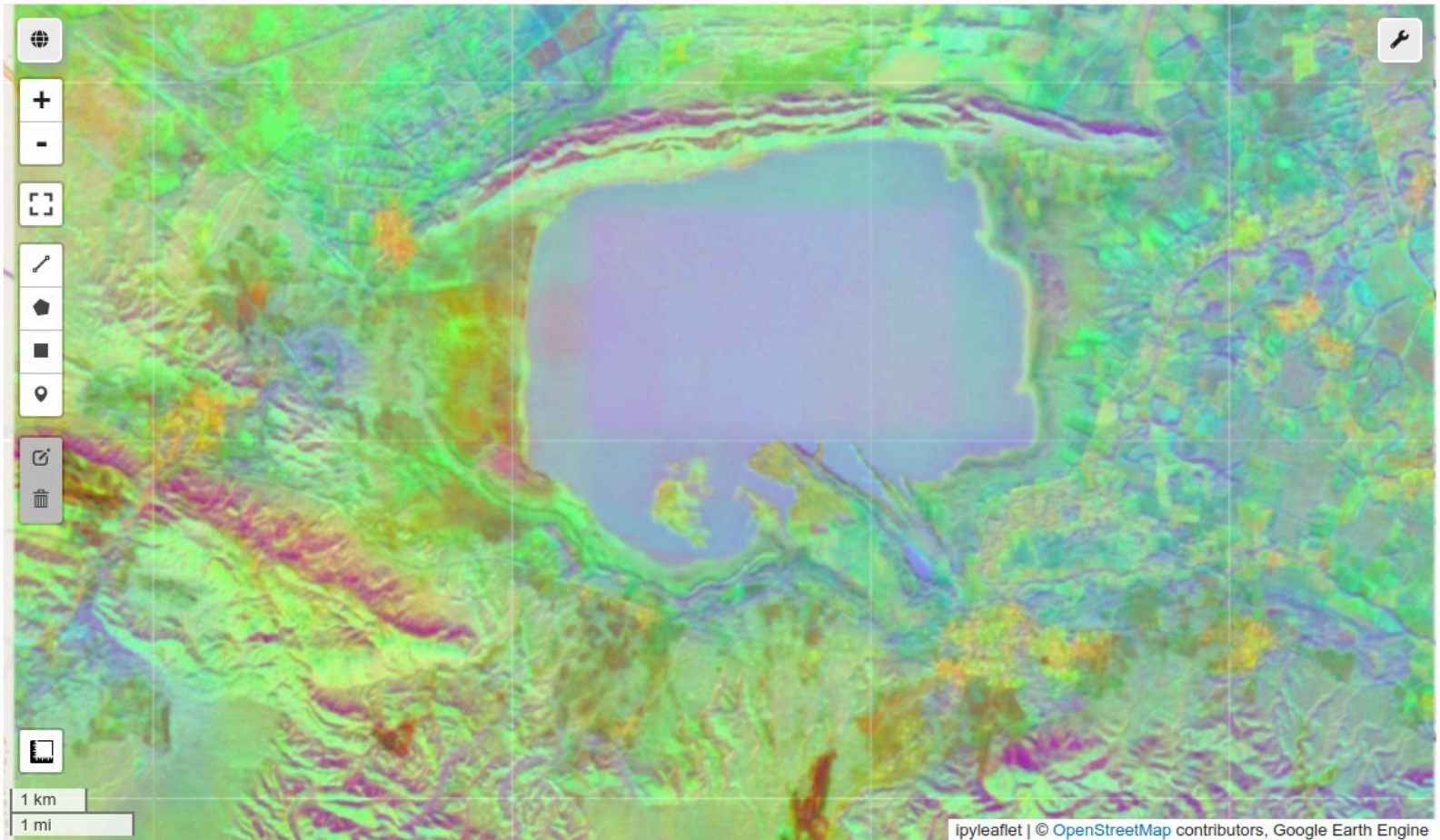
High-quality maps depend on high-quality labeled data, yet when working at global scales, a balance must be struck between measurement precision and spatial coverage. Many global mapping efforts focus on individual ecosystems like forests (Hansen et al., 2013), water (Pekel et al., 2016), tidal wetlands (Murray et al., 2022a) or other broad legends, e.g., (Brown et al., 2022; Zanaga et al., 2022). This simplifies the label collection process, allowing trained interpreters to collect larger volumes at scale at the expense of descriptive power for certain use cases. In the cases where high-quality annotations and/or field


Multi-source Earth observations
AlphaEarth Foundations
Date/Time
Annual embedding
-93.748°, 41.408°
Encodes
Weather
115
PPT mm
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Optical
LiDAR
32 m
Height
+Y
+X
Q95
32 m
Height
+Y
+X
Q70


A
AlphaEarth Foundations
improves error magnitude
Error Decrease From 2nd Best
2.6x
2.4x
2.2x
2.0x
1.8x
1.6x
1.4x
1.2x
1.0x
0.8x
0.6x
31.75 / 32.53 mosaiks
0.22 / 0.18 satclip
0.39 / 0.32 clay
0.35 / 0.26 clay
0.53 / 0.46 ccdc
0.52 / 0.45 satclip
0.49 / 0.34 ccdc
0.78 / 0.72 mosaiks
0.68 / 0.58 mosaiks
0.79 / 0.73 vit
0.03 / 0.04 mosaiks
0.84 / 0.74 mosaiks
0.79 / 0.64 xyz
0.89 / 0.78 ccdc
0.87 / 0.73 vit
ASTER global emissivity
US trees
LUCAS land use
LUCAS land cover
Canada crops (coarse)
Ethiopia crops
Canada crops (fine)
LCMAP land cover change
GLanCE land cover
LCMAP land use change
OpenET ensemble
LCMAP land cover
Desicals oil palm
Africa crop mask
LCMAP land use
B
time
2017/1/1
2018/1/1
2019/1/1
2020/1/1
2021/1/1
2022/1/1
2023/1/1
2024/1/1
Support Period
Valid Period
1.28km
SENTINEL-2
LANDSAT 8/9
SENTINEL-1
PALSAR-2
ERA5-LAND
GEDI
27.9 °C
-8.67mm
GRACE
GLO-30
NLCD
C
-95.25°, 16.30°
D
+0.15°
0.0°
-0.15°
R=A12
G=A23
B=A52
-95.28°, 16.24°
A63
A62
A02
A01
A00
E
S^{63}
A
$\{A_{00}, A_{01}, \ldots, A_{63}\} = \mathbf{A}$

Visualización “tal cual” de los embeddings (composición RGB)

```
embedding_rgb_bands = ["A00", "A21", "A42"]
emb_vis_params = {
    "bands": embedding_rgb_bands,
    "min": -0.25,
    "max":  0.25,
}
m2 = geemap.Map(center=[(bbox[1] + bbox[3]) / 2, (bbox[0] + bbox[2]) / 2], zoom=11)
# Embeddings en RGB pseudo-color
m2.addLayer(emb_image, emb_vis_params, "Embeddings AEF (RGB)")
# AOI
m2.addLayer(aoi, {"color": "black"}, "AOI", False)
m2
```



Objetivos

- Objetivo general: evaluar la utilidad práctica de los **embeddings de AlphaEarth Foundations** para apoyar diferentes procesos de análisis espacial en contextos ecológicos y territoriales de la República Dominicana.

- Objetivos específicos:

    - Probar su desempeño en la modelación de la **distribución espacial de especies** empleando registros de presencia y técnicas de aprendizaje supervisado.
    - Evaluar la capacidad para **discriminar coberturas** en zonas agrícolas, forestales y urbanas.
    - Analizar su aplicación en el **diseño de muestreo estratificado**, utilizando la similitud entre embeddings para optimizar la representatividad espacial y ambiental

Objetivo 1

Probar desempeño de embeddings de AEF en la modelación de la **distribución espacial de especies** empleando registros de presencia y técnicas de aprendizaje supervisado.

Colaboración para mis estudiantes del semestre 2023-02, asignatura biogeografía: Adrián Montás, Ángel González, Arisleydi De la Cruz, Bryan Ramos, Claribel Ramírez, Manuel Reyes, Ramona Muñoz, Saderis Carmona, Yenny Santana

Platform Solutions Resources Open Source Enterprise Pricing

biogeografia-202302 / **manuscrito** Public template

Code Issues Pull requests Actions Projects Security Insights

main 1 Branch 0 Tags Go to file Code

geofis ADD manuscript assignment Saderis data/report 59ec7a6 · 2 years ago 28 Commits

R	UPDATE advancing results template	2 years ago
data	ADD assignment 1 sample collection	2 years ago
fuentes/manuscrito	ADD manuscript assignment Saderis data/report	2 years ago
odk	UPDATE symlinks deleted	2 years ago
.gitignore	UPDATE .gitignore	2 years ago
LICENSE	Initial commit	2 years ago
README.md	UPDATE README	2 years ago

README GPL-3.0 license

Asignaciones de manuscrito

Asignación de manuscrito 1. Diseño de muestreo y colecta de datos de campo

Asignación de manuscrito 2. Técnicas de procesamiento y analíticas (subsección de la Metodología)

Asignación de manuscrito 3. Introducción

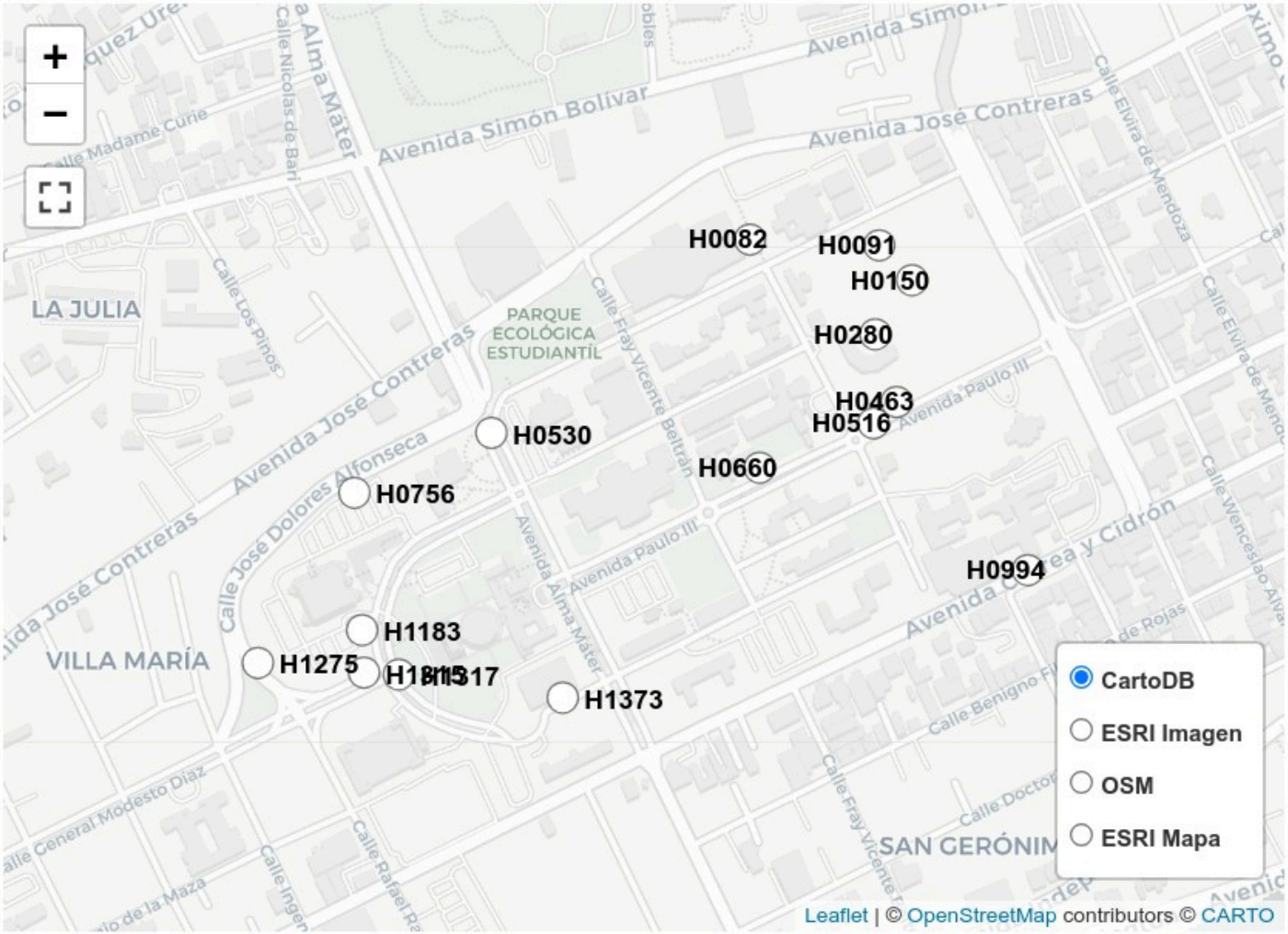
Asignación de manuscrito 4. Resultados

Estudiante	Ruta de informe
Adrian-Rafael-Diplan-Montas	https://biogeografia-202302.github.io/manuscrito/fuentes/manuscrito/informe-resultados-Adrian-Rafael-Diplan-Montas.html
Angel-Rolando-Gomez-Gonzalez	https://biogeografia-202302.github.io/manuscrito/fuentes/manuscrito/informe-resultados-Angel-Rolando-Gomez-Gonzalez.html
Arisleydi-Mejia-De-La-Cruz	https://biogeografia-202302.github.io/manuscrito/fuentes/manuscrito/informe-resultados-Arisleydi-Mejia-De-La-Cruz.html
Bryan-Josue-Funez-Ramos	https://biogeografia-202302.github.io/manuscrito/fuentes/manuscrito/informe-resultados-Bryan-Josue-Funez-Ramos.html
Claribel-Reyes-Ramirez	https://biogeografia-202302.github.io/manuscrito/fuentes/manuscrito/informe-resultados-Claribel-Reyes-Ramirez.html
Manuel-Enrique-Urena-Reyes	https://biogeografia-202302.github.io/manuscrito/fuentes/manuscrito/informe-resultados-Manuel-Enrique-Urena-Reyes.html
Ramona-Geraldo-Munoz	https://biogeografia-202302.github.io/manuscrito/fuentes/manuscrito/informe-resultados-Ramona-Geraldo-Munoz.html
Saderis-Carmona-Marte	https://biogeografia-202302.github.io/manuscrito/fuentes/manuscrito/informe-resultados-Saderis-Carmona-Marte.html
Yenny-Mabel-Santana	https://biogeografia-202302.github.io/manuscrito/fuentes/manuscrito/informe-resultados-Yenny-Mabel-Santana.html

2.2 Mapa de mis puntos colectados

Nota: sólo figuran los puntos con coordenadas. El total de formularios podría ser mayor al número de puntos mostrado en el mapa. Para un mapa comprensivo, ver más adelante el mapa bajo el texto “Mapa de hexágonos visitados”.

Code



2.3 Fotos

1696783022562.jpg

1696772480069.jpg

Code

Datos y análisis para resultados de Adrian-Rafael-Diplan-Montas

Elaborado por: José-Ramón Martínez-Batlle (jmartinez19@uasd.edu.do)
Facultad de Ciencias, Universidad Autónoma de Santo Domingo (UASD)
Santo Domingo, República Dominicana

1 Carga de paquetes

Son muchos los paquetes empleados en estos análisis. Puedes consultar en el ChatGPT qué hace cada uno. Considera un aspecto también importante: algunas funciones escritas por mí se cargan con `source_url` y `source`; dentro de algunas de dichas funciones, también se cargan paquetes adicionales.

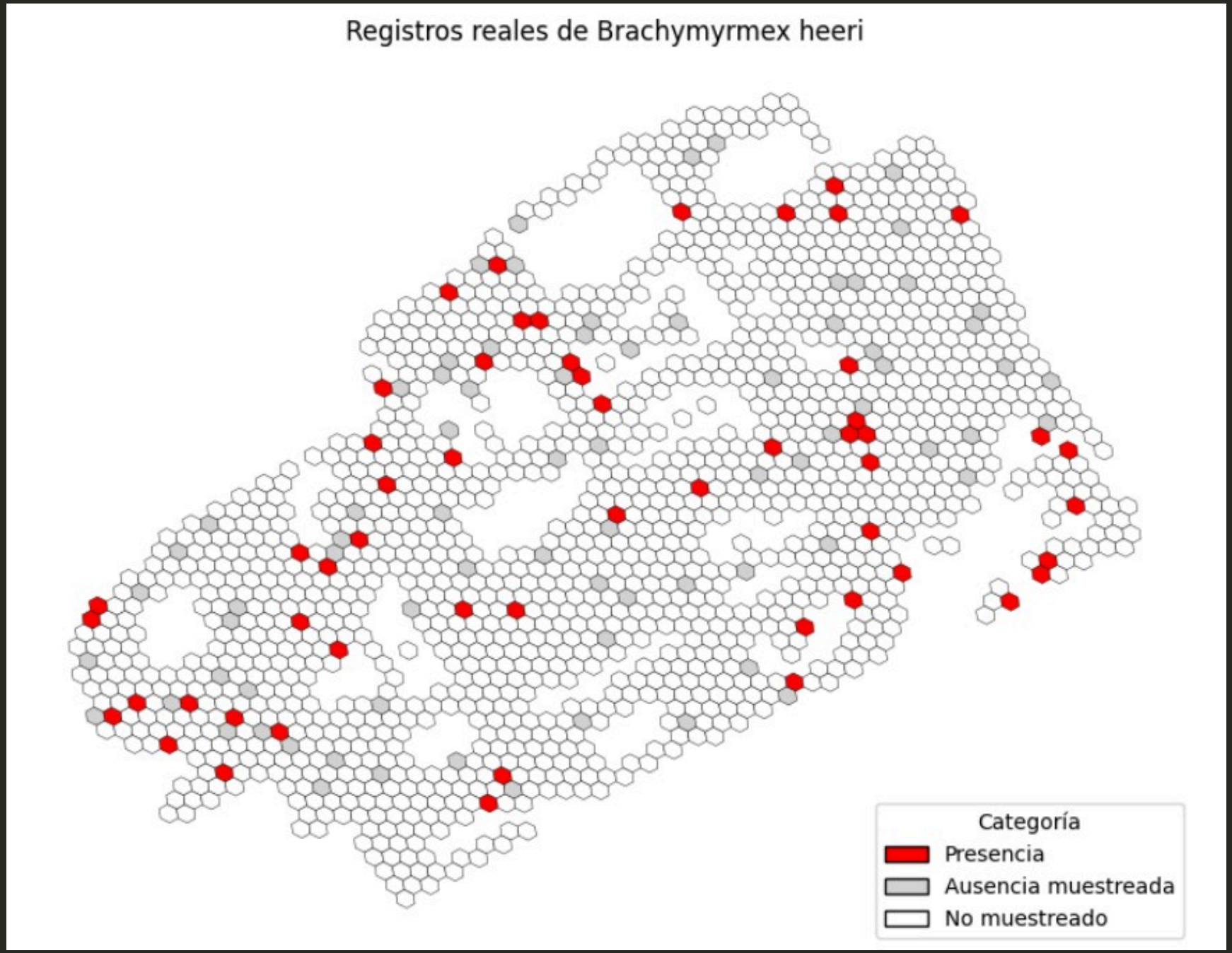
Code

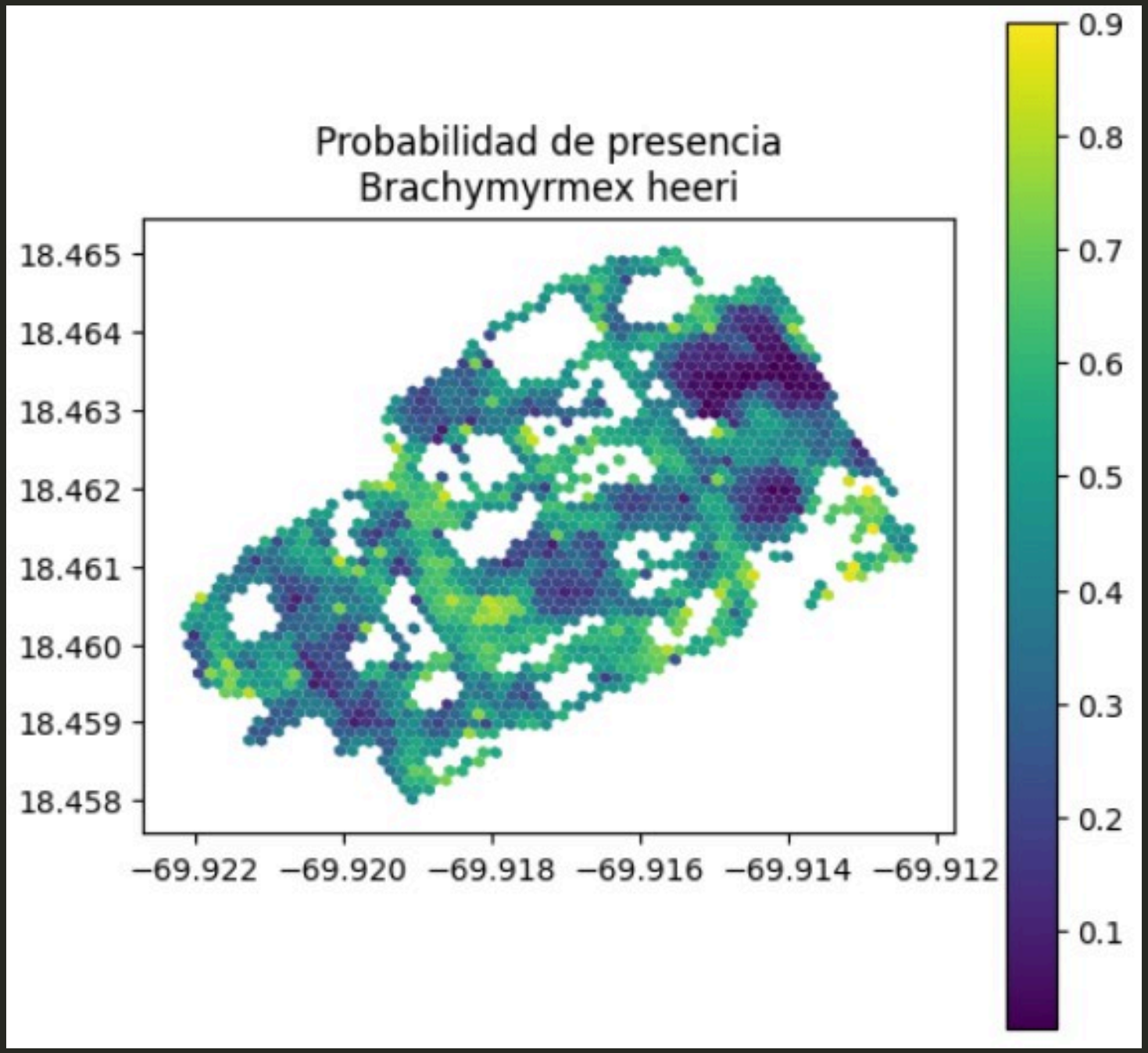
2 Leer y preparar datos

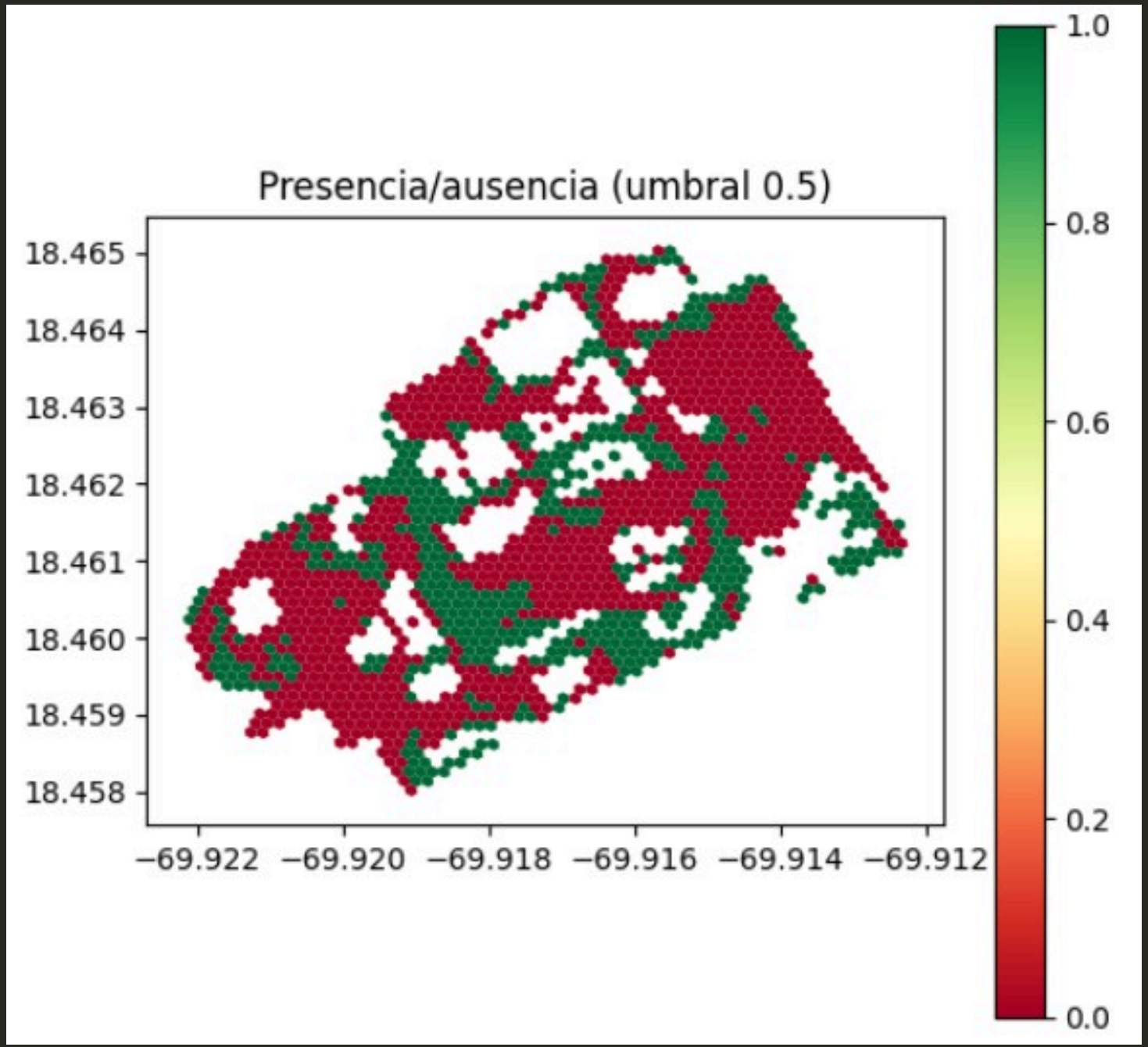
2.1 Formularios de Adrian Rafael Diplan Montas

Code


Parque Iberoamérica
Unidad Médica
Avenida Simón Bolívar
Avenida Pedro Henríquez Ureña
Calle Madame Curie
Calle Nicolás de Bari
Avenida Alma Máter
Calle Ricardo Robles
La Julia
Calle Los Pinos
Calle Cantera
Calle Santo Domingo
Avenida Sarasota
Calle Doctor Núñez y Domínguez
Calle Roma
Villa María
Avenida Abraham Lincoln
Calle General Moden
Calle Juan Tomás Díaz
Calle Ingeniero Huáscar Tejeda
Calle Rafael Ravelo
Avenida Independencia
Avenida George Washington
Calle Fray Vicente Beltrán
Calle Doctor Piñeiro
San Gerónimo
Calle Benigno Filomeno de Rojas
Calle Wenceslao Álvarez
Centro Medico Dominico Cubano
Malecon Center
parque
RD-2
Avenida Correa y Cidrón
Avenida Santiago
Avenida Máximo Gómez
Joaquín Balaguer
Calle Elvira de Mendoza
Calle José Desiderio Valverde
Calle Mahatma Gandhi
Calle Pedro Ignacio Espaillat
Calle Arístides Fiallo Cabral
Ensanche Inde
Calle 8
Calle 3
Francisco
200 m
500 ft
ipyleaflet | © OpenStreetMap contributors, Google Earth Engine


Registros reales de Brachymyrmex heeri
Categoría
Presencia
Ausencia muestreada
No muestreado


Probabilidad de presencia
Brachymyrmex heeri
18.465
18.464
18.463
18.462
18.461
18.460
18.459
18.458
−69.922
−69.920
−69.918
−69.916
−69.914
−69.912
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1


Presencia/ausencia (umbral 0.5)
18.465
18.464
18.463
18.462
18.461
18.460
18.459
18.458
−69.922
−69.920
−69.918
−69.916
−69.914
−69.912
1.0
0.8
0.6
0.4
0.2
0.0

Modelo

[29]:
```
results = model_species_distribution(
    gdf_full,
    especie_elegida,
    classifier_name="et",
    threshold=0.5
)
```

```
Presencias de Brachymyrmex heeri: 54
Filas finales para entrenar: 130
Nº de bandas de embeddings usadas: 64
Matriz de confusión (test):
[[16  7]
 [ 8  8]]
ROC AUC (test): 0.569
```

```
Especie: Brachymyrmex heeri
Hexágonos muestreados: 130 (presencias=54, ausencias=76)
Fold 1: Boyce = 0.517 | presencias val = 14
Fold 2: Boyce = 0.429 | presencias val = 9
Fold 3: Boyce = 0.810 | presencias val = 13
Fold 4: Boyce = 0.548 | presencias val = 12
Fold 5: Boyce = 1.000 | presencias val = 6

=== Resumen Boyce (validación cruzada) ===
Especie: Brachymyrmex heeri
Boyce medio = 0.660
Desviación estándar = 0.237
{'species': 'Brachymyrmex heeri',
 'boyce_mean': 0.6604761904761905,
 'boyce_sd': 0.23700703702824313,
 'folds': [{'fold': 1, 'boyce': 0.5166666666666667, 'n_pres_val': 14},
  {'fold': 2, 'boyce': 0.4285714285714286, 'n_pres_val': 9},
  {'fold': 3, 'boyce': 0.8095238095238096, 'n_pres_val': 13},
  {'fold': 4, 'boyce': 0.5476190476190477, 'n_pres_val': 12},
  {'fold': 5, 'boyce': 0.9999999999999999, 'n_pres_val': 6}]}
```

Conclusiones del Objetivo 1

- **Los embeddings AEF reflejan gradientes ambientales finos** dentro del campus, capturando variación espacial relevante.
- **La extracción de valores ambientales por hexágono H3 fue exitosa**, generando una base homogénea para integrar las presencias.
- **Los modelos exploratorios muestran potencial predictivo** de los embeddings como variables para la distribución

Objetivo 2

Evaluar la capacidad de los embeddings satelitales de Google AEF para discriminar coberturas agrícolas, forestales y urbanas

Colaboración para Kénnida Polanco.

Contexto

Para este objetivo se seleccionó un área con presencia simultánea de:

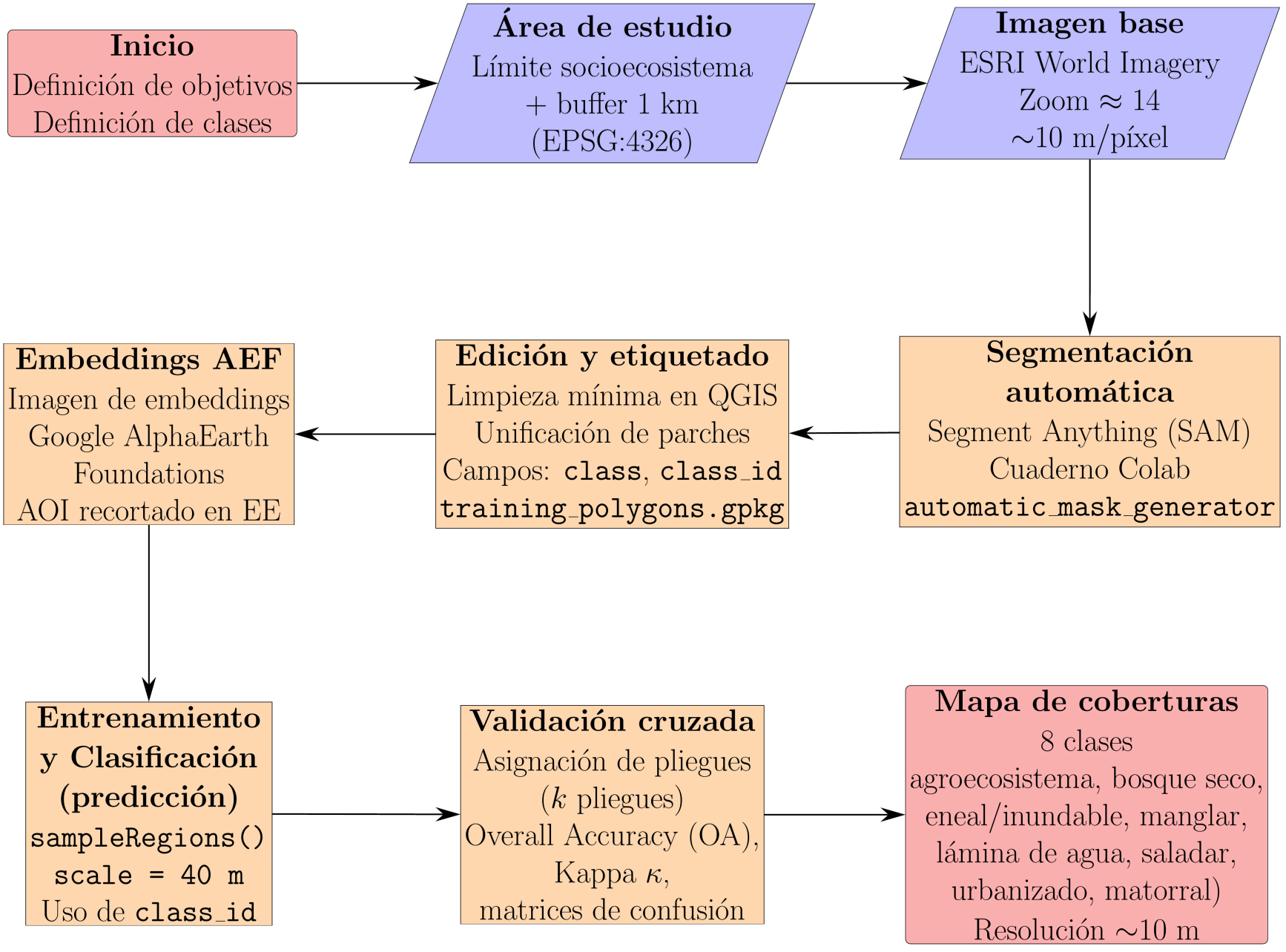
- **Agroecosistemas**
- **Bosque seco** / **matorral** (clases forestales/arbustivas)
- **Humedales** (varias)
- **Áreas urbanizadas**

Flujo metodológico general

A continuación se muestra el resumen metodológico seguido para este objetivo, basado en tres componentes:

1. **Obtención de las máscaras de entrenamiento** mediante SAM (Segment Anything) usando el cuaderno adaptado de *Qiusheng Wu*, paquetes `samgeo`, `geemap`, en Google Colab.
2. **Curación y unificación de clases** en QGIS.
3. **Clasificación supervisada con embeddings AEF** mediante Random Forest y validación cruzada *k-folds* en Google Earth Engine.

Metodología


Inicio
Definición de objetivos
Definición de clases
Área de estudio
Límite socioecosistema
+ buffer 1 km
(EPSG:4326)
Imagen base
ESRI World Imagery
Zoom ≈ 14
∼10 m/píxel
Segmentación automática
Segment Anything (SAM)
Cuaderno Colab
automatic_mask_generator
Edición y etiquetado
Limpieza mínima en QGIS
Unificación de parches
Campos: class, class_id
training_polygons.gpkg
Embeddings AEF
Imagen de embeddings
Google AlphaEarth Foundations
AOI recortado en EE
Entrenamiento y Clasificación (predicción)
sampleRegions()
scale = 40 m
Uso de class_id
Validación cruzada
Asignación de pliegues
(k pliegues)
Overall Accuracy (OA),
Kappa κ,
matrices de confusión
Mapa de coberturas
8 clases
agroecosistema, bosque seco, eneal/inundable, manglar, lámina de agua, saladar, urbanizado, matorral)
Resolución ∼10 m

Qiusheng Wu

giswqs

Follow

Sponsor

Associate Professor at the University of Tennessee, Knoxville | Amazon Scholar

7.2k followers · **178** following

University of Tennessee

Knoxville, TN

01:18 - 1h behind

qwu18@utk.edu

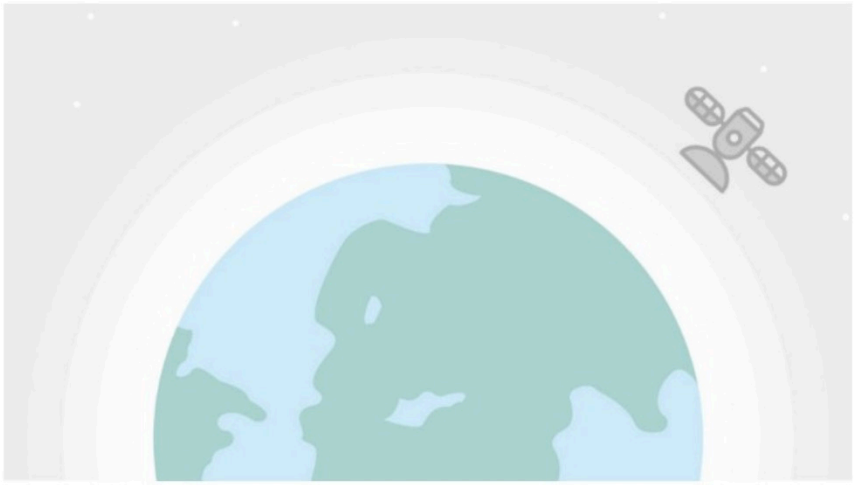
giswqs / README.md

Qiusheng Wu

Dr. Qiusheng Wu is an Associate Professor and the Director of Graduate Studies in the Department of Geography & Sustainability at the University of Tennessee, Knoxville. He also serves as an Amazon Scholar. Dr. Wu's research focuses on geospatial data science and open-source software development, with an emphasis on leveraging big geospatial data and cloud computing to study environmental change, particularly surface water and wetland inundation dynamics. He is the creator of several widely used open-source Python packages, including geemap, leafmap, segment-geospatial, and geoai, which support advanced geospatial analysis and interactive visualization. His open-source work is available at https://github.com/opengeos.

Open-source Projects

- **Linux:** manjaro-linux
- **R packages:** whiteboxR
- **Python packages:** geemap | leafmap | eefolium | geehydro | lidar | whitebox | whiteboxgui | geospatial | pygis | pypackage
- **ArcGIS Toolboxes:** WhiteboxTools-ArcGIS | Depression Analysis Toolbox | Wetland Hydrology Analyst
- **Google Earth Engine:** Awesome-GEE | earthengine-py-notebooks | qgis-earthengine-examples | earthengine-apps



Welcome to Google Earth Engine

1. Obtención de imágenes de alta resolución (ESRI World Imagery)

Se descargó una imagen de referencia del área seleccionada empleando *ESRI World Imagery* con resolución aproximada de **5–10 m**, de manera que fuera coherente con la escala representada en los embeddings AEF.

2. Segmentación automática con Segment Anything (SAM)

La segmentación se realizó usando cuaderno Jupyter. El procesamiento incluyó:

- Uso del modelo **ViT-H** (SAM), adecuado para imágenes de alta resolución pero útil también para imágenes de mediana resolución. Eficiente por su relación sensibilidad–costo computacional.
- Aplicación de `sam.generate()` con `unique=True` para obtener **instancias diferenciadas**.
- Ajuste de hiperparámetros (`points_per_side`, `pred_iou_thresh`, `crop_n_layers`, etc.) para aumentar la densidad de máscaras.

Automatic mask generation options

There are several tunable parameters in automatic mask generation that control how densely points are sampled and what the thresholds are for removing low quality or duplicate masks. Additionally, generation can be automatically run on crops of the image to get improved performance on smaller objects, and post-processing can remove stray pixels and holes. Here is an example configuration that samples more masks:

```
# sam_kwargs = {
#     "points_per_side": 32,
#     "pred_iou_thresh": 0.86,
#     "stability_score_thresh": 0.92,
#     "crop_n_layers": 1,
#     "crop_n_points_downscale_factor": 2,
#     "min_mask_region_area": 100,
# }

# Para zoom 14
sam_kwargs = {
    "points_per_side": 24,          # más que 16, menos que 32 (menos VRAM que 32)
    "pred_iou_thresh": 0.5,         # bajar de 0.86 -> acepta más máscaras
    "stability_score_thresh": 0.8,  # bajar de 0.92 -> idem
    "crop_n_layers": 1,             # mantiene una capa extra de crops
    "crop_n_points_downscale_factor": 2,
    "min_mask_region_area": 25,     # baja de 100 -> deja pasar más regiones pequeñas
}
```

```
sam = SamGeo(
    model_type="vit_h",
    sam_kwargs=sam_kwargs,
)
```

```
sam.generate(image, output="masks2.tif", foreground=False)
```


0
9 km
Mella
Cristóbal
Cabral - Duvergé
RD-46
Angostura
Las Salinas
Independencia
Nerba
Uvilla
Tamayo
La Canoa
Baoruco
Vicente Noble
Jaquimeyes
Palo Alto
El Peñón
Fundación
Refugio de Vida Silvestre Laguna Cabral o Rincón
Refugio de Vida Silvestre Humedales del Bajo Yuna
Pescadería
Habanero
Cabral
El Cachón
Barahona
Aeropuerto Internacional María Montez
RD-533
RD-44
El Batey Central
Santa Cruz de Barahona
N

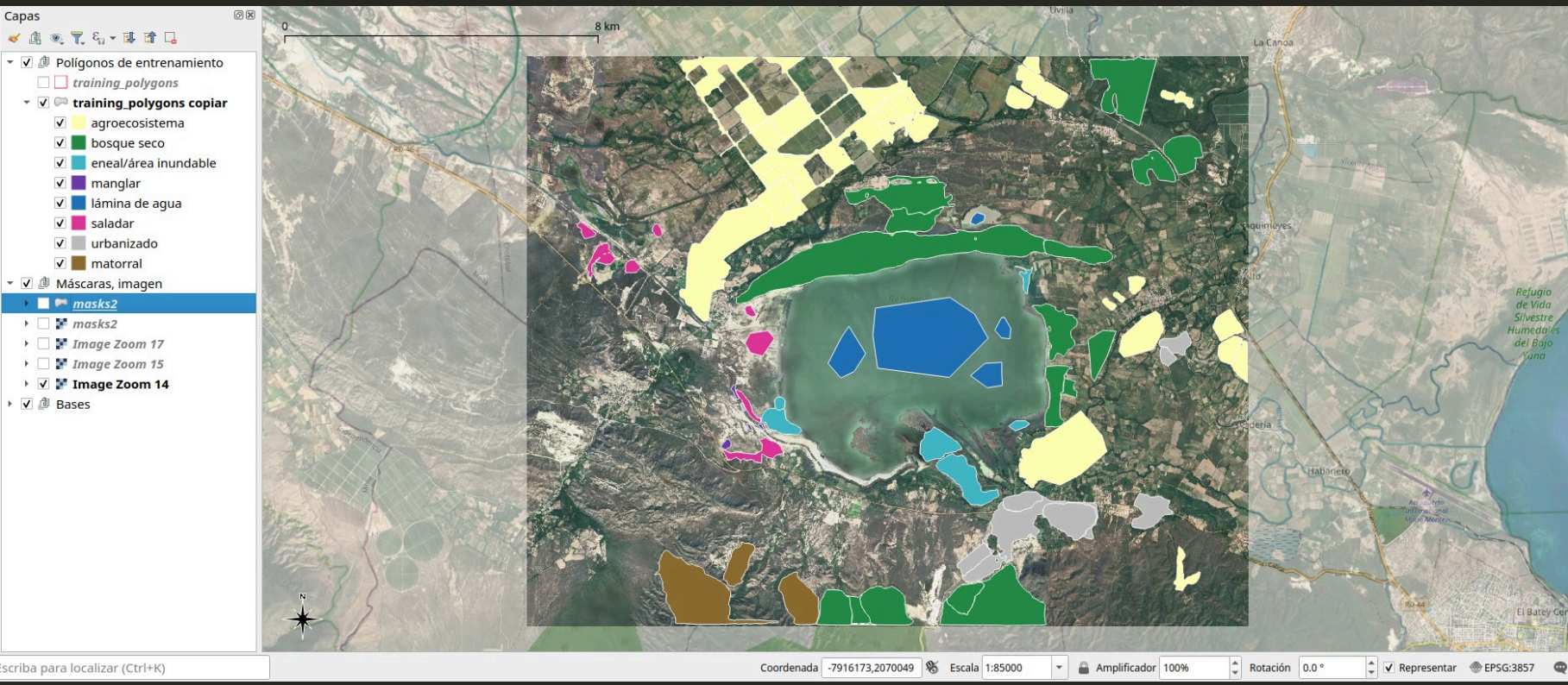
3. Limpieza y edición mínima en QGIS

Las máscaras obtenidas se exportaron a GeoPackage, donde se realizó una edición mínima:

- Corrección de pequeñas islas y polígonos erróneos.
- Eliminación de duplicados.
- Fusión de parches contiguos cuando correspondían a una misma cobertura.

Se añadieron dos campos:

- `class_id` (entero 1–8)
- `class_label` (nombre de la clase)


Capas
Polígonos de entrenamiento
training_polygons
training_polygons copiar
agroecosistema
bosque seco
eneal/área inundable
manglar
lámina de agua
saladar
urbanizado
matorral
Máscaras, imagen
masks2
masks2
Image Zoom 17
Image Zoom 15
Image Zoom 14
Bases
0
8 km
La Canoa
Habanero
Refugio de Vida Silvestre Humedales del Bajo Yuna
Escriba para localizar (Ctrl+K)
Coordenada
-7916173,2070049
Escala
1:85000
Amplificador
100%
Rotación
0.0 °
Representar
EPSG:3857

4. Clasificación supervisada con embeddings AEF en Google Earth Engine

Pasos principales

- Se definió un **AOI** a partir del *buffer de 1 km* del socioecosistema.
- Se extrajeron los embeddings mediante:

```
emb_image = ee.Image("GOOGLE/AE/IMG_V1").clip(aoi)
```

- Se muestrearon los píxeles según los polígonos curados:

```
sample = emb_image.sampleRegions(
    collection=fc,
    properties=["class_id"],
    scale=40,   # ajustado para evitar EEException: memory excee
    geometries=False
)
```

- Se añadieron **5 folds** para validación cruzada:

```
sample = sample.randomColumn("rand").divide(k).int()
```

- Se entrenó un clasificador:

```
classifier = ee.Classifier.smileRandomForest(
    numberOfTrees=100,
    minLeafPopulation=5
)
```

5. Resultados de la validación cruzada (5-fold CV)

Los resultados fueron evaluados en términos de:

- **Kappa de Cohen**
- **Exactitud global (OA)**
- **Matriz de confusión por pliegue**

Ejecutar k-fold CV

```
# Alternativa 1: usando list comprehension
results = [train_and_evaluate_fold(i) for i in range(k)]

for r in results:
    print(f"Fold {r['fold']}: "
          f"Train OA={r['train_accuracy']:.3f}, κ={r['train_kappa']:.3f} | "
          f"Val OA={r['val_accuracy']:.3f}, κ={r['val_kappa']:.3f}")
```

```
Fold 0: Train OA=0.986, κ=0.981 | Val OA=0.971, κ=0.961
Fold 1: Train OA=0.986, κ=0.981 | Val OA=0.972, κ=0.962
Fold 2: Train OA=0.986, κ=0.981 | Val OA=0.970, κ=0.960
Fold 3: Train OA=0.985, κ=0.980 | Val OA=0.972, κ=0.963
Fold 4: Train OA=0.986, κ=0.981 | Val OA=0.975, κ=0.967
```

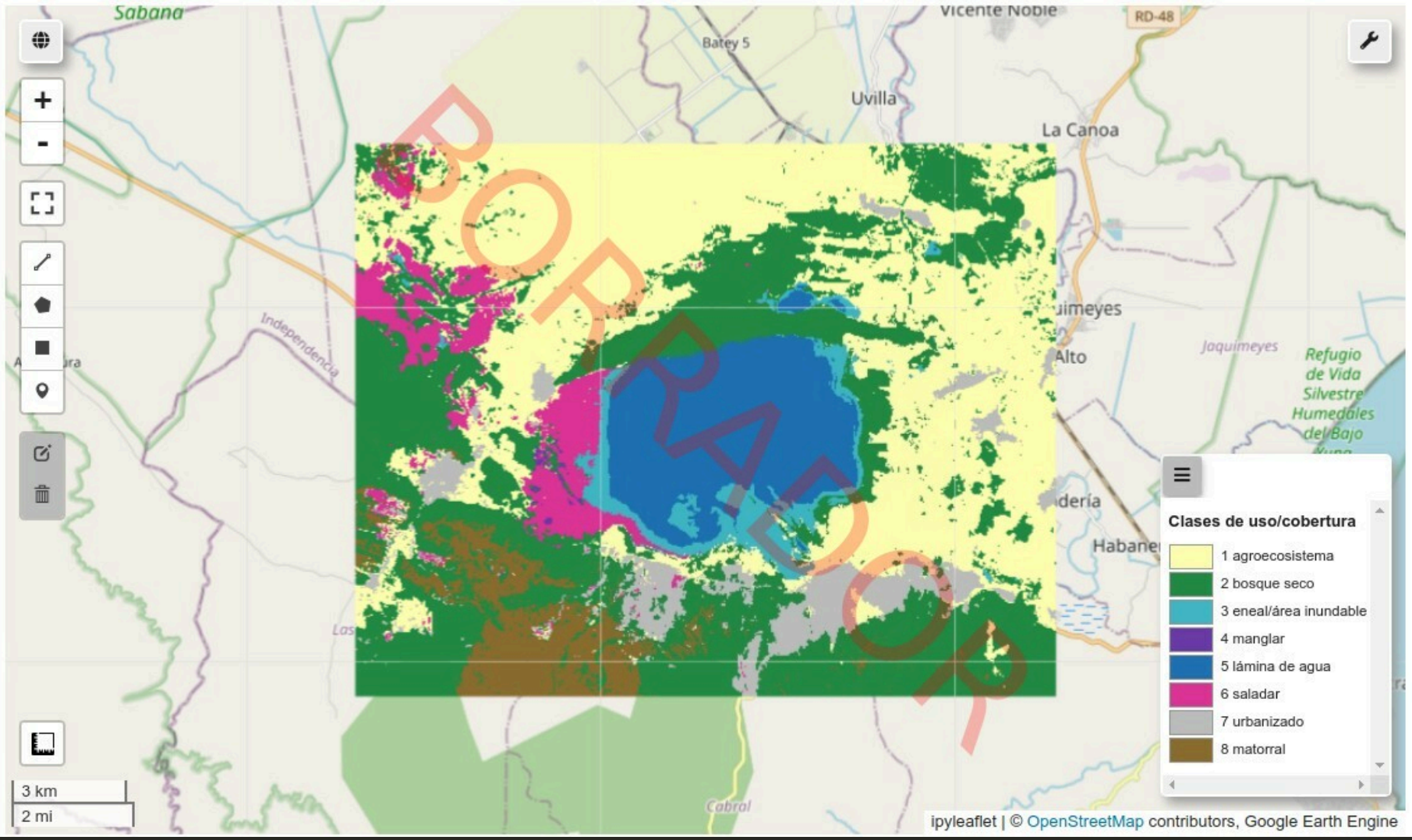
Métricas promedio de validación

```
val_accs  = [r["val_accuracy"] for r in results]
val_kappas = [r["val_kappa"] for r in results]

print("Promedio OA validación:", np.mean(val_accs))
print("Desv. estándar OA:", np.std(val_accs))
print("Promedio κ validación:", np.mean(val_kappas))
print("Desv. estándar κ:", np.std(val_kappas))
```

```
Promedio OA validación: 0.9718087715864383
Desv. estándar OA: 0.0017696703584463 98
Promedio κ validación: 0.9623802751491193
Desv. estándar κ: 0.002317866532848917
```

Mapa de clasificación


Sabana
Vicente Noble
RD-48
Batey 5
Uvilla
La Canoa
Independencia
Alto
Jaquimeyes
Refugio de Vida Silvestre Humedales del Bajo
Habane
Cabral
BORRADOR
Clases de uso/cobertura
1 agroecosistema
2 bosque seco
3 eneal/área inundable
4 manglar
5 lámina de agua
6 saladar
7 urbanizado
8 matorral
3 km
2 mi
ipyleaflet | © OpenStreetMap contributors, Google Earth Engine

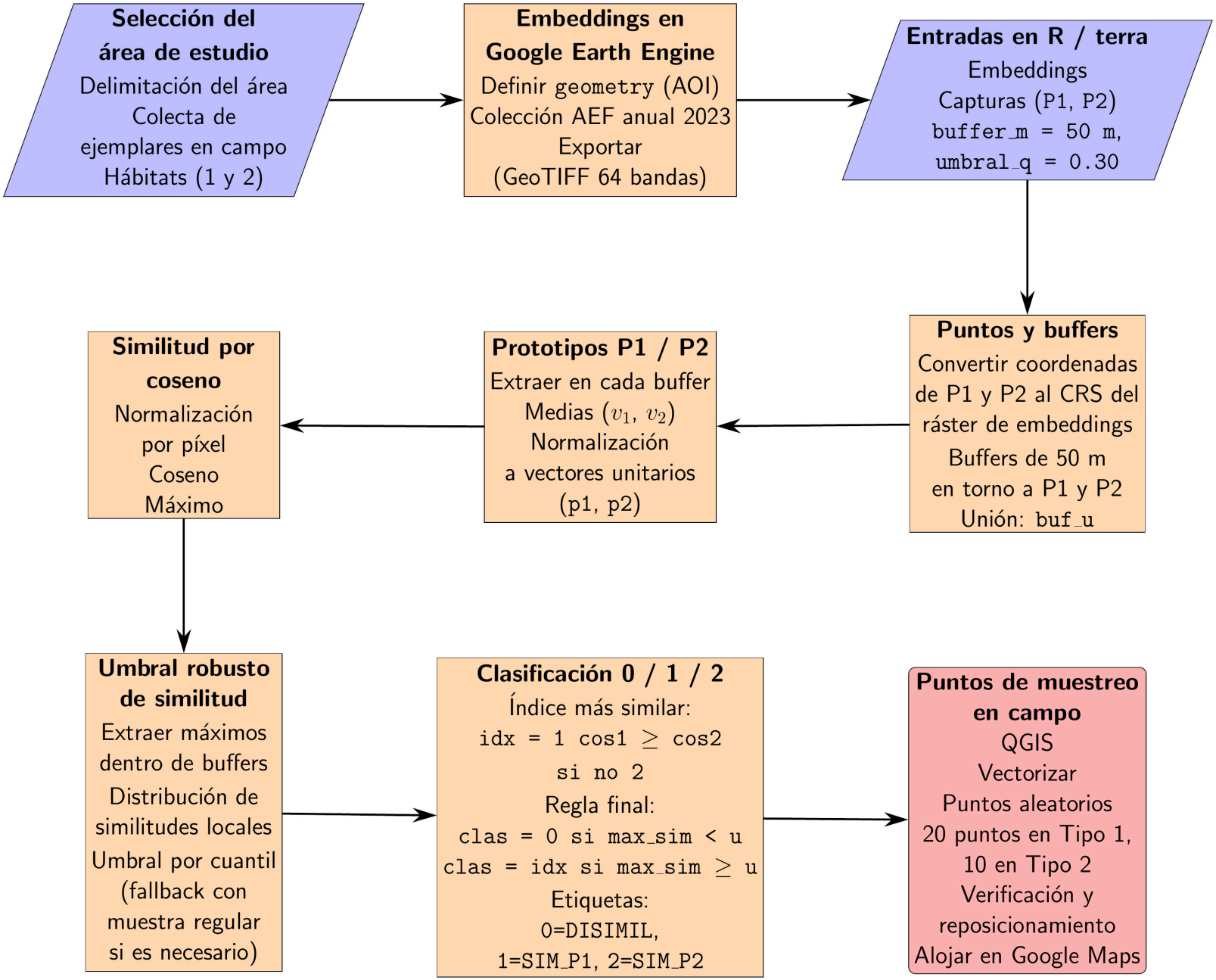
Conclusiones del Objetivo 2

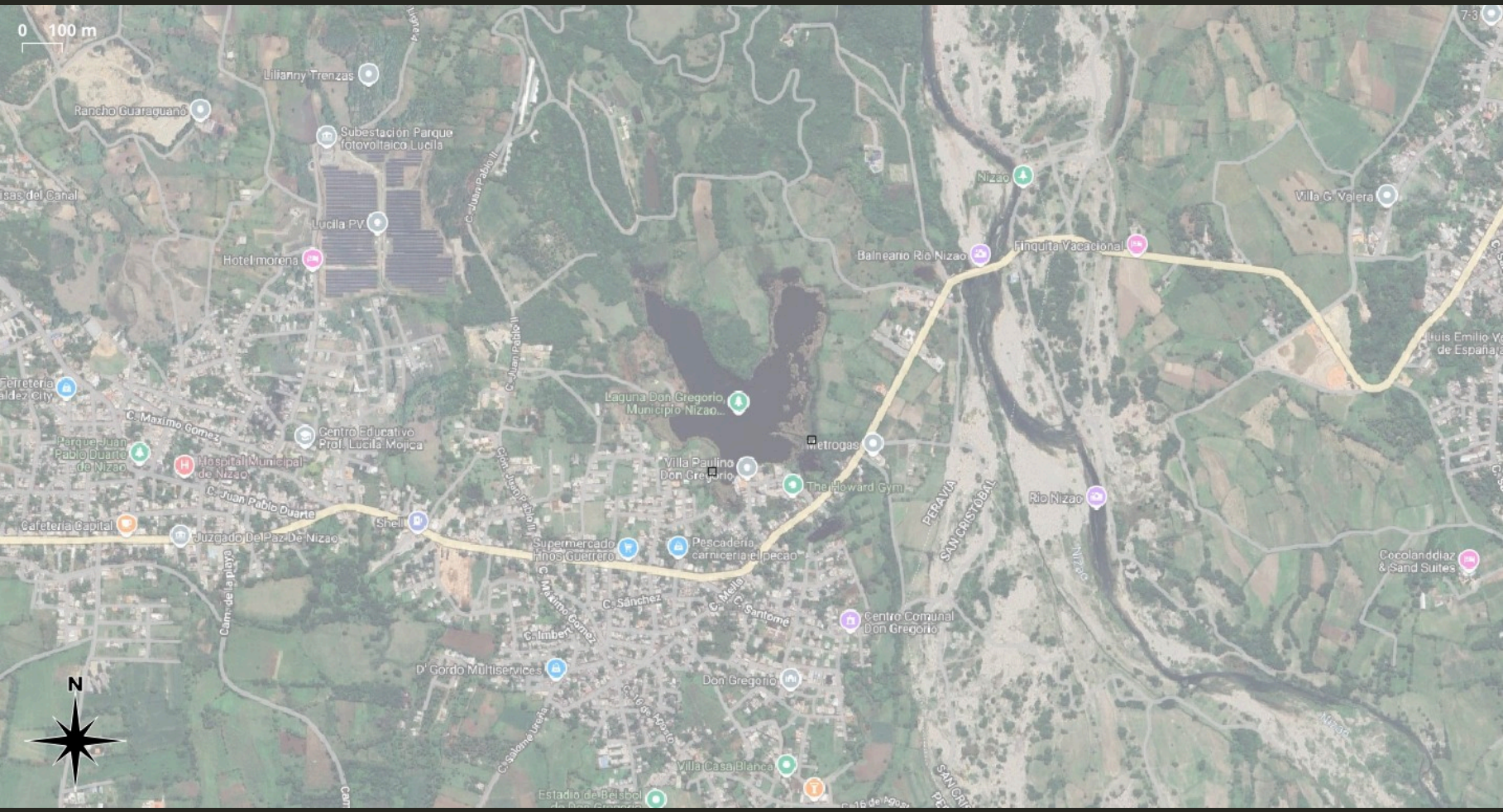
- Los embeddings de AEF **sí capturan diferencias estructurales** entre agroecosistemas, áreas forestales/arbustivas y zonas urbanas.
- Con solo unos pocos polígonos de entrenamiento (segmentados vía SAM), se logró una clasificación robusta.
- El uso combinado de **SAM** + **QGIS** + **AEF embeddings** + **Random Forest** constituye un flujo de trabajo reproducible y eficiente para análisis de cobertura sin necesidad de deep learning complejo.
- La validación cruzada mostró niveles consistentes de desempeño entre pliegues, indicando estabilidad del clasificador.

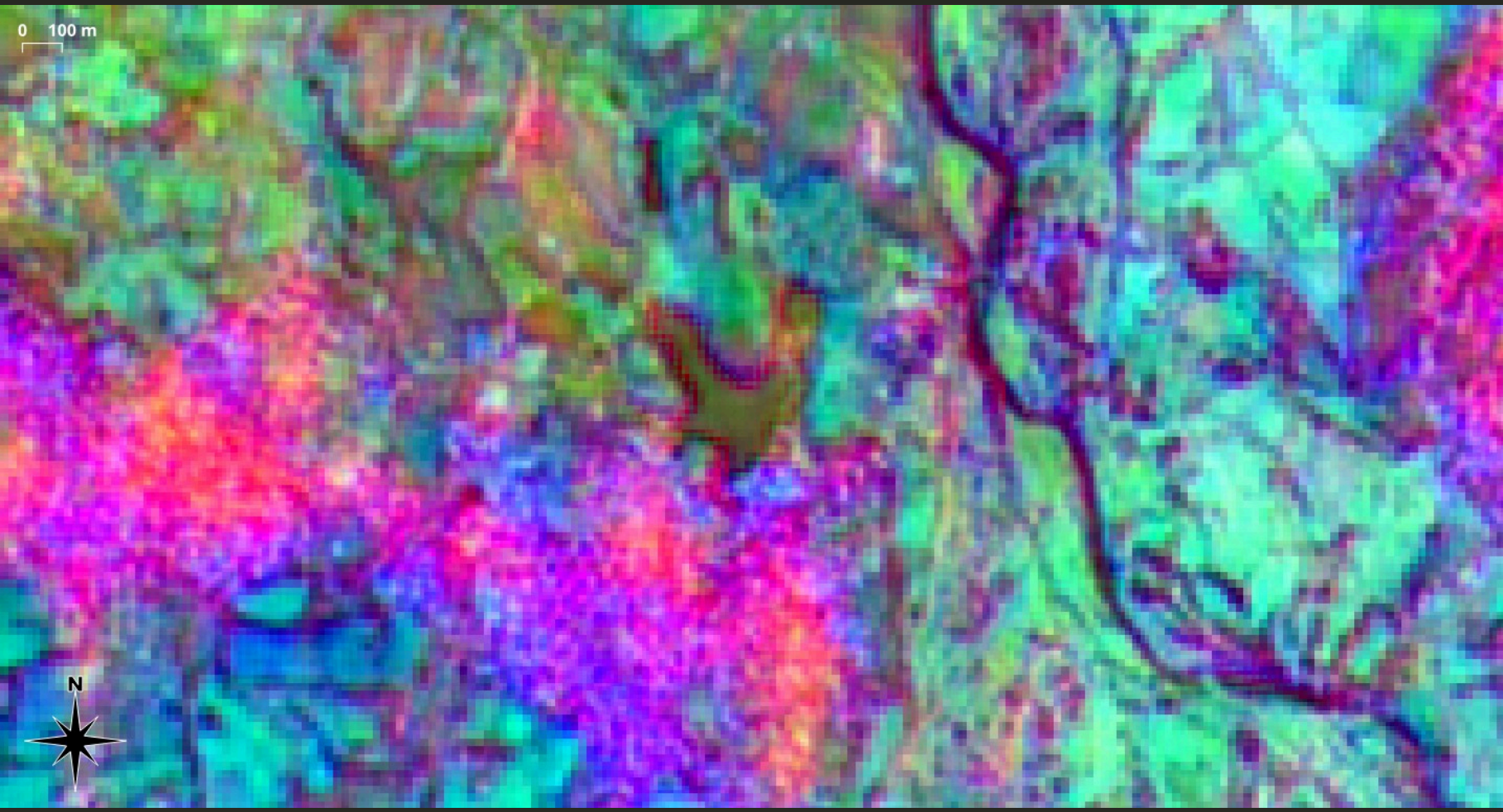
Objetivo 3

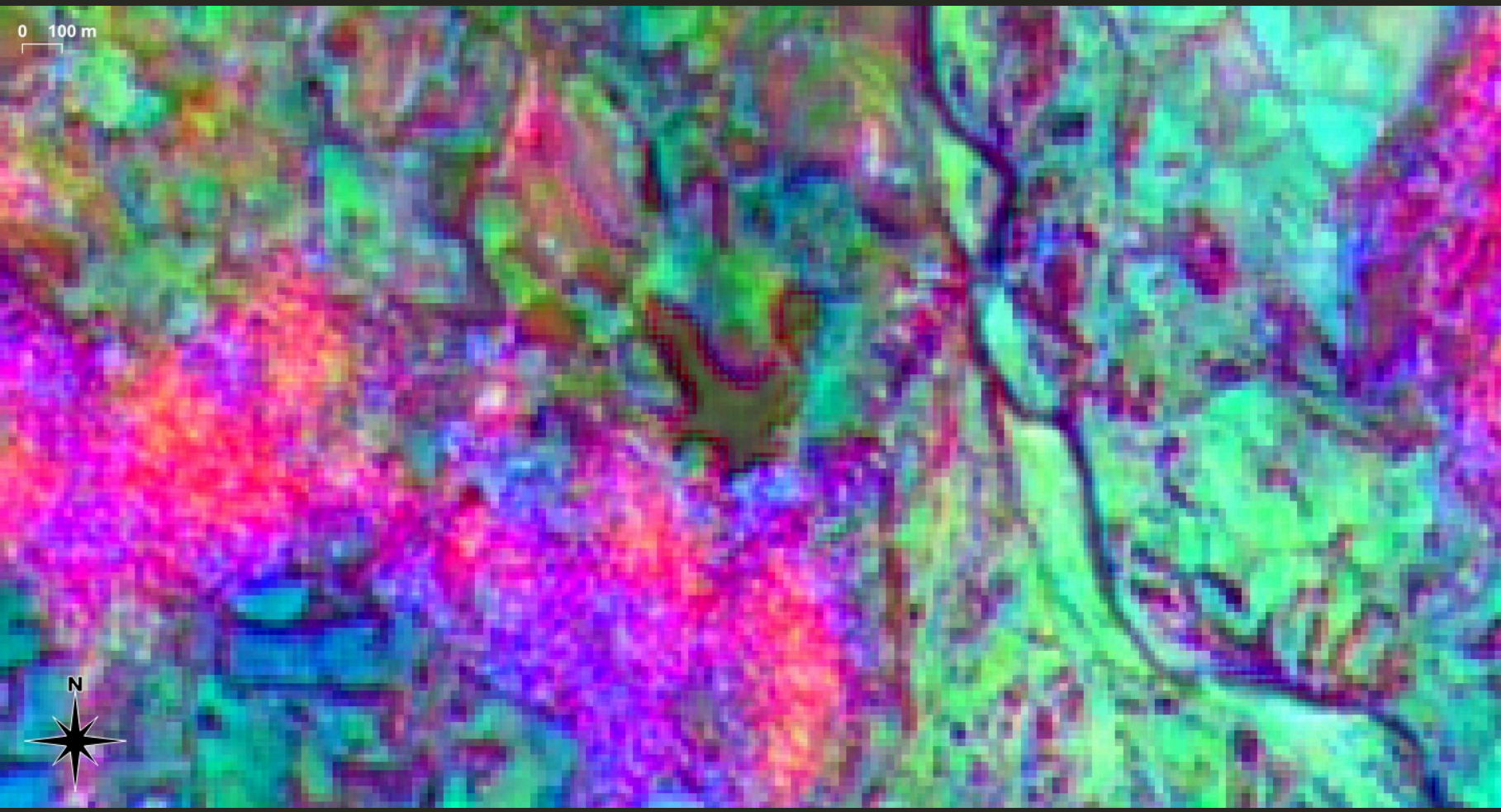
Analizar la aplicación de embeddings de AEF en el **diseño de muestreo estratificado**, utilizando la similitud entre embeddings para optimizar la representatividad espacial y ambiental

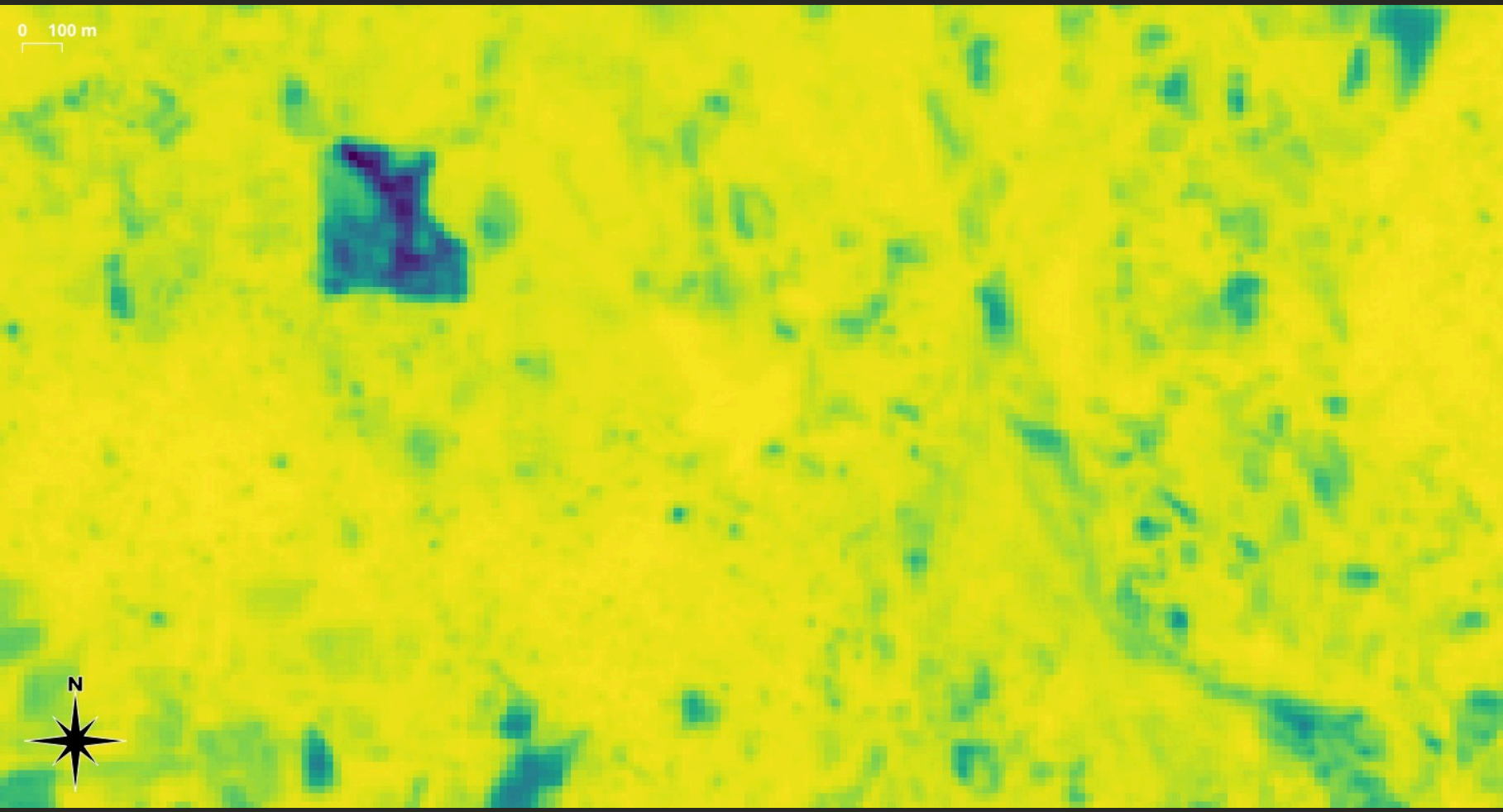
Colaboración para María Fernanda Rodríguez y Wellin Brito.

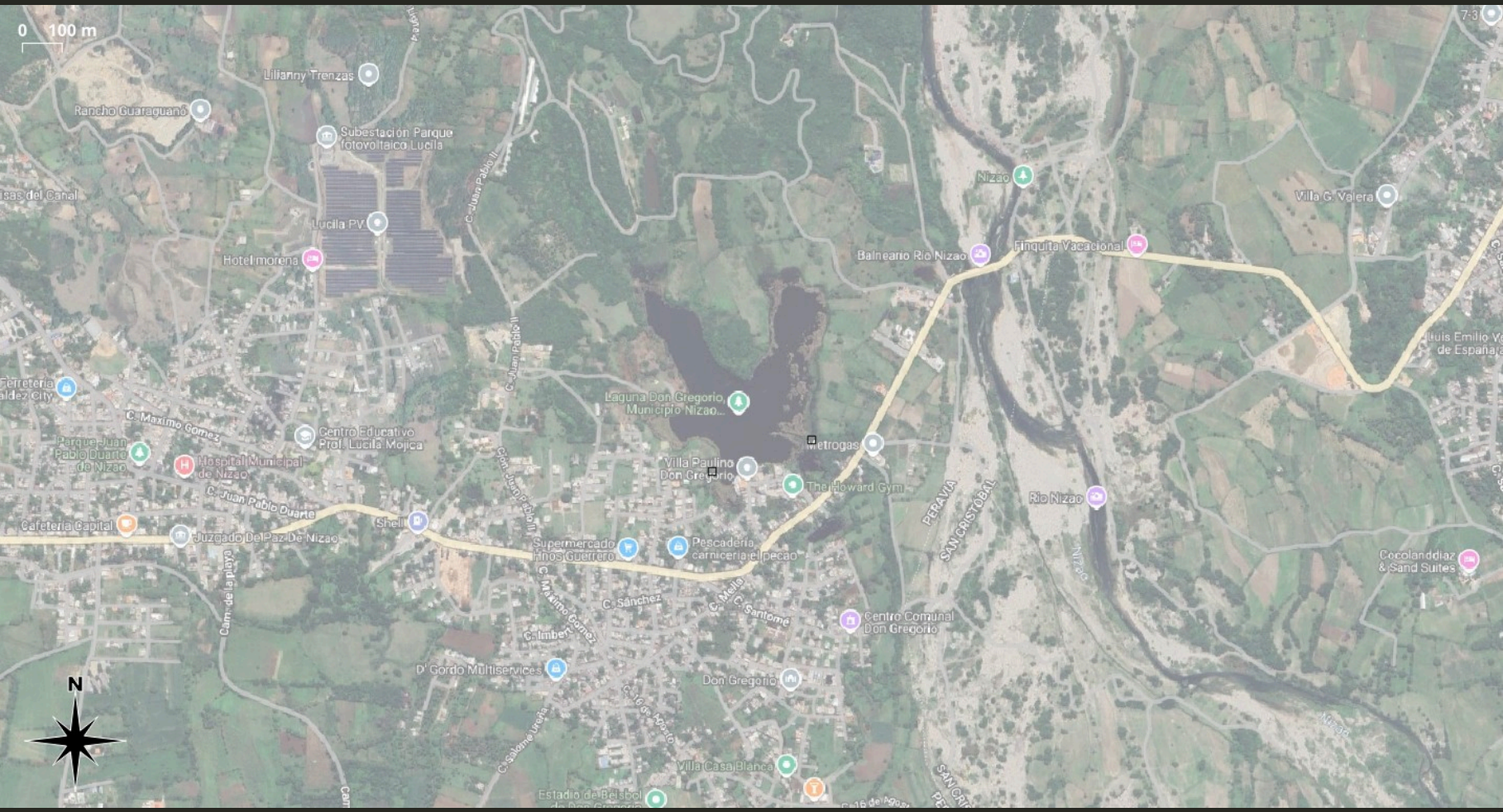

Selección del área de estudio
Delimitación del área
Colecta de ejemplares en campo
Hábitats (1 y 2)
Embeddings en Google Earth Engine
Definir geometry (AOI)
Colección AEF anual 2023
Exportar (GeoTIFF 64 bandas)
Entradas en R / terra
Embeddings
Capturas (P1, P2)
buffer_m = 50 m,
umbral_q = 0.30
Puntos y buffers
Convertir coordenadas de P1 y P2 al CRS del ráster de embeddings
Buffers de 50 m en torno a P1 y P2
Unión: buf_u
Prototipos P1 / P2
Extraer en cada buffer
Medias (v_1, v_2)
Normalización a vectores unitarios (p1, p2)
Similitud por coseno
Normalización por píxel
Coseno
Máximo
Umbral robusto de similitud
Extraer máximos dentro de buffers
Distribución de similitudes locales
Umbral por cuantil (fallback con muestra regular si es necesario)
Clasificación 0 / 1 / 2
Índice más similar:
idx = 1 cos1 ≥ cos2
si no 2
Regla final:
clas = 0 si max_sim < u
clas = idx si max_sim ≥ u
Etiquetas:
0=DISIMIL,
1=SIM_P1, 2=SIM_P2
Puntos de muestreo en campo
QGIS
Vectorizar
Puntos aleatorios
20 puntos en Tipo 1,
10 en Tipo 2
Verificación y reposicionamiento
Alojar en Google Maps


0 100 m
Lilianny Trenzas
Rancho Guaraguano
Subestación Parque fotovoltaico Lucila
Lucila PV
Hotel morena
Nizao
Villa G. Valera
Balneario Rio Nizao
Finquito Vacacional
Laguna Don Gregorio, Municipio Nizao...
Centro Educativo Prof. Lucila Mojica
Hospital Municipal de Nizao
Parque Juan Pablo Duarte de Nizao
C. Maximo Gomez
C. Juan Pablo Duarte
Villa Paulino Don Gregorio
Metrogas
The Howard Gym
Rio Nizao
Cafeteria Capital
Shell
Juzgado De Paz De Nizao
Supermercado Hnos. Guerrero
Pescaderia carniceria el pecao
PERAVIA
SAN CRISTOBAL
Cocolanddiaz & Sand Suites
C. Sanchez
C. Mella
C. Santome
Centro Comunal Don Gregorio
D' Gordo Multiservices
Don Gregorio
Villa Casa Blanca
Estadio de Beisbol
N


0 100 m
N

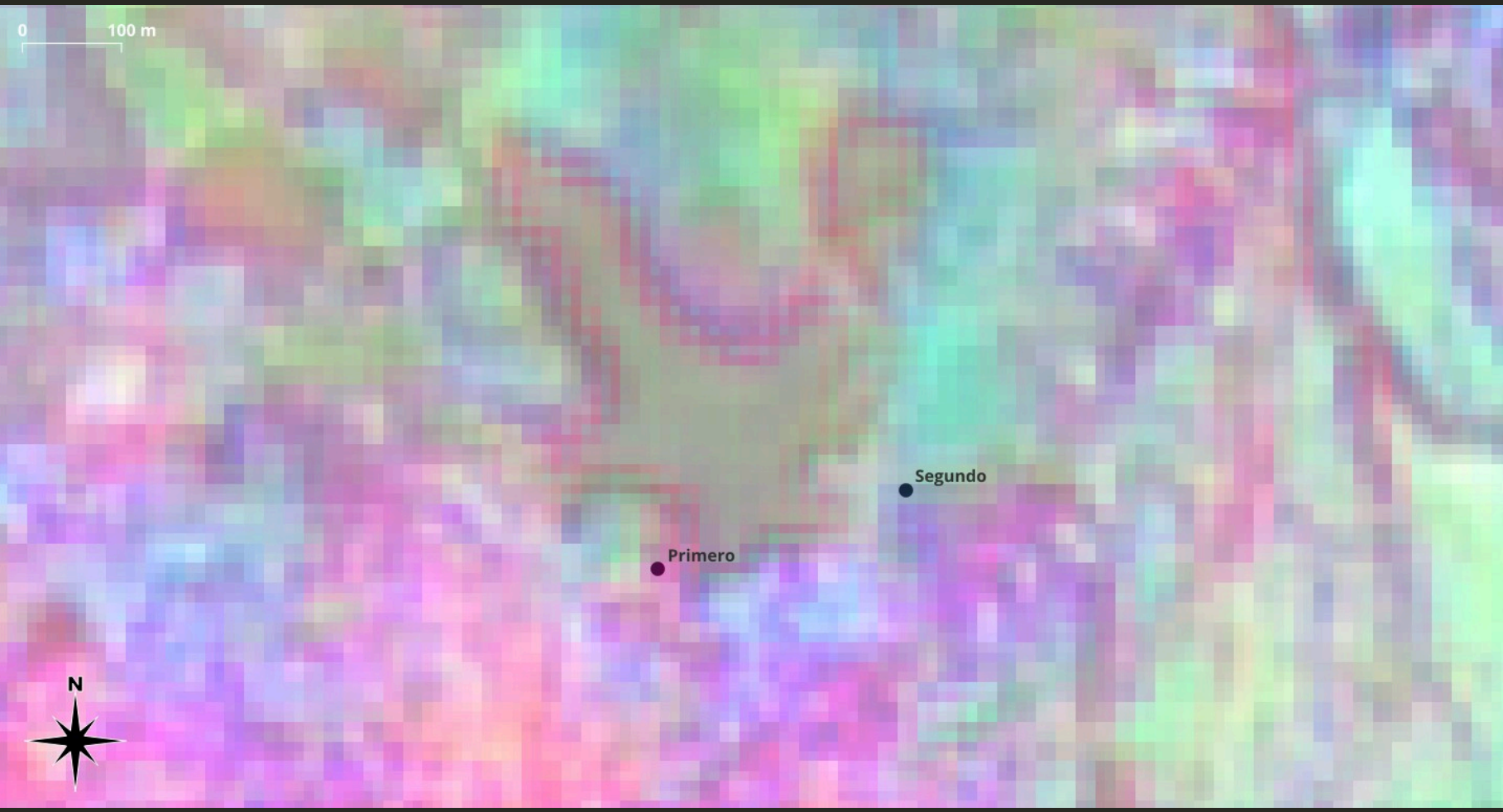

0 100 m
N


0 100 m
N


0 100 m
Lilianny Trenzas
Rancho Guaraguano
Subestación Parque fotovoltaico Lucila
Lucila PV
Hotel morena
Nizao
Villa G. Valera
Balneario Rio Nizao
Finquito Vacacional
Laguna Don Gregorio, Municipio Nizao...
Centro Educativo Prof. Lucila Mojica
Hospital Municipal de Nizao
Parque Juan Pablo Duarte de Nizao
C. Maximo Gomez
C. Juan Pablo Duarte
Villa Paulino Don Gregorio
Metrogas
The Howard Gym
Rio Nizao
Cafeteria Capital
Shell
Juzgado De Paz De Nizao
Supermercado Hnos. Guerrero
Pescaderia carniceria el pecao
PERAVIA
SAN CRISTOBAL
Cocolanddiaz & Sand Suites
C. Sanchez
C. Mella
C. Santome
Centro Comunal Don Gregorio
D' Gordo Multiservices
Don Gregorio
Villa Casa Blanca
Estadio de Beisbol
N


0
100 m
Balneario Río Nizao
Fabrica de Block la Patrona
C. Juan Pablo II
Laguna Don Gregorio, Municipio Nizao...
Segundo
Primero
D'Anyelina Gomas y Lubricantes
Metrogas
Don Gregorio
Plandy Milan Imports
Bloke Negro Studio
The Howard Gym
PERAVIA SAN CRISTÓBAL
N
Supermercado Hnos Guerrero
Pescadería carnicería el pecao
Patio de Frida


0
100 m
Balneario Río Nizao
C. Juan Pablo II
Metrogas
Plandy Milan Imports
Supermercado Hnos Guerrero
Pescadería carnicería el pecao
Patio de Frida
PERAVIA SAN CRISTÓBAL
N


0
100 m
Segundo
Primero
N

File Edit Code View Plots Session Build Debug Profile Tools Help

jose

Go to file/function | Addins | tesis-maria-fernanda

bitacora-diseno-muestreo.Rmd

Knit on Save | Knit | Run | Outline

Source | Visual

```
98  # SIMILITUD POR COSENO
99  # --------------------------
100 # Normalización por píxel: v / ||v||
101 # sum_sq <- sum(r^2)               # suma de cuadrados por píxel (1 capa)
102 sum_sq <- terra::app(r, fun = function(v) sum(v * v, na.rm = TRUE))
103 norm_r <- sqrt(sum_sq) + 1e-12
104 r_norm <- r / norm_r                    # cada banda dividida por ||v|| pixelar
105 r_norm
106
107 # cos(pixel, p1) y cos(pixel, p2)
108 cos1 <- app(r_norm, fun = function(v) sum(v * p1), cores = 0)
109 cos2 <- app(r_norm, fun = function(v) sum(v * p2), cores = 0)
110 names(cos1) <- "cos_P1"; names(cos2) <- "cos_P2"
```

46:1 (Untitled) | Copilot: No completions available. | R Markdown

Console | Terminal | Background Jobs

R 4.4.0 · ~/tesis-maria-fernanda/

```
coord. ref. : WGS 84 / UTM zone 19N (EPSG:32619)
source(s)   : memory
categories  : class
name        :   clase
min value   : DISIMIL
max value   :  SIM_P2
>
> # Plot
> plot(clas, col = c("grey80","dodgerblue","orange"),
+      main = "Clasificación: 0=Disímil, 1=Similar a P1, 2=Similar a P2")
> points(st_coordinates(pts_sf), pch = 21, bg = "red", cex = 1.2)
>
```

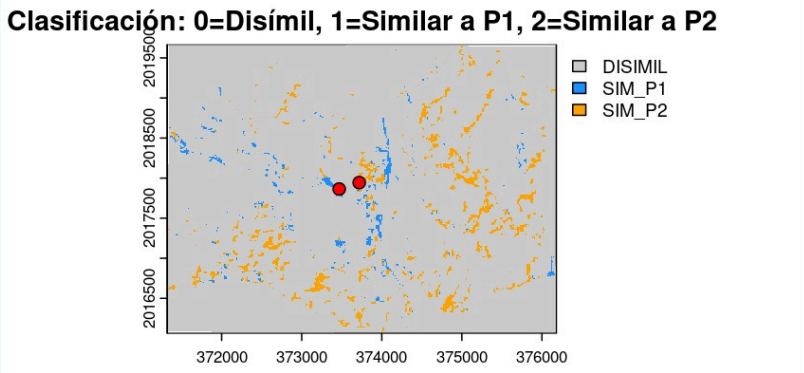
Environment | History | Connections | Git | Tutorial

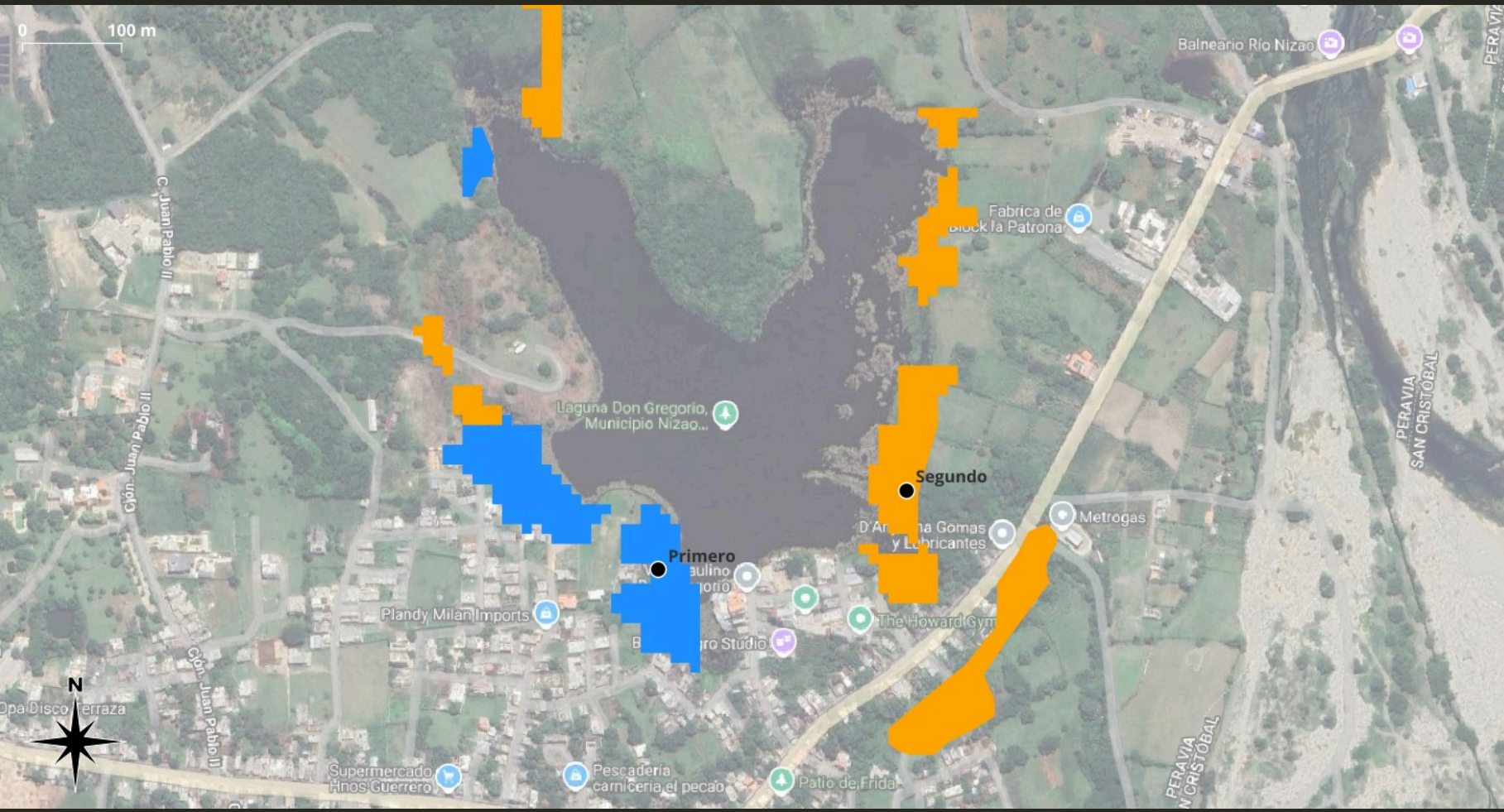
Diff | Commit | Pull | Push | History | More | main

Staged	Status	Path
☐	M	bitacora-diseno-muestreo.Rmd

Files | Plots | Packages | Help | Viewer | Presentation

Zoom | Export | Publish




0
100 m
Balneario Río Nizao
Fabrica de Block la Patrona
Laguna Don Gregorio, Municipio Nizao...
Segundo
Primero
Metrogas
Plandy Milan Imports
The Howard Gym
Supermercado Hnos Guerrero
Pescadería carnicería el pecao
Patio de Frida
C. Juan Pablo II
PERAVIA SAN CRISTÓBAL
N


0
100 m
Punto 27 (Tipo 2)
Punto 2 (Tipo 1)
Punto 4 (Tipo 1)
Punto 26 (Tipo 2)
Punto 22 (Tipo 2)
Punto 25 (Tipo 2)
Punto 24 (Tipo 2)
Punto 7 (Tipo 1)
Punto 6 (Tipo 1)
Punto 5 (Tipo 1)
Punto 11 (Tipo 1)
Punto 12 (Tipo 1)
Punto 9 (Tipo 1)
Punto 17 (Tipo 1)
Punto 18 (Tipo 1)
Punto 15 (Tipo 1)
Punto 14 (Tipo 1)
Punto 13 (Tipo 1)
Punto 3 (Tipo 1)
Punto 8 (Tipo 1)
Punto 29 (Tipo 2)
Punto 21 (Tipo 2)
Punto 28 (Tipo 2)
Punto 30 (Tipo 2)
Punto 23 (Tipo 2)
BORRADOR
Balneario Río Nizao
Plandy Milan Imports
The Howard Gym
Supermercado Hnos Guerrero
Pescadería carnicería el pecao
Patio de Frida
PERAVIA SAN CRISTÓBAL
N


0
3 km
VERDE PROVINCE
IAGO PROVINCE
Rio de mao
H Picao
VALVERDE PROVINCE
SANTIAGO PROVINCE
Super Colmado Núñez
Parque Solar
Sajoma EGE Haina
Comunidad
Terapeutica Hermanos...
Jaiqui picado
La Cuesta
N
BORRADOR

Conclusiones del Objetivo 3

- **Los embeddings AEF permiten definir estratos ambientales** sin mapas previos, usando solo similitud por coseno.
- **El umbral aprendido localmente genera estratos estables**, coherentes con los hábitats asociados a los puntos de terreno.
- **El muestreo estratificado resultante es más representativo**, asignando puntos según la estructura ambiental real del área.

Gracias por su atención

 jmartinez19@uasd.edu.do

 geofis