

International Syndromic Surveillance Project

Eduardo Rafael Cerqueira

Spring 2023

DSA 5900

4 credit hours

Dr. David Ebert, Dr. Gopichandh Danala

Data Institute for Societal Challenges (DISC)

**Introduction**

The project describes the International Syndromic Surveillance monitoring COVID-19 virus. Indeed, there is a need to identify the number of COVID cases, the number of COVID-related deaths, and the hospitalization rates. Access to this upcoming projection rates earlier would be of immense value to public health agencies, epidemiologists, decision-makers, and governmental organizations. The decision-makers or people responsible for health-related issues can quickly look at these surveillance systems and be more prepared to curb the spread of viruses like Covid 19.

**Objectives**

PanViz 2.0 is an interactive web-based visual tool incorporating various data streams across the US to simulate and model important disease spread metrics. These metrics include positive cases, deaths, hospitalization, and infection rate at different spatial hierarchical regions (country, state, county etc.) within the United States. Researchers at OU DISC, in collaboration with other departments across campuses of OU and OUHSC, developed the Panviz system in its current form. This interface could be used to project the cases, deaths, and hospitalization rates caused by COVID-19. In this project, the regional capacity of the current system will be expanded for countries like Peru, which have a different geographical hierarchy than the US. The hierarchy of Peru has four levels - (Country-level, Department-level, Province-level, and District-level). We have to explore the current data streams of the Peru Covid cases and find optimal ways to process the information into a meaningful aggregated form similar to the covid data provided by the New York Times.

Then we will evaluate the COVID-19 data from Peru with an understanding of its geographical hierarchy, aggregating the data at different temporal levels too, including daily, weekly, biweekly, monthly, and annual basis. In addition, we will run several machine learning algorithms with the context of Peruvian data for the projections of cases, deaths, and hospitalization rates and integrate the results into the PanViz system. Furthermore, this project also aims to give the knowledge of the Peru COVID data, the current PanViz system for the US, creating automated scripts to do data aggregation for Peru and run machine learning models for the integration.

**Plan**

The data for this project is given by the Peru government website (<https://covid19.minsa.gob.pe/sala_situacional.asp>). It contains the number of positive cases, deaths, and hospitalizations for COVID-19. The format is primarily numeric, yet strings are present for the locations, yes/no questions, and type of vaccines. These qualitative data will be converted to quantity data.

Data aggregation and processing will focus on understanding Peru's database and geographical hierarchy. My responsibilities in this project will be mainly data science-related duties that will aid the existing research team and will be included in the main part of the analysis.

The final deliverable is integrating the Peru data into the PanViz system. The functionalities of the existing PanViz system will be expanded to communicate the findings using standard plotting (line and bar charts) to predict the COVID cases for future dates. The project completion date will be April 28th, 2023. The following three months for this practicum study will be partitioned by understanding the codebase, converting the Peru data into more structured data, running the existing machine learning models on the Peru data and storing the results, and integrating all features into the PanViz system for final analysis.

**Deliverables:**

* Using the current codebase to finetune and scrape the raw data of Peru country.
* Data aggregation at different spatial and temporal scales
* Evaluating the performance of several machine learning models in doing predictive analysis.
* Integrating the final results in to the Peru version of the Panviz.

**Schedule:**

|  |  |
| --- | --- |
| **Week** | **Tasks** |
| 01/27/2023 | Project Proposal, Establishing GitHub repository |
| 02/03/2023 | Submitting the proposal draft, scheduling a weekly meeting |
| 02/08/2023 | Getting access to the codebase, and understanding data, Panviz system |
| 02/15/2023 | Web scraping to do data collection and script automation |
| 02/22/2023 | Data aggregation and solving any issues |
| 03/01/2023 | Continue to work on data collection, aggregation and different spatiotemporal scales |
| 03/08/2023 | Explore ML algorithms for predictive analytics |
| 03/05/2023 | Continue ML-related coding work |
| 03/20/2023 | Mid-Semester progress report; Asses the progress and adjust milestones if needed |
| 03/29/2023 | Continue ML-related coding work |
| 04/03/2023 | Access and finetune modeling results |
| 04/10/2023 | Continue finetuning, and integrate results into Panviz |
| 04/18/2023 | Integrate results into Panviz |
| 04/28/2023 | Final report |