Predicting the impact of earthquakes on the population and the economy

Team Members:

- Aaron Nova
- Till Nils Böhringer
- Moises De Jesus Maldonado Alonso
- Florian Hellwig

Proposal

1. General objective

Our study aims to predict the potential impact of earthquakes on the population and the economy of a given geographic region by calculating a risk factor score. This risk score is based on the damages caused by previous earthquakes, deaths, missing persons, injuries, houses destroyed and damaged. To estimate the risk, we use population density maps and earthquakes databases, as well as average regional income. A major part of the project will be to combine multiple data sources into one data base that can be used to train classic machine learning models.

2. Motivation

Rooted in Nassim Nicholas Taleb's "Black Swan" theory, earthquakes represent those highly improbable events that have a major impact because they are hard-to-predict, rare events that come as a surprise. Our motivation for this project is twofold: first, we intended to estimate the effects of a statistically unexpected event of great magnitude and consequence, and secondly, we aim to provide a practical tool to response to these events in a more effective manner.

3. What questions will you try to answer with the project?

3.1 General research question:

Is it possible to predict the impact of earthquakes on the population and economy of a given geographic region based on high population density maps and the record of earthquake damage in similar topographic regions around the world?

3.2 Other research questions to be consider:

Risk Factor

What are the Areas in which there is the highest earthquake Risk Factors for Humans based on our predictor variables?

Spatial

How do the demographics of the population correlate with the size of aftermath of the event?

Can we predict which areas are most likely to have these events?

Financial

Do the financial costs of an earthquake correlate to the population density and what about the magnitude of the earthquake?

Are the financial costs of an earthquake higher/lower in more wealthy countries? (E.g. Do more expensive buildings cause lower financial costs? How do different socioeconomic factors influence the vulnerability of populations to the impacts of earthquakes and volcanic eruptions? Are areas where there was a sizeable increase in population of the last years more or less prone to big destruction?)

Impact

How is the population impacted by an earthquake? (E.g. do the people leave after a big catastrophic event?)

Has the impact of similar earthquakes increased or decreased over the times?

4. What data will you be using (and what is the source of that data)?

Significant earthquake database (CSV): 1193 entries from 2000-2020

https://public.opendatasoft.com/explore/dataset/significant-earthquake-database/table/?sort=year

- Date
- Focal Depth
- Magnitude
- Coordinate
- Deaths and Injuries in absolute numbers
- Damages (categorical data, often no absolute numbers but a range e.g: 5-24 M$)
- Sometimes information about nr. of destroyed and damaged houses

Population density of the world, years 2000, 2005, 2010, 2015, 2020 (GeoTIFF (.tif) and ASCII (.asc))

https://sedac.ciesin.columbia.edu/data/set/gpw-v4-population-density-rev11/data-download

- Population density stored in WGS84, geographic coordinate system (latitude/longitude)

High Resolution Population Density Maps around the year 2020 (GeoTIFF sometimes CSV)

https://data.humdata.org/organization/meta?q=population%20density&sort=score%20desc%2C%20last_modified%20desc&ext_page_size=25

- Data for each country is in a new file.

Average national income from 2000-2020 (CSV)

https://wid.world/data/

- Average national income per country, normalized to euros.

5. What techniques will you use for the analyses?

5.1 Data Modelling process

We will use the standard SEMMA (Sample, Explore, Modify, Model, Assess) methodology mixed with the CRISP-DM methodology, encompassing similar steps for the full data science pipeline. Since the data of interest comes from different sources, compiling it into a useful master data set will be a focus of this project. Furthermore, we are aspiring to learn how to handle lesser-known data types, such as geoTIFF for spatial data.

5.2 Methods

We will use supervised machine learning methods for regression, and we will apply PCA to reduce the dimensionality of the features of the original datasets, because the intended focus is on prediction accuracy rather than on statistical significance.