Satellite extractor: Towards a Smart Eco-epidemiological Model of Dengue in Colombia using Satellite Imagery*.

MIT CRITICAL DATA COLOMBIA
Sebastian Andres Cajas Ordoñez, David Restrepo, Kuan-Ting Kuo, Dana Moukheiber, Atika Rahman Paddo, Saptarshi Purkayastha, Leo Anthony Celi, Po-Chih Kuo, Juan Sebastián Osorio-Valencia, Kuan-Ting Ku, Braiam Escobar, Diego M. López, Cheng Che Tsai, Wilson Arbey Diaz, Luis Jesús Martínez, Alessa Álvarez, Siyi Tang, Amara Tariq, Imon Banerjee, Aakanksha Rana, Maria Patricia Arbelaez-Montoy, Cheng Che Tsai, Laura Sofía Daza Rosero, Jhon Fredy Romero Núñez, Wilson Arbey Diaz, Luis Jesús Martínez, Saketh Sundar, Alessa Álvarez, Siyi Tang, Amara Tariq, Imon Banerjee, Aakanksha Rana, Ivan Darío Velez, Maria Patricia Arbelaez-Montoya.
*Project supported by ESA Network of Resources Initiative.

MIT Critical Data

Project Goals

- Set up a database including satellite images, related sociodemographic, and entomological metadata associated with dengue outbreaks in Colombia. This will be used to reproduce the success in this project and for further research and educational purposes.
- Develop and validate unsupervised/semi-supervised and supervised deep learning models to identify the areas with the highest risk of dengue outbreak in Colombia.
- Build a community around dengue where multidisciplinary teams collaborate, do research, educate and prevent dengue outbreaks.

MIT Critical Data

Our Solution: Satellite Extractor

Satellite extractor allows the download of satellite imagery on any coordinates and fixed timestamps. The best image is obtained with least cloud cover per epi-week window using a forward-backward artifact removal algorithm. The images extracted in the framework are also related with relevant information of the region like demographic, economic, climatic or epidemiological data in a json file. The framework also provides a contrastive analysis using image hash encryption to avoid duplicates.

MIT Critical Data

What does *it do*?

- We used hash encryption to improve the quality of the images.
- We developed a framework for collecting and processing Sentinel-2 satellite data using modified Copernicus Sentinel data. We extracted satellite images of **+80** cities by utilizing the SentinelHub API and a scalable, dockerized framework that incorporates Google Earth Engine (GEE) to generate regions of interest (ROI).
- The framework includes a novel recursive forward-backward artefact removal algorithm with inter-band data augmentation on satellite imagery, cloud removal based on LeastCC, and Nearest Interpolation for spatial resolution that reduces camera capturing noise generated by the Sentinel 2-L1C orbit transit per week
- Once the images are extracted, the epidemiological week and the name of the region are used to generate a JSON file including relevant metadata, like climatic, sociodemographic or socioeconomic data.
- Data stored in GCP and Oracle and further pre-processed and trained in Oracle Cloud.

MIT Critical Data

GCP

ORACLE

Satellite imagery via satellite extractor

- Image size: 750 x 750
- Spatial resolution: 10m/px. Nearest neighbor interpolation used in bands with less resolution.
- Temporal resolution: 1 image per week (epiweek) selected using Least cloud coverage.
- Format: tiff
- Bands: 12 bands in total (Table).
- Bands 2-8 are most commonly used.

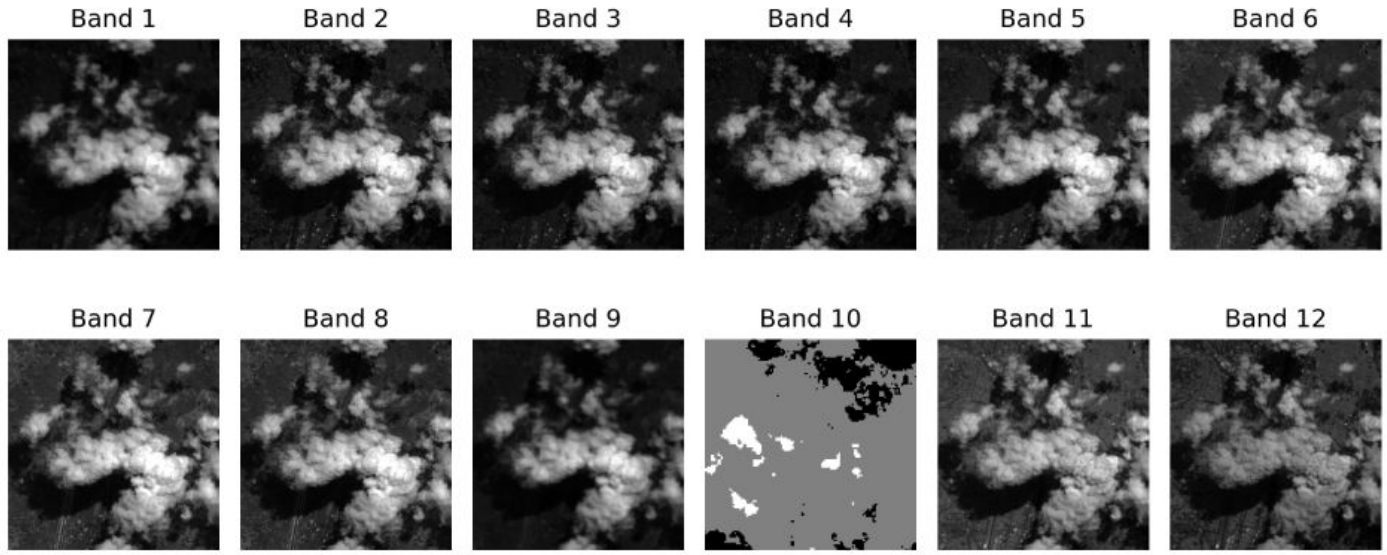
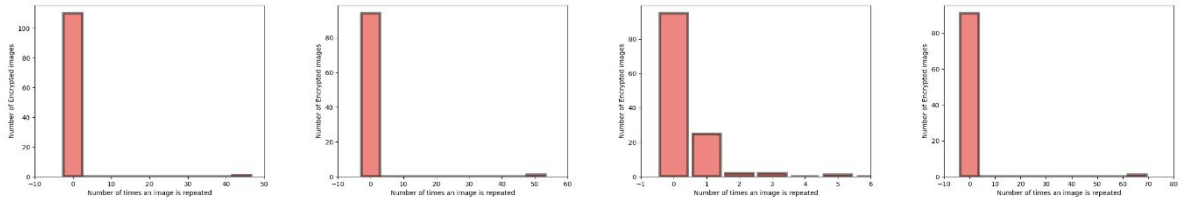


Figure 2: Example of 12 bands in a satellite imagery.

MIT Critical Data

GCP ORACLE

Satellite imagery acquisition via satellite extractor



(e) 5360 (f) 68001 (g) 73001 (h) 23001

(i) 54001 (j) 8001

Figure 3. dhash encryption analysis to visualize the quality the downloaded satellite images. Figures a-j depict the number of repeated images on the x-axis, and the number of images for each case on the y-axis across the ten Colombian municipalities identified by the municipality code.

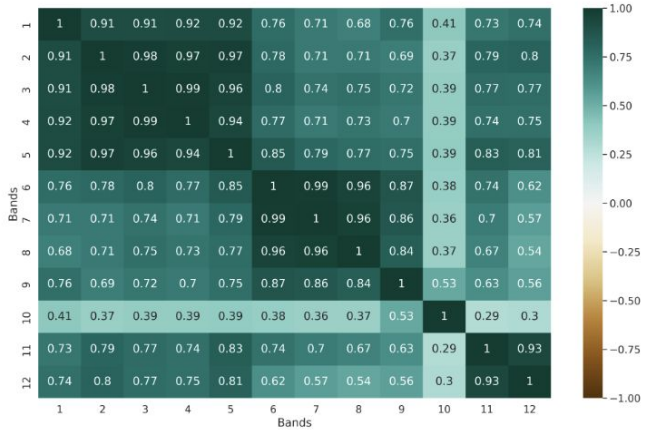
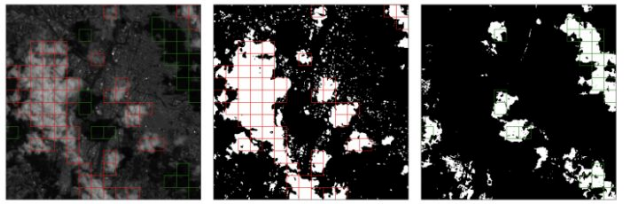


Figure 3: Average Pearson's correlation of the 12 bands for the Sentinel-2 satellite images across five Colombian municipalities in the training set from 2016 to 2018. The majority of correlations in this plot are statistically significant ($p < 0.001$).



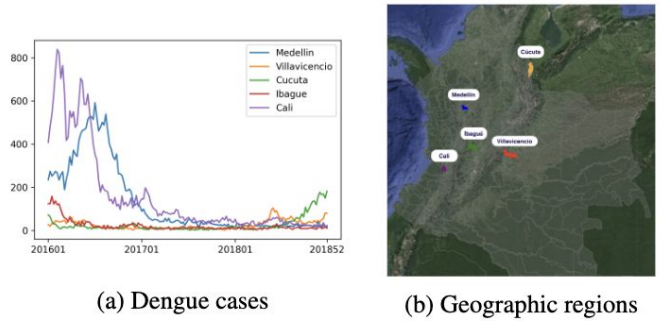
(a) Original image (b) Cloud mask (c) CS mask

Figure 4: Cloud and cloud shadow masks for tile swapping. (a) Original image with the abnormal tiles that will be swapped with the average of normal tiles. (b) Cloud mask with detected abnormal cloudy pixels in white and normal pixels in black. Abnormal tiles detected by the cloud mask are highlighted in red. (c) Cloud shadow (CS) mask with detected abnormal shadowy pixels in white and normal pixels in black. Abnormal tiles detected by the shadow mask are highlighted in green.

MIT Critical Data

Metadata Description: Extracting the number of dengue cases and corresponding factors in Colombia

- Dengue cases are extracted from the National Institute of Health (INS) and their surveillance system (SIVIGILA)
- Sociodemographic and Socioeconomic data was extracted from the National Administrative Department of Statistics (DANE)
- Climatic data was extracted from satellites MODIS and CHIRPS using Google Earth Engine



(a) Dengue cases

(b) Geographic regions

Figure 5: Municipality-level dengue case numbers and geographic locations. (a) Dengue cases from 2016 to 2018 obtained from the SIVIGILA database for the top five affected municipalities in Colombia. (b) Geographic locations from satellite imagery for each municipality.

Table 2. Distribution of Stable, Increased and Decreased labels across ten Colombian municipalities associated with each epi-week satellite image in 2016-2018.

Municipality	Stable	Increased	Decreased
Medellín	113	20	23
Cali	127	12	17
Villavicencio	97	31	28
Cúcuta	121	20	15
Ibagué	104	23	29
Bucaramanga	90	34	32
Neiva	105	26	25
Montería	110	24	22
Barranquilla	95	30	31
Itagüí	128	14	14

Google Earth Engine

INSTITUTO NACIONAL DE SALUD

SIVIGILA

DANE

Preliminary results

Deep learning models have been used and are currently being evaluated in order to assess and improve the quality of the data.

- This research has also allowed the exploration of the use of machine learning for satellite images in epidemiology and public health, as well as allowing experimentation in new data fusion techniques for the improvement of existing machine learning models and techniques.
- Novel techniques and models for multimodal data fusion using satellite imagery and metadata are being explored and implemented.
- Pretrained models and codes will be released to encourage research in the community.

MIT Critical Data

Preliminary results

Table 3. AUC for novel satellite imagery model architectures and standard meta data based models. The first four rows use time-series satellite imagery as inputs, and the last two rows use meta data as inputs. Models with the best performance for both, tabular and satellite images data results are bolded.

Model	Medellín	Villavicencio	Ibagué
CL + LSTM	0.539±0.021	0.388±0.027	0.550±0.030
VAE + LSTM	**0.738±0.177**	0.560±0.078	0.413±0.051
AE + LSTM	0.617±0.094	0.527±0.062	0.510±0.049
ResNet50V2 + LSTM	0.517±0.154	**0.605±0.008**	**0.795±0.087**
Temperature + Precipitation (LSTM)	**0.603±0.087**	0.577±0.010	0.472±0.026
Cases (LSTM)	0.504±0.025	**0.617±0.012**	**0.522±0.020**

Table 4. sMAPE for novel satellite imagery model architectures and standard meta data based models. The first four rows use time-series satellite imagery as inputs, and the last two rows use meta data as inputs. Models with the best performance for both, tabular and satellite images data results are bolded.

Model	Medellín	Villavicencio	Ibagué
CL + LSTM	**62.689±46.98**	**55.295±14.868**	33.078±3.384
VAE + LSTM	132.645±24.418	64.779±7.744	35.020±7.905
AE + LSTM	157.499±44.780	59.831±15.552	**29.303±3.022**
ResNet50V2 + LSTM	129.876±49.827	100.747±18.522	57.512±18.035
Temperature + Precipitation (LSTM)	80.451±0.149	80.451±0.149	106.738±0.343
Cases (LSTM)	**79.477±10.613**	**44.626±9.092**	**38.679±3.522**

Metrics	Medellín	Ibagué	Villavicencio	Cali	Cúcuta
MAE	28.05±3.60	4.59±0.17	12.99±2.42	33.28±9.60	16.37±0.06
sMAPE	49.23±2.23	29.85±0.89	49.86±10.49	40.60±7.02	45.47±0.65
RMSE	38.70±6.33	6.22±0.21	21.76±2.69	43.24±13.29	40.55±0.02

(a) Evaluation of the DengueNet across five municipalities

Models	MAE	sMAPE	RMSE
DengueNet w/ TS	19.06± 4.71	43.00±6.11	30.09±6.69
AR Model	27.77±2.11	39.78±3.54	49.95±1.89
DengueNet + AR Model	**12.21±2.84**	**31.37±4.35**	**20.71±2.69**

(b) Comparison of different models' performances

Table 1: Evaluation and model comparison. The model comparisons compare our proposed method DengueNet (DengueNet + Tile Swapping(TS)) using satellite images to an LSTM-based autoregressive (AR) Model using dengue cases and the DengueNet + LSTM-based autoregressive Model with satellite images and dengue cases combined. All experiments are repeated three times, with the average value reported with the standard deviation (the best model scores are bolded).

TS	FEng	MobileNetV2	Medellín	Ibagué	Villavicencio	Cali	Cúcuta
✓	✓	✓	**28.05±3.60**	4.59±0.17	12.99±2.42	**33.28±9.60**	16.37±0.06
	✓	✓	28.42±3.07	**4.22±0.14**	**12.71±2.68**	36.00±3.62	16.43±0.21
✓	✓		22.58±4.28	4.40±0.21	16.39±0.05	59.06±16.86	16.48±0.07
✓		✓	101.54±0.11	4.45±0.20	14.08±0.25	80.20±1.42	**16.23±0.16**

(a) MAE

TS	FEng	MobileNetV2	Medellín	Ibagué	Villavicencio	Cali	Cúcuta
✓	✓	✓	49.23±2.23	29.85±0.89	49.86±10.49	**40.60±7.02**	45.47±0.65
	✓	✓	51.93±2.02	**28.08±0.52**	**49.35±11.22**	44.83±3.21	47.34±1.55
✓	✓		**36.12±1.72**	28.66±1.10	64.76±0.14	53.63±13.45	45.43±0.31
✓		✓	83.97±0.20	28.72±0.50	54.46±0.43	64.08±1.37	**44.82±1.39**

(b) sMAPE

TS	FEng	MobileNetV2	Medellín	Ibagué	Villavicencio	Cali	Cúcuta
✓	✓	✓	38.70±6.33	6.22±0.21	21.76±2.69	**43.24±13.29**	40.55±0.02
	✓	✓	39.19±4.52	**5.78±0.22**	**21.08±3.28**	44.32±5.08	**39.95±0.20**
✓	✓		**33.36±8.73**	6.17±0.36	25.46±0.07	87.27±27.92	40.79±0.12
✓		✓	170.80±0.12	6.35±0.49	23.78±0.35	124.20±2.71	40.38±0.21

(c) RMSE

Table 2: Ablation studies for different modules for the five municipalities. TS indicates applying tile swapping to the samples. FEng indicates using the features from the feature-engineering pipeline. MobileNetV2 indicates the features extracted from the MobileNetV2 are used. All experiments are repeated three times, with the average value being given ± the standard deviation (the method with the best metric is bolded).

MIT Critical Data

Project achievements

- Notice of Award for Oracle for Research Cloud Credits: “Towards a Smart Eco-epideiological Model of Dengue in Colombia using Satellite Images” project by the Oracle for Research Program.
- MISTI grant from MIT-Colombia - Cali Seed Fund.
- Paper under review at IJCAI-23: DengueNet: Prediction of Dengue Cases using Satellite Images
- Paper under working progress: DengueBench: A benchmark dataset integrating time-series satellite imagery with metadata for dengue outbreak prediction per epidemiological week

Project achievements

- Conference presentation: Osorio Valencia, Juan Sebastian. "Towards the Implementation of Eco-epidemiological Models for Dengue in Colombia Using Machine Learning and Satellite Images: Policy Advocacy and Open Data Repositories". Nov 24-26, 2021.

Project achievements

- Capacity Building: Two datathons were organized on the in Colombia and Latin America (Chile) under Make Health organization and two datathons are being planned for octover (Mexico) and november (Argentina).
- The founding universities of the National Center for Health Information Systems (CENS), MIT, Harvard University (USA), and the Digital Health Network of State Universities (RSDUE) organized this event, which includes three days of conferences, datathons, workshops, and panel discussions with renowned national and international experts.
- The focus was on how data science creates new knowledge and transforms our approach to health and quality of life.

MIT Critical Data

Conclusions

Using the satellite imagery sponsored by ESA, we have been able to obtain the following results:

- One papers on review (IJCAI-23) built using satellite imagery using our adaptation of SentinelHub API “Satellite.extractor”.
- Dataset benchmark paper under working progress, publicly available code on GitHub: https://github.com/sebasmos/satellite.extractor
- We obtained $100K Oracle credit to support this project, which is still on progress of execution.
- Two 2022 Make Health Datathon held in Colombia and Latin America with satellite images based on models weights trained on satellite.extractor.
- $29.7 MISTI-Colombia, Cali seed fund

MIT Critical Data

Our Team


Atika


Braiam


Wilson


Siyi


Maria


Ivan


Kuan-Ting


Po-Chih


Dana


Diego


Luis


Cheng Che


Sebastian


David


Juan


Leo

MAYO CLINIC

UAM Universidad Autónoma de Madrid

université de BORDEAUX

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

UNIVERSIDAD DE ANTIOQUIA Facultad de Medicina

PECET UNIVERSIDAD DE ANTIOQUIA Programa de Estudio y Control de Enfermedades Tropicales

HARVARD MEDICAL SCHOOL DEPARTMENT OF Biomedical Informatics

MIT Critical Data

UNIVERSIDAD CES Un compromiso con la excelencia

University at Buffalo

Universidad del Cauca

UNIVERSITY of WASHINGTON

Colombia CENTRE FOR THE FOURTH INDUSTRIAL REVOLUTION

Stanford University

UNIVERSIDAD NACIONAL DE COLOMBIA

國立清華大學 NATIONAL TSING HUA UNIVERSITY

EMORY UNIVERSITY

IUPUI

MIT Critical Data