Towards the Implementation of Eco-epidemiological models of Dengue in Colombia using Machine Learning and Satellite Images

Juan Sebastián Osorio-Valencia

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Public Health

University of Washington

2021

Committee:

Abraham D. Flaxman

Peter M. Rabinowitz

Leo A. Celi

Program Authorized to Offer Degree:

Global Health

University of Washington

Abstract

Towards the Implementation of Eco-epidemiological models of Dengue in Colombia using Machine Learning and Satellite Images

Juan Sebastián Osorio-Valencia

Chair of the Supervisory Committee:

Abraham Flaxman

Department of Global Health

Machine learning (ML) presents countless opportunities for population and public health research, and infectious disease modeling is among those. Dengue is a climate-sensitive disease, and, over the last 50 years, its incidence has increased 30-fold, with a distinctive high burden in countries like Colombia. ML and using deep learning on satellite images have gained more attention in recent years due to the amount of heterogeneous data that could inform dengue disease modeling. We introduced a project that aims to build responsible and explainable ML-based dengue models that supports later deployment and implementation. It includes a global health data science approach in Colombia, with the development of open databases, a spatial model for disease mapping, political incidence, and multi-stakeholder collaboration.

Table of Contents

List of Figures

List of Tables

Acknowledgements

To my parents: Libardo Antonio and Luz Margarita, who have been unconditional and supportive during my entire career. To all the members of the *Dengue and Satellite Image Team*, who are integral to the project I am introducing here and with whom we are moving data science within the public health community in Colombia forward. To Colombia Científica, ICETEX and Fulbright, that provided the financial support. To all my mentors, family, and friends, particularly my amazing DGH cohort and Blue House. Even though the last year was one of the toughest, it was the year, in my lifetime, that science and public health stood out, and the need for better leadership was reinforced. I would like to acknowledge as well my commitment to that, not only by leading but finding our future leaders.

Introduction

Artificial intelligence and machine learning (ML) are disrupting healthcare and present a myriad of opportunities and threats for population and public health research. Infectious disease modelling and outbreak prediction are among those opportunities. The COVID-19 pandemic has demonstrated the importance of urban intelligence: analyzing city-level information using data science methods and exploring its role in an outbreak response (Lai, Yeung and Celi, 2020). On that note, rapid spread of vector-borne disease (VBD) outbreaks, like dengue, poses a major problem for countries that are most affected by global warming (Ebi and Nealon 2016), a surging population with disorderly urbanization, high migration rates, and growing socioeconomic gaps (Hagenlocher, et al. 2013). Colombia has around 25 million people at risk of dengue, which is more than half of its population, and a number of large-scale outbreaks have occurred in recent years (PAHO, 2021). The number of cases reported in 2019 was the highest registered in the history of dengue in the Americas, exceeding the number of reported cases by 30% in the epidemic year 2015, and the country level's cumulative incidence rate was 475.4 cases per 100,000 inhabitants (Gutierrez-Barbosa et al. 2020). Even though the country has surveillance and control strategies in place, mean case fatality rate is still significantly high (0.186%) compared to other countries in the America's region (4.84 times higher), posing a high political pressure to decision-makers.

To control, prevent and respond to dengue outbreaks, the World Health Organization (WHO) recommends preparedness and response plans, patient care plans, rapid laboratory diagnoses, risk communication, and integrated vector management (WHO, 2012). Implementing improved outbreak prediction and detection required coordinated surveillance and integrated multi-stakeholder vector management to deploy locally adapted control measurements. In this respect, vulnerability, defined as the predisposition of the health system and its population to be affected by a disease, and eco-epidemiology, that pertains to the macro-determinants influencing dengue infection, are concepts that become relevant (Birkmann 2013; Cutter et al. 2003; Füssel 2007). A dengue's ecoepidemiology vulnerability model considers a more holistic approach, including environmental, socioeconomic, and epidemiological factors. To build such models it is required to perform first descriptive and ecological studies of dengue, linking georeferenced cases with urban/social data at census. Silva et al. (2017) demonstrated how dengue deaths are concentrated in areas of social vulnerability at the municipality of São Luis (Maranhão, Brazil), and da Conceição Araújo et al. (2020) found that spatial modelling

explained 40% of the influence of social inequalities on dengue incidence in the state of Sergipe (Brazil). Conversely, Fuentes-Vallejo (2017) did not find a linear association of dengue with poverty or with vulnerable peripheral spaces in intra-urban settings in Girardot, Colombia. Although these studies have shown opposite results regarding the association between dengue and socioeconomic conditions, this relationship remains to be established. Hence, the spatial dimension is essential to better understand the transmission of this disease.

Machine learning and particularly, deep learning (DNN) approaches that are based on artificial neural networks with representation learning, has been used recently to analyze satellite images to obtain important additional information (i.e., deep landscape features) for such eco-epidemiological models (Abdur Rehman et al, 2019; Elisavet et al, 2019). In a recent study, utilizing satellite images to conduct remote sensing of the environment has shown to be proved as an effective process to estimate poverty and obesity rates in Africa and the United States, respectively (Jean et al. 2016; Maharana and Nsoesie 2018). Accelerating the translation of safe, ethically, responsible and meaningful ML-based models in health care and public health, requires engaged stakeholders and a systematic process from the problem formulation to the widespread deployment stage (Wiens et al. 2019). In that sense, having explainable models are preferred and would facilitate reproducibility. Hence, there is a pressing need of eco-epidemiological models for dengue vulnerability mapping in Colombia that include information from ML-based satellite image processing. Additionally, intuitive explanations are also needed to increase uptake of such ML-based disease transmission modelling in public health (Flaxman and Vos 2018).

MIT Critical Data is a consortium that consists of healthcare practitioners, computer scientists, engineers and social scientists who believe that data and learning are the best medicine for population health. This effort builds communities across disciplines to derive knowledge from data routinely collected in the process of care in order to understand health and disease better, and in the local context. The consortium is led by the Laboratory for Computational Physiology (LCP) at MIT. Sana MIT, an initiative to advance global health data science, is an arm of MIT Critical Data and focuses on democratizing knowledge and global networks of multidisciplinary experts. Our long-term goal is to build AI solutions for global health that are responsible and explainable that supports later deployment and implementation. The burden of dengue in Colombia and Latin America represents a motivation to advance eco-epidemiological models using machine learning approaches in public health research. At the same time that we

demystify AI solutions and reduce the hype around them for public health research, we want to reduce the potential for algorithms to reinforce social inequities. Consequently, two initial aims for such endeavors are presented: releasing an open database related to dengue outbreaks in Colombia and identifying areas with the highest risk of dengue in Colombia with novel ML-based modelling approaches that include heterogeneous data sources.

Materials and Methods

Setting the Stage for a Data Science Project in Global Health

MIT Critical Data has worked on strengthening the partnerships in Colombia. Student exchange, organization of workshops and hackathons, and invitation of guest speakers to the courses in Boston have been part of the initiatives. An effective multidisciplinary data science project needs early and solid partnerships, and the approach from organizations like MIT Critical Data, on a position of power and privilege, should be as enablers of such collaborations and allies to the local institutions, understanding that the actual change makers are in Colombia. Hess et al. (2020) reinforced the importance of aligning local priorities and capacities to tackle cross-scale problems and co-design the best modelling approaches with public health climate practitioners. Our process for setting the stage involved four quarters and specific milestones. The first quarter, starting in July 2020, we focused on the team formation, leveraging our contacts in the country, particularly those working on vector-borne diseases. The last quarter of 2020 and the first one of 2021 was mainly driven by funding and political action, as application to grants and conversations with local stakeholders, as well as data collection and curation. The second quarter of 2021, we started working on the model development and currently have three approaches: one is a purely machine learning regression, the second is a time series analysis, and the model we introduced in this paper is a spatial one that uses deep-landscape features.

Data collection and management

For this study, we followed the six phases of CRISP-DM (CRoss-Industry Standard Process for Data Mining), a process model that describes the data science life cycle (Shearer 2000). It is defined as a set of techniques and technologies that allow exploration of large databases, either automatically or semi-automatically, to find patterns, trends, or rules that explain the behavior of the data in a given context. We used block population information (Census 2018) and a mandatory case report database. In Colombia, dengue cases are reported through SIVIGILA, the national public health surveillance system. Cases are identified and reported by the healthcare system but corresponds only to symptomatic patients who are seen in a healthcare institution. Besides, the recorded residence does not necessarily represent the places of transmission. The data for reported dengue cases are recorded as aggregated totals on a weekly basis and collected from 2007 to 2019.

The population data, along with other socioeconomic and demographic variables, were downloaded from the National Administrative Department of Statistics (DANE 2018) and curated in a database uploaded to an open GitHub repository[1]. Hagenlocher et al. (2013) provided a methodology for the spatial assessment of socioeconomic vulnerabilities to dengue in Cali, Colombia, and followed the MOVE framework (Methods for the Improvement of Vulnerability Assessment in Europe) to obtain susceptibility and lack of resilience indicators from the census (Birkmann et al. 2013). We included these indicators in the dataset. We downloaded shapefiles from an ArcGIS online repository from DANE[2], containing the different spatial polygons at the municipality level (admin 2) and other socioeconomic indicators.

We obtained climate variables from two different data sources. The first one was *static*, consisting of multiannual values averaged for the 1981-2010 period for temperature and precipitation, which came from IDEAM's (Institute of Hydrology, Meteorology and Environmental Studies) Climatological Atlas of Colombia (IDEAM, 2015). The second one was *dynamic*, consisting of historical monthly temperature and precipitation values from 2007 to 2018, which we obtained from the CRU-TS 4.03 dataset (Harris et al., 2014) downscaled with WorldClim 2.1 (Fick and Hijmans, 2017). We also obtained elevation data from the WorldClim 2.1 dataset, derived from the Shuttle Radar Topography Mission (SRTM). We performed a spatial join with the DANE National Geostatistics Framework to obtain the average value per municipality, either multiannual or monthly average.

We extracted the images from open data sources such as satellite imagery from Google or Apple. We select these datasets as they are publicly available and provide better spatial resolution (~5 m) with the caveat of being processed for visual or scientific commercial use and not having higher spectral (RGB only) nor temporal resolution compared to acquiring the image directly from MODIS or Landsat. Tile images had a broader coverage of the landscape (around 1.3km^2) and required post-processing for cloud identification (although these already processed imagery had less percentage of clouds). All the imaging data is hosted on Google Cloud through LCP academic licenses. We used a pre-trained classifier (DenseNet 121) on the *"Understanding the Amazon from Space"* Kaggle competition[3] to obtain a possibility score per tile for the following features: agriculture, primary forest, roads, habitation, bare-ground, and waterway.

[1] https://github.com/MITCriticalData-Colombia/Dengue-MetaData
[2] https://dane.maps.arcgis.com/home/item.html?id=bb63fb9c6bc84cbeb2399e77ff20f504
[3] https://www.kaggle.com/c/planet-understanding-the-amazon-from-space

We followed the FAIR principle for scientific data management and stewardship (Wilkinson et al. 2016). Data is findable, accessible, interoperable, and reusable. We included consistent identifiers and rich metadata describing our data to support findability, citation and reuse, and to provide the context for the interpretation of the data and results. We uploaded data to trusted repositories, both GitHub and Google Cloud (for the tile images), and we provided accurate information on data provenance and included the codes used for data retrieval and processing aiming for future reusability, reproducibility and further research and educational purposes. As we are presenting health estimates, we also followed the GATHER checklist (Stevens et al. 2016).

Statistical methods: disease mapping

The proposed approach for integrating metadata and image modelling to predict cases (health outcome counts) over a set of disjoint geographical areas in the country, was to perform disease mapping of dengue cases to obtain relative risk (RR) estimates for each study area. We included the whole features as covariates per municipality, to smooth over covariate space and evaluate whether the spatial variability in risk is attributed to either a specific covariate or the spatial random effects. To provide more reliable estimates in each of the municipalities, we smoothed the RR using hierarchical/random effects models that use the data from the totality of areas. We first fit a classic three-stage (Poisson-lognormal non-spatial random effect) model as follows:

Stage 1. The Likelihood $Y_i|\theta_i \sim Poisson(E_i\theta_i)$, i = 1, …, n with

$$Log\,\theta_i = \beta_0 + x_1\beta_1 + \cdots x_p\beta_p + e_i$$

With area-specific independent random effects e_i, that capture the residual or unexplained (log) relative risk of the event in area i, and E_i as the expected number of cases in municipality i, with $E_i = \frac{\sum_i y_i}{\sum_i pop_i} \times pop_i$, where pop_i is the population size of municipality i.

The covariates $x_{1,\ldots,}x_p$ are image, climate and socioeconomic data. We fitted the model with and without the covariates to compare the results.

Stage 2. The random effects (prior distributions) is $e_i|\sigma_e^2 \sim_{iid} N(0, \sigma_e^2)$. With a single parameter controlling the spread of the random effects, σ_e^2.

Stage 3. The hyperprior on the hyperparameters $\beta_0, \beta_1, \ldots \beta_p, \sigma_e^2$:

$$p(\beta_0, \beta_1, \ldots \beta_p, \sigma_e^2) = p(\beta_0)\, p(\beta_1) \ldots\ p(\beta_p) p(\sigma_e^2)$$

Here we have assumed independent priors and use the integrated nested Laplace approximation (INLA) computational approach (Bakka et al. 2018) implementation in the R-package (R-INLA). INLA is an approximate method for Bayesian inference for latent Gaussian models developed by Rue et al. (2009).

Our final (Spatial) model follows the convolution prior model by Besag, York and Mollié (1991), also known as *BYM*, wherein an independent (across areas) random effect is added to the equation:

$$\eta_i = \beta_0 + x_i^T \beta + e_i + S_i$$

With independent random effects e_i and intrinsic conditional autoregressive (iCAR) random effects S_i. The BYM model uses both spatial and non-spatial error terms to account for over-dispersion not modelled by the Poisson variables. Distribution of S_i, the spatial structured residual, is modelled using iCAR with values taken by the neighboring random variables: areas which share boundaries with the *i*-th one. To define neighbors, we used the *poly2nb* function in R from the *spdep* package, that builds a neighbors list based on regions with contiguous boundaries, that is sharing one or more boundary point. We parameterized (i.e., *BYM2* model within INLA (Simpson et al. 2017), that follows a *Penalized Complexity* framework), in terms of total variance (σ_b^2), with $b_i = e_i + S_i$, and proportion that is spatial (ϕ), and placed a penalized complexity prior on these two parameters.

We fit three different models with the covariates. Model 1 with image data alone, Model 2 with image and climate data, and Model 3 with all the covariates: image, climate and socioeconomic. We measured the model performance using a deviance information criterion (DIC) and Watanabe-Akaike information criterion (WAIC), to find the one with better prediction quality. For the visualization of results, we calculated first the empirical averages and map the raw relative risk (RR), accompanied by a plot of raw RR values and the corresponding standard error. Then we fit the non-spatial random effects model (Poisson-lognormal, IID) and plot the resulting posterior median. To finalize, we fit the ICAR+IID spatial model and added the covariates and

map the final posterior medians and the residuals. Posterior medians for β_0, σ_e and ϕ are reported, with their respective 95% credible intervals. For the scales, we use the projection EPSG:3118 (MAGNA-SIRGAS).

We used R (version 3.6.1) for the analysis and visualization. We merged the shape files and the dengue data using the same municipality code, performing a minor change to the variables in order for them to match accordingly. We used choropleth maps for visualization.

Results

Team formation and Data Repository

At this writing, the project has recruited more than 15 members, global researchers from different institutions located both at Colombia and the U.S., representing the academia and public sector (S1 Appendix). The group includes clinicians, public health experts, machine learning scientists and data engineers, that apply their methodological and domain expertise. Accredited universities, public and private, located at one of the largest metropolitan area, as well as a recognized university in a region, are part of the partnership, with the intent to decentralized research in the country, another aspect that global health data science projects should aim to. Among the stakeholders, policymakers and civil society have been included as key members in the conversations for policy action. The data repository is stored and published using GitHub, where both open, machine readable datasets and the code to download and/or curate the data is included (summary of the data and data sources at S2 Appendix). The code for the disease mapping is also extensively documented and uploaded as a tutorial to GitHub, explaining the statistics and the methodology followed.

Disease mapping in the Department of Antioquia for the 2019 Outbreak

Descriptive Statistics

The Department of Antioquia has 125 municipalities and a total population of 6,550,206, with 5,812 dengue cases registered for the outbreak year 2019. Figure 1 shows the relative risk estimates (raw RR, left figure, smoothed right figure) and indicates a larger spread with an increasing trend in the south-north direction (and even from the center to the periphery). For this data, the expected cases E_i are highly skewed, with a range 6.95-6,377.33 and median of 40.93. Since the variance of the estimator will be larger if E_i is small, the variability of E_i suggest that the extreme RR may be based on small expected numbers. Figure 2 shows the raw RR values and the standard error, showing how the highest RR tend to have the largest standard errors.

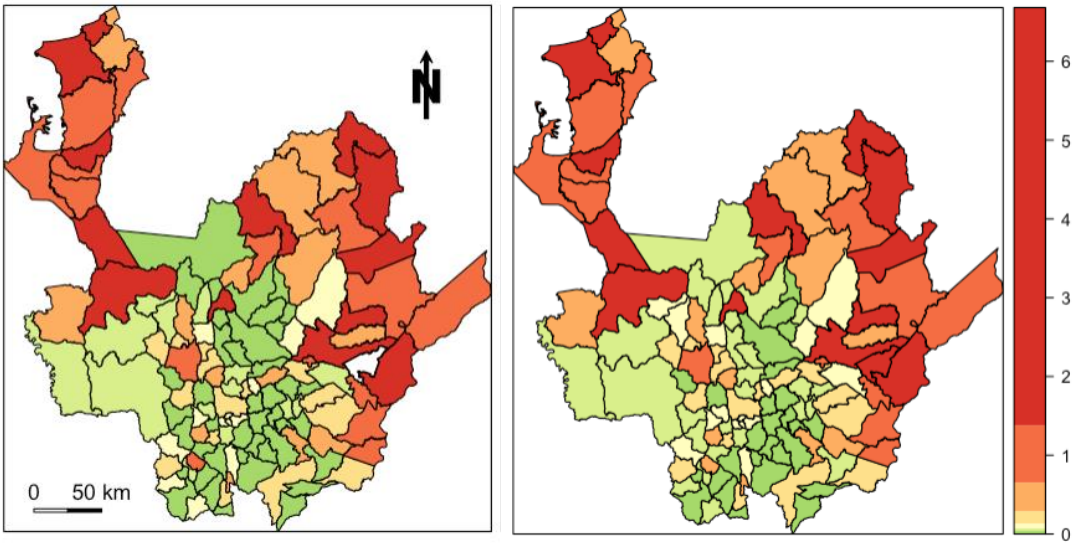


Figure 1. Observed (left) and smoothed (right) RR estimates in 125 municipalities of Antioquia. Smoothed values correspond to the posterior median of the BYM2 model (Poisson-LogNormal-Spatial). The color palette was used to differentiate the changes in RR (nine quantile style breaks using the observed RR). Smoothing is clearer at the Municipalities that had zero reported cases (observed RR=0); therefore, the right figure has less Municipalities colored with dark green, and the left figure more well-defined regions. On the other hand, one Municipality had a very high observed RR estimate (i.e., an outlier, white spot) but could be mapped in the smoothed figure.

We considered the alternative lognormal model for the relative risk, but still independent (Poisson-lognormal, IID), specifying that there is a 5% chance that the standard deviation σ_e is greater than 1, and plotted the posterior median estimates against the observed RR. Figure 3 (left) shows the "shrinkage" of the Bayes estimates, particularly at the low values of RR. The posterior median for β_0 is exp(-2.23)= 0.107 (credible intervals: 0.072-0.15), and for σ_e is $\frac{1}{\sqrt{0.26}} = 1.92$ (credible intervals: 1.65-2.26).

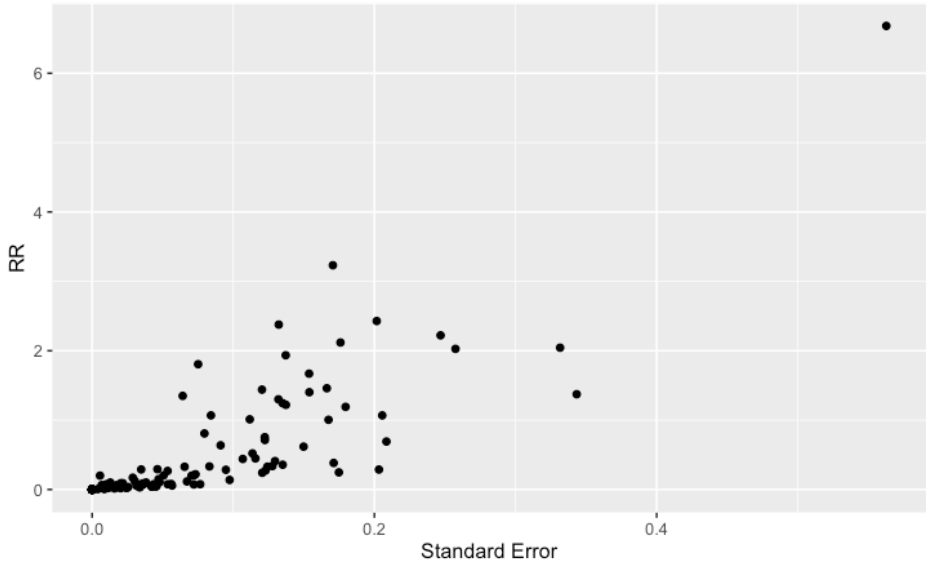


Figure 2. Relative risk and standard error in 125 municipalities of Antioquia.

For the ICAR+IID spatial model (BYM2), the choices of the parameters correspond to the prior belief that there is a 1% chance that the total residual standard deviation is greater than 0.3, and a 50% chance that the proportion of the variance that is spatial is bigger than 0.5. Figure 1 (right) shows the posterior median RR values and how municipalities with an observed RR equal to zero, increased their value after the smoothing. Nevertheless, there are still some municipalities with a low RR, but concentrated near the center. On the other hand, a few municipalities with a high observed RR, though closer to municipalities with a low RR, "shrink", and the posterior median is now lower. Figure 3 (right) shows that "shrinkage" after the BYM smoothing with independent and ICAR random effects, particularly at the low values of RR, and the difference to the Poisson-Lognormal-non-spatial model. The posterior median for β_0 is exp(-2.12)= 0.1198 (credible intervals: 0.095-0.14), and the posterior median of the total standard deviation (on the log relative risk scale) is $\frac{1}{\sqrt{0.46}} = 1.47$ (credible intervals: 1.27-1.71), and the posterior median for the proportion of the residuals that is spatial is 0.54 (credible interval: 0.30-0.78).

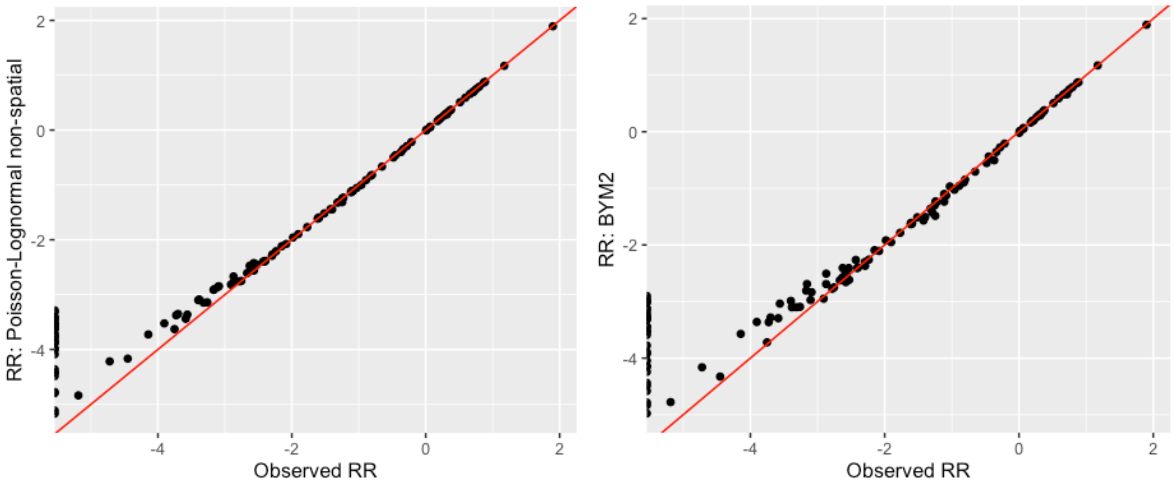


Figure 3. Log-Posterior median of the RR from Poisson-LogNormal non-spatial (left) and BYM2 (right) model estimates in 125 municipalities of Antioquia, compared to the observed RR.

We added the covariates and repeated the three-stage (Poisson-lognormal non-spatial random effect) model as well as the BYM2. Table 1 presents the descriptive statistics for the covariates and Table 2 the model selection criteria with the performance metrics DIC and WAIC. Model 3 had better prediction quality. Fitting a quasi-likelihood model with the original counts, the whole set of covariates (the same as Model 3) and the naive log-linear ecological regression model gave the following significant estimates: $\beta_{Waterway} = \exp(-0.074) = 0.927$ ($p < 0.05$), $\beta_{Temperature} = \exp(0.203) = 1.22$ ($p < 0.001$), and $\beta_{Hospitalsperkm2} = \exp(0.74) = 2.09$ ($p < 0.05$). Therefore, we would expect that an area whose waterway is one unit higher has a 7.3% lower relative risk, whose temperature is one unit higher has a 22% higher relative risk, and whose number of hospitals per km2 increase by one has a 2-fold increase in the relative risk. Nevertheless, we have to realize that these are not individual-level associations (i.e., ecological fallacy), and that the higher RR risk with the number of hospitals per km2 could be related to an increased rate of *detecting* dengue, nor disease transmission.

Table 1. Descriptive statistics for covariates

	Mean	SD	Min	Q1	Median	Q3	Max
Agriculture (%)	49.7	9.12	21.13	44.89	49.7	55.12	68.34
Bare ground (%)	3.92	0.84	1.81	3.53	3.92	4.32	6.62
Habitation (%)	37.28	14.6	10.5	27.8	36.37	43.43	81.81

Primary (%)	94.23	1.5	83.2	93.71	94.23	95.09	97.18
Roads (%)	34.32	6.91	18.66	29.73	34.32	39.04	49.96
Waterway (%)	26.83	7.84	10.4	21.62	26.02	30.4	50.43
Elevation (m.a.s.l.)	1386	741	32	930	1503	1970	2883
Temperature (°C)	20.01	4.01	14	18	18	21	29
Rainfall (mm)	2744	824.56	1750	2250	2750	2750	6000
Age 0-4 y/o (%)	6.52	1.58	3.82	5.55	6.2	7.47	13.4
Age 5-14 y/o (%)	16.63	3.69	9.27	14.01	16.14	18.6	28.78
Houses without potable water access (%)	18.56	14.35	0.21	7.17	15.91	27.2	86.4
Number of hospitals per km2	0.041	0.1795	0	0.0023	0.0046	0.0104	1.5553
Unemployed population (%)	3.38	1.95	0.82	2.19	2.87	4.23	13.52
Secondary/Higher Education (%)	45.11	10.2	26.26	37.81	42.83	50.12	77.33

Table 2. Model selection criteria for the best fitted models in INLA: deviance information criterion (DIC) and Watanabe-Akaike information criterion (WAIC). Model 1 with image data alone, Model 2 with image and climate data, and Model 3 with all the covariates: image, climate and socioeconomic.

	Model 1	**Model 2**	**Model 3**
DIC	661.96	655.61	**651.32**
WAIC	647.26	639.01	**633.26**

As we were interested in smoothing over covariate space and generate hypothesis with informal comparison of risk maps with exposure maps, we also performed an ecological correlation study. We examined the posterior summaries for the non-spatial random effects model on the exponentiated scale, finding that the estimate for waterway, temperature and hospitals per km2 were 0.950 (CI: 0.89-1.013), 1.129 (CI: 0.993-1.285) and 1.046 (CI: 0.201-5.367), respectively. Waterway was similar, temperature had a reduction of 10% and hospitals per km2 had an important reduction of almost half compared to the previous result. Examination of these intervals helped us determining whether the association is “significant”, and here we did not have strong evidence that the RR is associated with these covariates. The posterior median of σ_e is $\frac{1}{\sqrt{0.60}} = 1.29$ and a 95% interval (1.08-1.55). A more interpretable quantity, the interval on

the residual relative risk that follows a LogNormal(0, $\sigma_e{}^2$), is a 95% interval $\exp(\pm 1.96 \times \sigma_e) = [0.079, 12.53]$, which is quite wide.

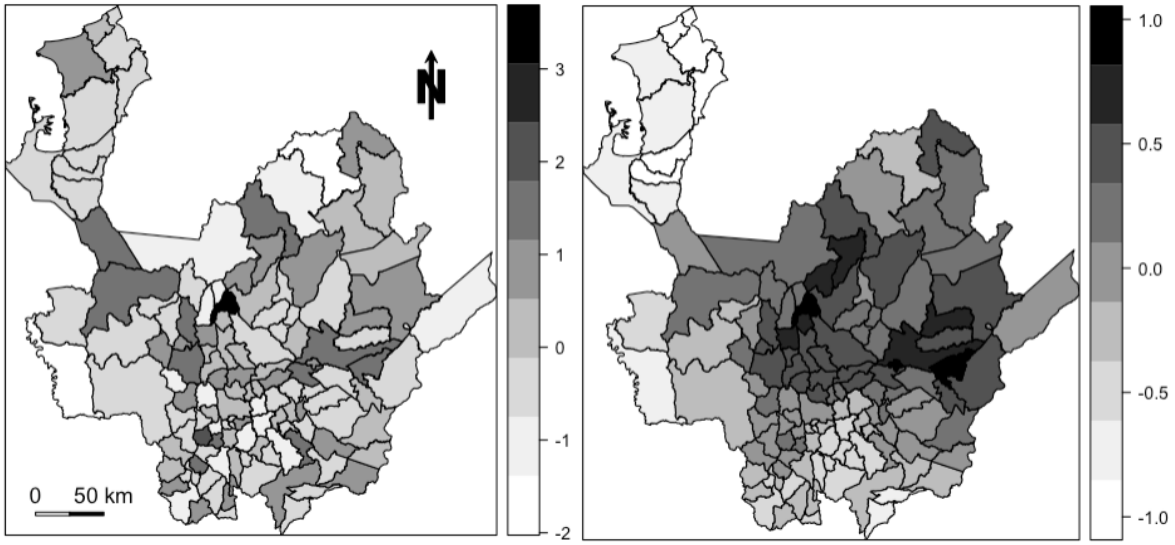


Figure 4. Map of posterior medians of non-spatial e_i (left) and spatial S_i (right) random effects for the LogNormal spatial model with the whole set of covariates (Model 3) in 125 municipalities of Antioquia.

Once we added the spatial (ICAR) random effects to the model with the covariates, we found that the posterior median of the total standard deviation (on the log relative risk) is $\frac{1}{\sqrt{0.58}} = 1.31$ (credible interval: 1.08-1.59), and the posterior median for the proportion of the residual variation that is spatial is 0.27 (credible interval: 0.04-0.71). Note that the posterior mean estimates of β (Table 3) associated with the previously studies covariates did not reduce significantly when moving from the non-spatial to spatial model, not showing confounding by location. Surprisingly, we found that the covariates waterway and elevation showed strong evidence of association with RR after examining the credible intervals. Finally, the map for the non-spatial and spatial random components of the log residual relative risk (e_i) can be found in Figure 4, where we can observe a similar scale between them, indicating that it is not clear whether the spatial component dominates or not in this data. We can observe also how the posterior median estimate of σ_e was reduced from 1.92 (Poisson-LogNormal) to 1.31, when the spatial random effect and covariates are added.

Table 3. Summary statistics for the parameter estimates of Model 3 for the relative risk of dengue. Posterior mean (standard deviation) together with 2.5% and 97.5% quantiles.

Covariate	**Mean**	**SD**	**2.5%**	**97.5%**
$exp(\beta_{Agriculture})$	0.989	1.030	0.932	1.048
$exp(\beta_{BareGround})$	1.308	1.289	0.794	2.157
$exp(\beta_{Habitation})$	1.005	1.023	0.962	1.050
$exp(\beta_{Primary})$	1.071	1.123	0.859	1.357
$exp(\beta_{Roads})$	0.987	1.035	0.923	1.057
$exp(\beta_{Waterway})$	0.929	1.036	**0.866**	**0.995**
$exp(\beta_{Elevation})$	0.999	1.000	**0.998**	**0.999**
$exp(\beta_{Temperature})$	1.122	1.071	0.980	1.284
$exp(\beta_{Rainfall})$	1.000	1.030	0.932	1.000
$exp(\beta_{Age0\text{-}4})$	0.940	1.273	0.588	1.521
$exp(\beta_{Age5\text{-}14})$	0.980	1.151	0.740	1.287
$exp(\beta_{HousesWithoutWater})$	1.017	1.016	0.986	1.049
$exp(\beta_{Hospitalsperkm2})$	1.067	2.232	0.217	5.122
$exp(\beta_{UnemployedPopulation})$	1.056	1.084	0.900	1.238
$exp(\beta_{Secondary/HigherEducation})$	1.044	1.029	0.989	1.106

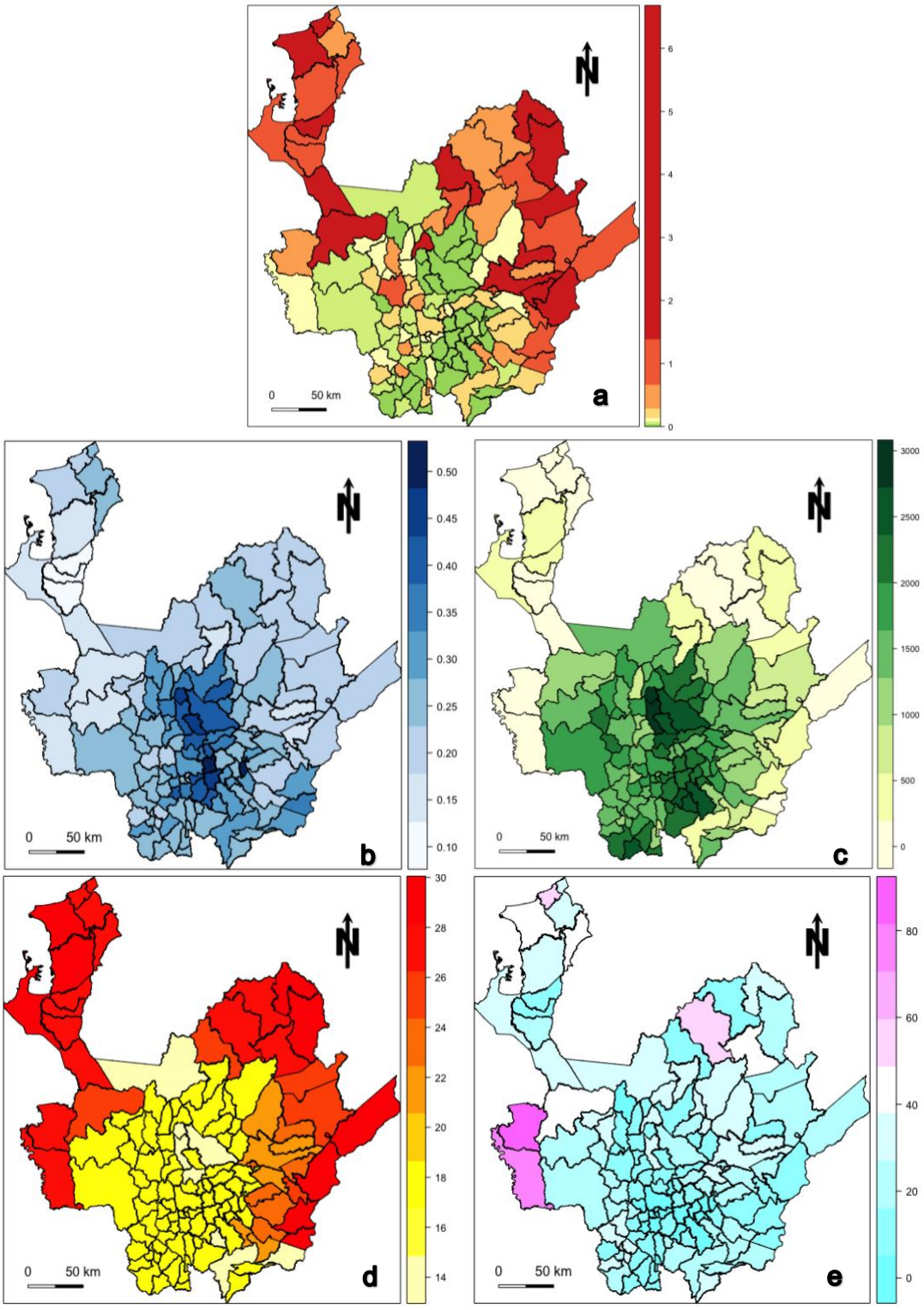


Figure 5. Map of posterior median relative risks (a) for the 125 Municipalities for the Department of Antioquia after smoothing the data with the Poisson-LogNormal spatial model, with shrinkage particularly at low RR values. Two covariates, waterway (b) and elevation (c) have strong evidence of a negative

association with RR. On the other hand, the covariates temperature (d) and houses without water (e) have a positive association with log(RR), although in a subsequent analysis the credible intervals included the RR value of 1, therefore no strong evidence for association was found.

Discussion

Data science projects in global health require a multidisciplinary approach that involves early stakeholder partnerships, political action and FAIR data repositories. This study is innovative since it includes ML-based satellite image processing for dengue modelling and presents a methodology that is explainable for future ML approaches in the field. We provided a dengue relative risk mapping tool as the initial result of a multidisciplinary collaboration. In particular, there is an open repository of both data and codes and an explainable model that maps dengue outbreaks, accounting for geographical, environmental, and socioeconomic factors, and considers future deployment and implementation.

There are well-documented difficulties with the mapping of raw disease estimates since, for small areas and rare events in particular, these estimates will be dominated by sampling variability. Likewise, if data are sparse in an area, averages and totals are unstable because of the small denominators. This study performs disease mapping of dengue cases in the Department of Antioquia, Colombia, for the outbreak year of 2019. As Wakefield (2007) mentioned, surveillance is greatly helped with disease mapping as we could know the distribution of the disease in the absence of a “hot spot”, by including the information about the variability in residual spatial risk and the nature of that variability (spatial versus nonspatial). The smoothed posterior median RR estimates after a multi-stage, hierarchical modeling, represent a reliable estimate as it uses the totality of the data to inform on the distribution, both locally and globally. Many of the large, sparsely populated areas in the north (and periphery) of Antioquia have high RR. Conversely, those municipalities in the central region, close to Medellin (the capital), have low RR. After smoothing the RR using the random-effects models, we found that municipalities close to these areas exhibited a similar behavior in terms of RR estimates.

Although the “shrinkage” means that hierarchical models are not good for “hot spot” detection and each area’s estimate is biased (in a frequentist sense), it is important to understand that “zero” cases of local transmission of dengue in certain geographical areas could be an anomalous data point, and therefore the use of disease mapping for providing better estimates (with lower variance). From an ecological context, *A. aegypti* is the principal mosquito vector spreading dengue viruses in Colombia (with a 90% presence up to 2,300m), is well adapted to urban life and an efficient epidemic vector (Rocklöv and Dubrow 2020). There are municipalities where the presence of *A. aegypti* is expected and imported cases due to high population

movement is facilitating the transmission. From a diagnostic context, the lack of access to healthcare services requires patients to travel to the capital or major towns to get diagnosed, with some cases are missed in that process. From a symptom's context, the presence of other arboviruses could facilitate misclassification from the medical personnel and inapparent disease is likely to happen in younger individuals experiencing their first dengue virus infection (WHO, 2018). Finally, the epidemiological history of cases could inform us of previous reported cases at these municipalities, where a "zero" value is very likely related to bad data quality.

There was no clear confounding by location found when including the covariates in the model (estimates did not vary considerably), indicating that the model attributes spatial variability in risk to the covariates more than to the spatial random effects. Figure 5 presents the mapping of the smoothed posterior median RR estimates and four selected covariates (waterway, elevation, temperature and houses without water), using a visualization scheme that highlights areas above the RR=1. We can easily observe the impact of the outbreak in a considerable number of municipalities in Antioquia, and the visual relationship of such municipalities with the covariates. For instance, municipalities with a temperature above 22 °C have a tendency to have RR higher than 1. Also, the elevation map shows the high variability in terms of this covariate, with a range of 32 to 2883 m.a.s.l. (meters above the sea level) along the entire Department of Antioquia. Although the proportion of the residuals that is spatial is considerably reduced by half from 54 to 27% from the LogNormal spatial model without covariates to the Model 3 with covariates, spatial location is still acting as an important surrogate for unobserved covariates, not surprising as dengue is a complex disease with many other risk factors.

Limitations of the study dataset is that reported cases in SIVIGILA correspond to symptomatic patients who present at healthcare centers and are registered with a residence addresses that do not correspond necessarily with the places of transmission. In addition, the SIVIGILA database does not contain specific serotype infection information, another important factor for disease surveillance. Symptomatology can change depending on mechanisms arising from infection with heterologous viral serotype (Nealon et al. 2020), and there is a clear differential healthcare access in the country. Altogether, underreporting of cases on surveillance data is common, and Colombia is specially challenging due to potential misclassification bias of arboviral diagnostics (Carabali et al. 2021). Data can also be biased by age and sex (Silva et al. 2016), and indigenous health data might be missing. Entomological data was requested to the National Institute of Health but due to the COVID-19 pandemic, the response has been highly

delayed. It has been included to the policy advocacy plan designed with the partners in the country (S3 Appendix). Regarding the use of the naïve Poisson model, it is important to consider extra-Poisson variability resulting from unmeasured confounders, data anomalies in numerator and denominator, and model misspecification (Wakefield, 2007). Also, there is always a trade-off when a geographical scale is chosen. In the case of Departments (admin 1), which are larger geographical areas, it provides more stable rates and less problems of migration, but the relative risk summaries may be distorted due to the large aggregation of individuals.

The status quo to dengue vulnerability mapping is the Water Disease-Associated Index (WADI) proposed by Dickin et al. (2013), that integrates a range of social and biophysical determinants in a map format. This tool was designed to provide stakeholders with a long-term understanding of subnational vulnerabilities and claims to be pragmatic and not geographically constrained due to data availability, a limitation that many risk models and early warning systems have (Racloz et al. 2012). However, one of its main components, land environment, heavily depends on the data source. On the other hand, ML models in public health are well-known for yielding a limited understanding of their results (i.e., black boxes), that limits their deployment and implementation (Wiemken and Kelley, 2020). A solution is explainable ML. Explainable ML for computer vision as the data modality (e.g., DNN to satellite images), are post hoc explanation methods with feature importance approaches (Linardatos et al. 2021). When structured data is used, the methods also include feature importance, but saliency maps and gradient-based methods are not very meaningful. We provided an initial ML-based model that includes an explainable mechanism with both data modalities: computer vision and structured data. Satellite image processing generates DNN-derived data that, along with structured environmental and socioeconomic metadata, are entered as covariates to the spatial model. This spatial model does not represent any black box for decision-makers, facilitating its deployment and implementation, opening new horizons in the use of these approaches for population and public health research. Once more ML-based eco-epidemiological models for dengue that includes satellite images are developed, public health researchers can translate the technique to other climate sensitive diseases where landscape or image-based features might have an impact on the model performance. Moreover, public health decision-makers can use our pragmatic model to classify areas of high risk for dengue with data that is more readily available and that requires fewer resources to be obtainable.

Future studies could include mapping the posterior probabilities, indicating when the RR in each area exceeds certain thresholds (Wakefield, 2017), to evaluate areas with high probabilities and discover characteristics of the individuals or health hazards. Also, sensitivity analysis is needed. For instance, changing the E_i or expected number of cases in municipality *i*, for a more data-supported number, using the average number of the interepidemic years of the same dataset (2015-2018). Another sensitivity analysis could be performed with the modification of priors and the parameters used for the BYM2 model specification. Lastly, the choice of the model for covariates, like the large-term spatial trend modelling, could also modify notably the results.

For a successful completion of the study, deployment and implementation of the model in Early Warning Systems in Colombia is required. Two additional aims are in the works and consist of a qualitative research to understand barriers and facilitators for implementation, and the execution of an implementation strategy of learning-collaborative through of datathon in Colombia (Aboab et al. 2016), to continue building a cross-disciplinary community for machine learning in public health in the country. We then expect to have a potentially important impact in population and public health research, particularly at having accurate and real-time information on outbreak risk/vulnerability to inform policies for prevention and control.

Conclusions

This study presents the potential to advance the field of eco-epidemiology in global health by leveraging a multi-stakeholder collaboration for data science that provides open data repositories and explainable ML-based models. As an initial approach for disease mapping, we included deep landscape features from satellite images as predictors along with climate and socioeconomic variables in eco-epidemiological models for dengue transmission, with the aim of reducing the black-box nature of deep learning in public health and therefore increasing the likely of adoption by stakeholders. It is intended that the ML-based ecoepidemiology model of dengue will provide accurate and in-time information on disease surveillance at the municipality level in Colombia to public health authorities' decision-making. This information could notably support planning intervention strategies that help identify and mitigate the risks related to dengue infection. It would also allow efficient allocation of scarce resources and facilitate timely response strategies with precise geographic targeting during outbreaks. Finally, the COVID-19 pandemic has shown us the importance of data and knowledge sharing (Cosgriff et al. 2020). The climate crisis is also a reality, dengue and other climate-sensitive diseases should be approached with the same determination we have done for COVID-19: joining forces and data, understanding the event but also what works and what doesn't, is a priority.

References

Abdur Rehman, N., Saif, U. and Chunara, R., 2019. Deep landscape features for improving vector-borne disease prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 44-51).

Aboab, J., Celi, L.A., Charlton, P., Feng, M., Ghassemi, M., Marshall, D.C., Mayaud, L., Naumann, T., McCague, N., Paik, K.E. and Pollard, T.J., 2016. A “datathon” model to support cross-disciplinary collaboration. *Science Translational Medicine, 8*(333), pp.333ps8-333ps8.

Bakka, H., Rue, H., Fuglstad, G.A., Riebler, A., Bolin, D., Illian, J., Krainski, E., Simpson, D. and Lindgren, F., 2018. Spatial modeling with R-INLA: A review. *Wiley Interdisciplinary Reviews: Computational Statistics, 10*(6), p.e1443.

Besag, J., York, J. and Mollié, A., 1991. Bayesian image restoration, with two applications in spatial statistics*. Annals of the institute of statistical mathematics, 43*(1), pp.1-20.

Birkmann, J., 2006. *Measuring vulnerability to natural hazards: towards disaster resilient societies* (No. 363.34 M484m). New York, US: United Nations University Press.

Birkmann, J., Cardona, O.D., Carreño, M.L., Barbat, A.H., Pelling, M., Schneiderbauer, S., Kienberger, S., Keiler, M., Alexander, D., Zeil, P. and Welle, T., 2013. Framing vulnerability, risk and societal responses: the MOVE framework. *Natural hazards*, *67*(2), pp.193-211.

Carabali, M., Jaramillo-Ramirez, G.I., Rivera, V.A., Mina Possu, N.J., Restrepo, B.N. and Zinszer, K., 2021. Assessing the reporting of Dengue, Chikungunya and Zika to the National Surveillance System in Colombia from 2014–2017: A Capture-recapture analysis accounting for misclassification of arboviral diagnostics*. PLoS neglected tropical diseases, 15*(2), p.e0009014.

Cosgriff, C.V., Ebner, D.K. and Celi, L.A., 2020. Data sharing in the era of COVID-19. *The Lancet Digital Health*, *2*(5), p.e224.

Cutter, S.L., Boruff, B.J. and Shirley, W.L., 2003. Social vulnerability to environmental hazards. *Social science quarterly, 84*(2), pp.242-261.

DANE 2018. *Censo Nacional de Población y Vivienda 2018*, Departamento Administrativo Nacional de Estadística, viewed 14 March 2021, <https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/censo-nacional-de-poblacion-y-vivenda-2018>.

Dickin, S.K., Schuster-Wallace, C.J. and Elliott, S.J., 2013. Developing a vulnerability mapping methodology: applying the water-associated disease index to dengue in Malaysia. *PLoS One, 8*(5), p.e63584.

Ebi, K.L. and Nealon, J., 2016. Dengue in a changing climate. *Environmental research, 151*, pp.115-123.

Fick, S.E. and Hijmans, R.J., 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International journal of climatology*, 37(12), pp.4302-4315.

Flaxman, A.D. and Vos, T., 2018. Machine learning in population health: opportunities and threats. *PLoS medicine*, 15(11), p.e1002702.

Fuentes-Vallejo, M., 2017. Space and space-time distributions of dengue in a hyper-endemic urban space: the case of Girardot, Colombia. *BMC infectious diseases, 17*(1), pp.1-16.

Füssel, H.M., 2007. Vulnerability: A generally applicable conceptual framework for climate change research*. Global environmental change, 17*(2), pp.155-167.

Gutierrez-Barbosa, H., Medina-Moreno, S., Zapata, J.C. and Chua, J.V., 2020. Dengue Infections in Colombia: Epidemiological Trends of a Hyperendemic Country. *Tropical Medicine and Infectious Disease, 5*(4), p.156.

Hagenlocher, M., Delmelle, E., Casas, I. and Kienberger, S., 2013. Assessing socioeconomic vulnerability to dengue fever in Cali, Colombia: statistical vs expert-based modeling*. International journal of health geographics, 12*(1), pp.1-14.

Harris, I.P.D.J., Jones, P.D., Osborn, T.J. and Lister, D.H., 2014. Updated high-resolution grids of monthly climatic observations–the CRU TS3. 10 Dataset. *International journal of climatology*, *34*(3), pp.623-642.

Hess, J., Boodram, L.L.G., Paz, S., Ibarra, A.M.S., Wasserheit, J.N. and Lowe, R., 2020. Strengthening the global response to climate change and infectious disease threats. *bmj, 371*.

IDEAM 2015. *Atlas Interactivo - Climatológico – IDEAM*, atlas.ideam.gov.co, viewed June 3 2021, <http://atlas.ideam.gov.co/visorAtlasClimatologico.html>.

Jean, N., Burke, M., Xie, M., Davis, W.M., Lobell, D.B. and Ermon, S., 2016. Combining satellite imagery and machine learning to predict poverty. *Science, 353*(6301), pp.790-794.

Lai, Y., Yeung, W. and Celi, L.A., 2020. Urban intelligence for pandemic response: viewpoint. *JMIR Public Health Surveill.* Apr 14; 6 (2): e18873. doi: 10.2196/18873.

Linardatos, P., Papastefanopoulos, V. and Kotsiantis, S., 2021. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy, 23*(1), p.18.

Maharana, A. and Nsoesie, E.O., 2018. Use of deep learning to examine the association of the built environment with prevalence of neighborhood adult obesity. *JAMA network open, 1*(4), pp.e181535-e181535.

Nealon, J., Bouckenooghe, A., Cortes, M., Coudeville, L., Frago, C., Macina, D. and Tam, C.C., 2020. Dengue endemicity, force of infection and variation in transmission intensity in 13 endemic countries. *The Journal of infectious diseases*.

PAHO 2021, Dengue - *PAHO/WHO* | *Pan American Health Organization*, PAHO Topics, viewed 10 March 2021, <https://www.paho.org/en/topics/dengue>.

Parselia, E., Kontoes, C., Tsouni, A., Hadjichristodoulou, C., Kioutsioukis, I., Magiorkinis, G. and Stilianakis, N.I., 2019. Satellite earth observation data in epidemiological modeling of malaria, dengue and west nile virus: a scoping review. *Remote Sensing, 11*(16), p.1862.

Racloz, V., Ramsey, R., Tong, S. and Hu, W., 2012. Surveillance of dengue fever virus: a review of epidemiological models and early warning systems. *PLoS Negl Trop Dis*, *6*(5), p.e1648.

Rocklöv, J. and Dubrow, R., 2020. Climate change: an enduring challenge for vector-borne disease prevention and control. *Nature immunology, 21*(5), pp.479-483.

Rue, H., Martino, S. and Chopin, N., 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology), 71*(2), pp.319-392.

Shearer, C., 2000. The CRISP-DM model: the new blueprint for data mining. *Journal of data warehousing*, *5*(4), pp.13-22.

Shiffman, J. and Smith, S., 2007. Generation of political priority for global health initiatives: a framework and case study of maternal mortality. *The lancet*, *370*(9595), pp.1370-1379.

Simpson, D., Rue, H., Riebler, A., Martins, T.G. and Sørbye, S.H., 2017. Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical science*, pp.1-28.

Silva, M.M., Rodrigues, M.S., Paploski, I.A., Kikuti, M., Kasper, A.M., Cruz, J.S., Queiroz, T.L., Tavares, A.S., Santana, P.M., Araújo, J.M. and Ko, A.I., 2016. Accuracy of dengue reporting by national surveillance system, Brazil. Emerging infectious diseases, 22(2), p.336.

Stevens, G.A., Alkema, L., Black, R.E., Boerma, J.T., Collins, G.S., Ezzati, M., Grove, J.T., Hogan, D.R., Hogan, M.C., Horton, R. and Lawn, J.E., 2016. Guidelines for accurate and transparent health estimates reporting: the GATHER statement. *PLoS medicine*, *13*(6), p.e1002056.

Wakefield, J., 2007. Disease mapping and spatial regression with count data. *Biostatistics, 8*(2), pp.158-183.

Wiemken, T.L. and Kelley, R.R., 2020. Machine learning in epidemiology and health outcomes research. *Annual review of public health*, *41*, pp.21-36.

World Health Organization 2012, *Global strategy for dengue prevention and control 2012-2020*, WHO, viewed 10 March 2021, <https://www.who.int/immunization/sage/meetings/2013/april/5_Dengue_SAGE_Apr2013_Global_Strategy.pdf>

World Health Organization, 2018. *A toolkit for national dengue burden estimation* (No. WHO/CDS/NTD/VEM/2018.05). World Health Organization.

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V.X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M. and Ossorio, P.N., 2019. Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine, 25*(9), pp.1337-1340.

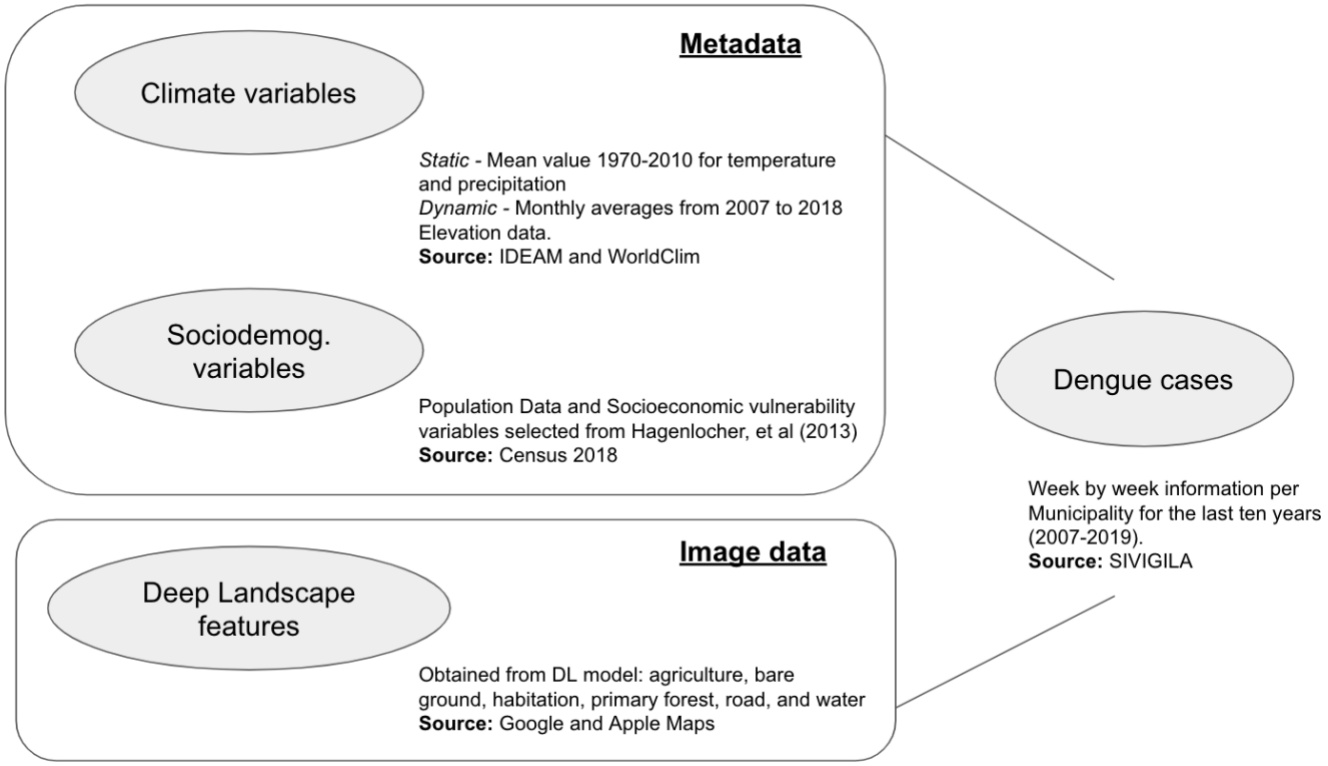
Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E. and Bouwman, J., 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*(1), pp.1-9.

Appendices

S1 Appendix. Collaborators

Name	Affiliation
Dr. Leo Celi MD MS MPH	Principal Research Scientist / Associate Professor, MIT / Harvard Medical School, USA
Dr. Ivan Darío Velez MD MSc PhD	Director of PECET, Professor, Universidad de Antioquia, Colombia
Dra. Maria Patricia Arbelaez-Montoya MD MPH PhD	Professor, Universidad de Antioquia, Colombia
Dr. Diego López MSc PhD	Professor, Universidad del Cauca, Colombia
Braiam Escobar, PhDc	Leader of the Center for Innovation and Digital Transformation at Universidad CES, Colombia
Dr. Cheng Che Tsai, MD MBI	Department of Biomedical Informatics at Harvard Medical School, USA
Dana Moukheiber	Master student at University at Buffalo, USA
Sebastian Andres Cajas Ordonez	Erasmus Joint Master student, Europe
Laura Sofía Daza Rosero	Student Electronics and Telecommunications Engineering, Universidad del Cauca, Colombia
Jhon Fredy Romero Núñez	Student Electronics and Telecommunications Engineering, Universidad del Cauca, Colombia
David Restrepo	Student Electronics and Telecommunications Engineering, Universidad del Cauca, Colombia
Luis Jesús Martínez	Data and GIS Coordinator at World Mosquito Program, PECET, Universidad de Antioquia, Colombia
Saketh Sundar	MIT Critical Data, USA
Alessa Álvarez, MBA	Operations and Management Support, WEF Centre for the Fourth Industrial Revolution, Colombia
Kavya Ravichandran	MEng Student at MIT Computer Science and Artificial Intelligence Laboratory (CSAIL), USA
Wilson Arbey Diaz	Environmental engineer. Student hydraulics resource management at Universidad Nacional de Colombia, sede Medellín.

S2 Appendix – Summary of data and sources



S3 Appendix. Policy Advocacy Plan.

The goal of this policy advocacy plan is to increase the commitment from the Colombian National Institute of Health (*Instituto Nacional de Salud* - INS) in releasing findable and accessible data for public health research, particularly dengue data. The policymaker with the authority to adopt this recommendation is the INS Director. The list of activities the advocacy team will engage to build the political priority of the recommended policy intervention follows Shiffman and Smith (2007) and are as follows:

Priority Factor	Description of Activity
1. Policy Community Cohesion	One of the project collaborators has joined the Colombian Radiology Association (ACR) Committee on Artificial Intelligence. This committee is having high-level meetings with directors at the Ministry of Health and other governmental agencies. It is also organizing the Congress in AI in health in the country and reviewing the regulatory environment to assure a responsible implementation of AI in the country. The message to the policy community is that responsible AI requires FAIR databases for training the models and that public health data should prioritized.
2. Leadership	Dr. Ivan Darío Velez, Director of PECET from University of Antioquia and another collaborator, is a world recognized leader in vector-borne diseases. The country leadership has him as an individual capable of uniting the policy community. PECET is also part of the World Mosquito Program, a non-for-profit initiative that is working to protect local communities from mosquito-borne diseases. This is not only important as a factor determining Dr. Velez as a champion for the cause, but an important aspect for global advocacy.
3. Guiding institutions	University of Antioquia (UdeA) is the institutional partner with more political power in the country. It is one of the major public institutions in Colombia, and one of the oldest and more internationally recognized for their academic achievements. Not only Dr. Velez is part of the University but Dra. Maria Patricia Arbelaez, another member of the project. She is a former vice-president for research at UdeA and another leader in the country, currently serving at the board of directors for the *Instituto Nacional de Salud* (our target institution). We might need to start a coordinating mechanism for open data in public health research, dealing with the FAIR principles and possible data bias.
4. Civil society	Participation of civil society at events organized by the ACR has been crucial to communicate the importance of open and heterogeneous data for better modelling using machine learning. We are also starting conversations with the Open Data for Development network (https://www.od4d.net/) which aims to foster change by working with multiple stakeholders and particularly under-represented communities.
5. Internal frame	The policy community agrees as this issue has been on the policy agenda for long time. We need to emphasize that the Law 1712 of 2014 (*Transparency and Access to National Public Information*), that standardizes data sharing and interoperability to facilitate access and use, fails in the implementation.
6. External frame	We could start framing climate sensitive diseases as threat to security, due to their exponential increase in the last 50 years and its close relationship with the climate crisis. Although the public portrayal of the issue is that Colombia is

	predestined to suffer dengue due the endemicity of the disease, it is actually a major concern that 80% of children become dengue seropositive at 8 y/o, providing a large pool of individuals at risk of secondary infection and therefore major risk of death.
7. Policy window	COVID-19 has opened a policy window to work on tackling infectious diseases in the country. Currently many governmental institutions are working on internal policies to deal with the pandemic, and collaboration among them are more feasible. We will facilitate such collaboration leading a commission for open data to inform infectious diseases in the country, as a way to prepare the country for future outbreaks.
8. Global governance structure	The CONPES 3072 (National Council for Economic and Social Policy) from 1999 aimed to build a new economy and the construction of a modern and efficient State. This CONPES lead to the Law 1712 for open data. These norms and the Ministry for ICT should provide a platform for effective collective action. Our advocacy is to align them and present to the INS their responsibility in fulfilling these regulations.
9. Credible indicators	We could improve the situation analysis and include further references for dengue in the country. So far, the data show the severity of the problem with more than 12x increase in the number of cases in the last 20 years, the very high incidence in the outbreak year of 2019, considering that the data collected that year might miss counting some asymptomatic patients, and the high case fatality compared to other countries in the America's region (4.8x higher). Besides, the idea with the novel modelling is to improve the health estimators and reduce the uncertainty within them.
10. Severity	Dengue is already considered in the country as one of the major public health problems. We should highlight even more the burden of the disease and its economic impact. Something that policymakers normally miss is the commentary from the people facing the disease on daily basis, both medical personnel and under-represented communities in hot spot areas.
11. Effective interventions	The proposed intervention is simple to implement and uses the resources from the Colombian government (there is no need to create a platform). It will require some time and specific skills for data management and curation, but the advocacy team is offering their services to support the joint commission of experts, providing training and helping with the data curation.