Grant Results Summary

Overview

the grant funding was used for two main purposes:

1. Sponsor an event in Bogotá, Colombia, focusing on the use of analytic tools in response to different humanitarian health crises at a national scale. This event was intended to socialize the different approaches and strategies taken by the local agencies and form a network of collaboration for future emergencies.
2. Design an analytic pipeline that combines different data sources to create coherent units of study for malaria at a national level. This tool is intended as an example of the kind of software tools that enable research to policy pathways.

Event

The event took place in Bogotá, Colombia on the 1st of August 2022. It was organized by Crisis Ready, Universidad Nacional de Colombia and Universidad del Rosario and resulted in a two-day workshop with more than 26 participant organizations. Please refer to the event's policy brief and summary for more information regarding the event's contents and results.

Key Takeaways of the Event Planning and Execution

- Even having solid connections within public institutions, it proved very difficult to have them actively participate in the workshop. Having a social event (the coffee tasting) made a big difference in terms of participation and ease of networking.
- In our experience, the willingness of government agencies to participate comes from being able to provide a specific solution to an existing government initiative. Specifically within the Colombian context, there is willingness to agree to "get involved" however it has been difficult to translate this into actual collaboration. Based on some of the input of the workshop we think initiatives like co-funded grant schemes could be a good way to facilitate collaboration between stakeholders by structuring collaboration around a specific project.
- The possibility of collaboration between different actors is variable depending on the person in charge of different government sectors, meaning that continuity of projects is vulnerable to changes in power. This was a concern that came up often in during the event.

- The event would have benefited greatly from having a pre-workshop survey of actors. This would have allowed us to better understand the landscape of participants to better design an event to encourage collaboration.
- Although there was rigorous notetaking during the event, documenting the workshop proved difficult. It would have been a good resource to have the participants document the workshop themselves.

Analytic Pipeline

Background and Necessity

It is not uncommon for governments to have redundant data repositories, where each branch has its own database, often functionally isolated from each other. This hinders the development of technological solutions that require data from different institutions and interoperability of the public sector, and Colombia is no exception[1].

On the other side, when data is available it is often, if not always, sourced or agreggated at an administrative-unit level. Although this makes sense from a logistics point of view, it discourages the use of more organic analysis-units which can be better suited for understanding and addressing emergencies. Restricting analytic inference to administrative units, can artificially exclude relevant dynamics that are not determined by administrative boundaries (migration, climate, informal settlements etc...).

Furthermore, impact evaluation as a tool to inform crisis response, benefits from having flexibility when choosing the temporal and spatial resolution of the data. This is true in other contexts of public policy where the questions of *where* and *when* directly determine the budget, timeline and expected outcomes of a given policy.

Thus, it would be useful to be able to study emergencies using case-specific geographical units. This would entail a collection of algorithms and datasets that, given a particular mechanism, would divide the affected space and cluster it into a set of homogeneous geographic areas, each with distinct *behavior*. These geographic areas would then be treated as analysis units that can be further evaluated to design area-specific interventions, giving this approach an advantage over the use of predetermined (administrative) units.

---

[1] https://www.bbva.com/es/que-le-falta-a-colombia-para-ser-lider-en-transformacion-digital/

Developed Tool

The proposed analytic pipeline is a collection of scripts and datasets designed to generate geographical units for the study of Malaria in Colombia. This tool is intended as a first approach to the problem, designed with enough flexibility to make it adaptable to other emergencies given enough data.

Included Databases

- **Coca:** The annual coca field seizure at a national scale as reported by the army. The data consists of annual 1000mx1000m pixel map with the amount of area dedicated to coca plantations. Dataset courtesy of the Universiadad del Rosario.
- **Weather:** Daily aggregates from the ERA5 datasets of the ECMWF[2]. This dataset includes precipitation and temperature at a 1000mx1000m resolution. This dataset is publicly available and is acceded through Google Earth Engine (GEE).
- **Deforestation:** Yearly deforestation information from Hansen et al. Global Forest Change Dataset.[3] This dataset is publicly available and is acceded through Google Earth Engine (GEE).
- **Human Mobility:** Daily human movement aggregated between tiles from the Data for Good Dataset for the COVID-19 Emergency from Meta. The dataset consists of the number of devices traveled between predetermined 10kmx10km tiles covering the entire Colombian map. This dataset was made available to the group as part of an academic data partnership with Meta after the start of the pandemic. It will soon become a private Legacy dataset.
  This dataset will soon become legacy
- **Gold Mining:** Gold mine production by municipality as reported by the Observatorio Colombiano de Minería[4]. It will be replaced by a dataset of gold mining area at a 1000mx1000m resolution.
- **Malaria:** Daily malaria cases as reported by the *Instituto Nacional de Salud.* The data is at a municipality level and is part of their public datasets. The database also includes metadata from the reported case, including parasite, gender, age and ethnicity.

---

[2] https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5

[3] https://www.science.org/doi/10.1126/science.1244693

[4] https://pure.urosario.edu.co/es/projects/colombian-mining-observatory
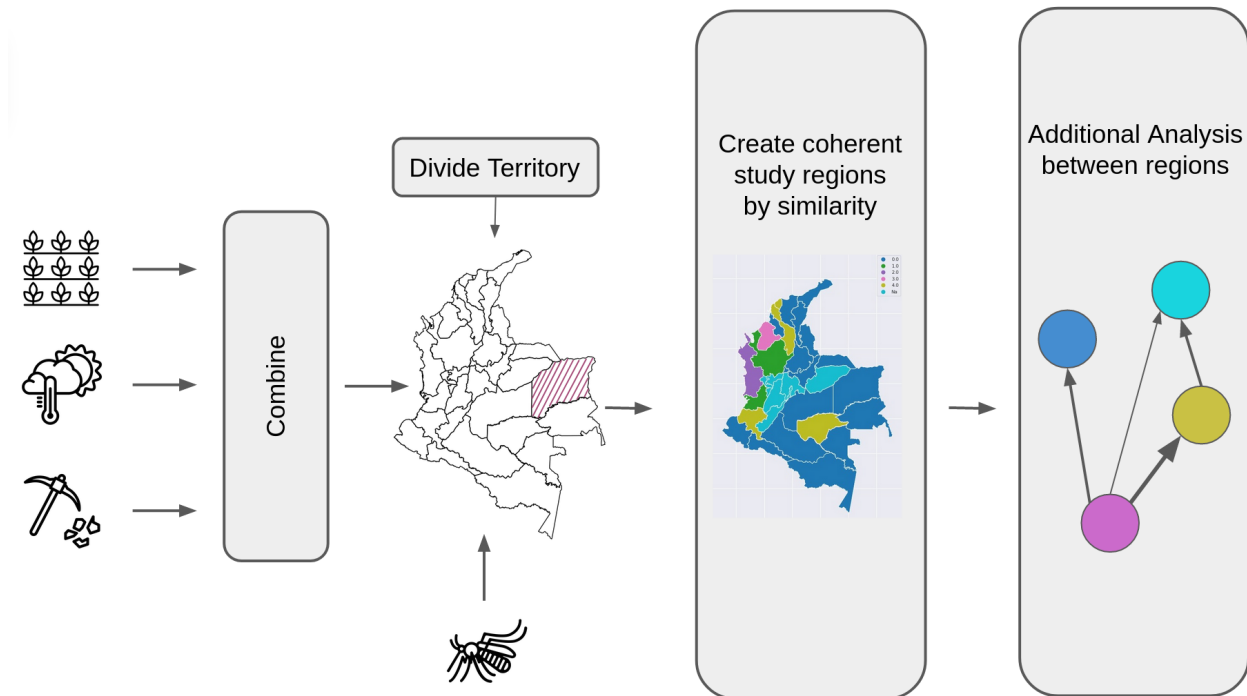
General Use Case Scenario

Figure 1. General use scenario.

The general use scenario, as showed in Figure 1, follows the scheme:

1. Combines the different available datasets, so that they have the same geographical and temporal resolution. The geographic division is set as a parameter and can be, in principle, any division of Colombian territory (states, municipalities, a square grid, rivers etc…)
2. Runs the model that relates the source variables with the target variable (this model can be as simple as a linear regression).
3. Performs a clustering algorithm on the different models, where similarity is determined from the interaction between the explanatory variables and the target variable (malaria cases) over each geographical unit.
4. The resulting clustered regions already represent a coherent division of the territory in the context of malaria, given a particular understanding of the model relating source and target variables, but can be further analyzed to understand its behavior.
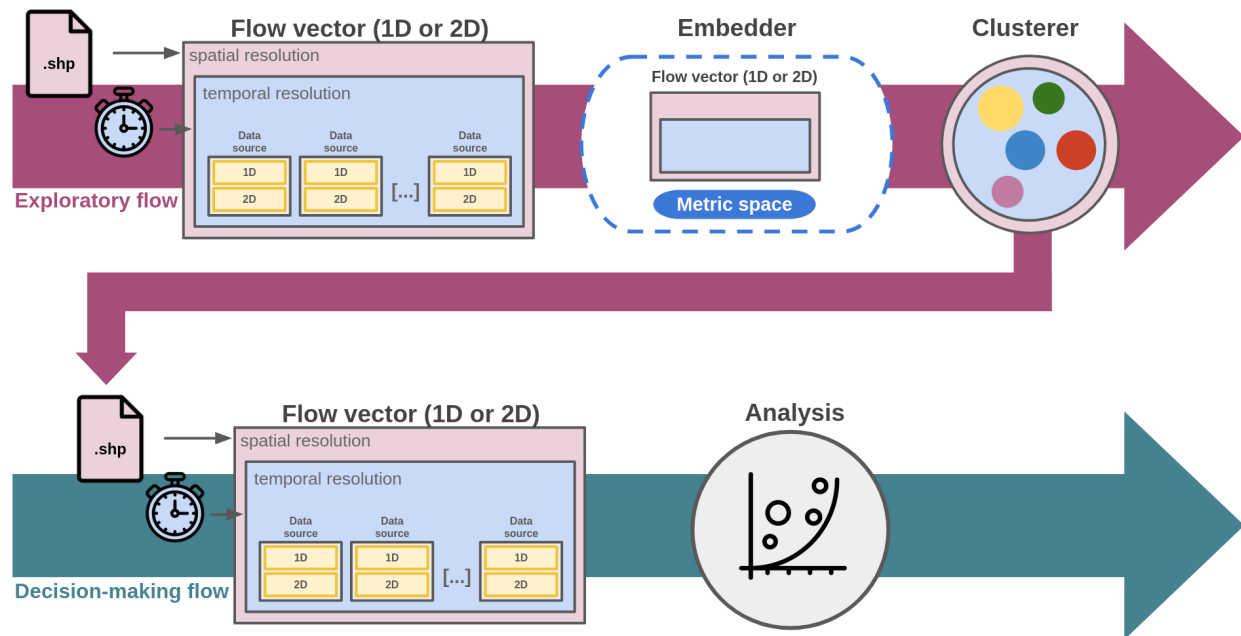
Figure 2. Pipeline's input/output scheme

It is important to note that the pipeline's output can serve as an input in a following iteration, obtaining not only new "meta" geographical units of study from the previous ones, but clean tabular data associated to them. Thus, the pipeline can also be used as a simple tool to unify and construct datasets at a desired geographical unit.

Technical Specifications

The totality of the developed code is in Python and made publicly available at: https://github.com/Data-Lama/pathogen_study_regions_generator. The code is under MIT license for free distribution.