Written Assessment: Data Scientist

Alexandra Ruth

Earthquake Hazard Assessment at U.S. Embassy and Consulate Locations

Executive summary

Significant earthquake events pose both safety and security risks to U.S. embassy personnel overseas. U.S. embassies and consulates are distributed across a wide range of geographic locations, some of which are closer to areas of high earthquake activity than others depending on their location relative to the earth's tectonic plates. For this analysis, the **estimated percent probability of one or more significant earthquakes occurring within the next five years within 20- or 100-miles of an embassy/consulate location** was calculated to determine the highest-risk embassies and consulates that should be prioritized for planning purposes. These probabilities were calculated using the Poisson distribution, a statistical distribution that can be used to estimate the occurrence of relatively rare events, such as earthquakes, based on previous event counts over a given time period. Counts of previously-occurring earthquake events within 20- and 100-miles of embassy locations were derived from a dataset sourced by the National Oceanic and Atmospheric Adminstration (NOAA) significant earthquakes database and included earthquakes from the past ~50 years. U.S. Embassy and Consulate geolocations were obtained by scraping text addresses from the U.S. State Department website and geocoding these addresses in R to obtain latitudes and longitudes.

The top three locations with the highest estimated probability of a significant earthquake event occurring within a 20-mile radius in the next five years are the **U.S. Embassy in Albania** (31.93%); the **U.S. Embassy in Algeria** (25.06%); and the **U.S. Embassy in Quito, Ecuador** (25.06%); these embassies should therefore be highest-priority for earthquake contingency planning given the probabilities of a significant earthquake occurring in close proximity within the next five years. Further results for these calculations - including top ten locations at highest risk for each radius - are described in **Tables 1 & 2** of this report.

INTERACTIVE MAP TOOL

To further explore these data and calculations, an interactive mapping tool was created and can be accessed at the link below. Light blue stars indicate embassy and consulate locations, and yellow dots indicate past significant earthquake events. Clicking on an embassy or consulate location will reveal estimated risk probabilities and hovering over an earthquake location will reveal magnitudes:

https://rpubs.com/aruth3/921267

I. QUESTION 1: Study Design

i. Research question

Which U.S. embassies and consulates are at the highest risk for **at least one significant earthquake event occurring in the next five years** within a 20-mile radius of their location? Within a 100-mile radius of their location?

ii. Study design and approach

Overall approach

This study combined web-scraping and mapping techniques to calculate probabilities of at least one significant earthquake event occurring in the next 5 years within a geographic radius of a given embassy or consulate location. Physical addresses of embassies/consulates were scraped from U.S. State Department webpages and then geocoded to get latitude and longitude coordinates. Nearest-neighbor mapping techniques were used to obtain counts for previous significant earthquake events that had occurred within a radius of 20 miles and 100 miles of each embassy location in the time period from the NOAA dataset. These radii were selected because they can reasonably be expected to be meaningful if a significant earthquake event were to occur, particularly given that civil unrest following an earthquake event - even 100 miles away - might pose security risks.

Poisson probabilities for a significant earthquake event occurring within those radii in the next five years were then calculated for each embassy. The primary outcome(s) of interest were the probabilities of at least one significant earthquake event occurring in the next five years within 20mi and 100mi of an embassy/consulate.

Rationale

The Poisson probability distribution is used for calculating probabilities in cases where event-count data are the outcome of interest, where events can be assumed to be unrelated to each other, and where events being counted are relatively rare, making it a suitable distribution for earthquake count data. The Poisson distribution is represented by the following equation, where λ equals the mean occurrences of an event over a particular time interval: $P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$

The outcome of interest calculated for each embassy using the Poisson distribution was $P(X \geq x)$, where P represents the predicted probability of one or more significant earthquake events occurring within a specified radius of an embassy's location within the next 5 years. For example, the variable for the percent probability of one or more earthquake occurring within a 20mi radius of a given embassy location within five years was calculated using the formula `ppois(q=0, lambda = 5*(counts_20mi/52), lower.tail=F)`, where `q` represents the count threshold, `lambda` represents the rate calculated by the counts from the NOAA dataset that covers a 52-year timespan, and `lower.tail=F` ensure that the upper tail of the probability (i.e. *more* than zero earthquakes) is the output.

iii. Data sources

Earthquake data

Earthquake data were obtained from an Excel file derived from the NOAA significant earthquake database. This file contained earthquakes from a ~50 year period and included earthquake attributes such as geolocations, magnitude, casualties, and damages.

U.S. Embassy and Consulate location data

A vector of countries with U.S. embassy or consulate locations was created by scraping text from the U.S. Embassies webpage provided in the assignment (https://www.usembassy.gov). A for-loop was constructed from this list of countries to paste country names into urls, navigate to the associated country embassy

webpages, and scrape text of physical addresses from these pages according to the address locations in the html code using the `rvest` package. The text for embassies/consulates' physical addresses was then geocoded using the `tidygeocoder` package in R, which calls the Google Maps API with an associated API key to get latitudes and longitudes for each of these physical addresses. The final number of geocoded embassy and consulate locations using this method was **N=209**.

II. QUESTION 2: Data Analysis

i. Key findings

Estimated probability of a significant earthquake within a 20-mile radius of an embassy or consulate within the next five years

The ten embassies/consulates at highest risk of a significant earthquake occurring within a 20-mile radius in the next five years are outlined in **Table 1**. The **counts** variable represents the number of significant earthquake events that have occurred within a 20mi radius of each location within the time period included in the NOAA dataset, and the **prob** variable represents the Poisson probability of a significant earthquake occurring within a 20mi radius of the embassy within the next 5 years based on earthquake frequencies given their locations in the NOAA dataset.

A vast majority of the embassies/consulates in this sample (N = 180; 86.1%%) had no estimated risk ($P = 0$) of one or more significant earthquakes occurring within a 20-mile radius in the next five years; this indicates that resources should be focused on the relatively small fraction of embassies/consulates at highest risk of a significant earthquake occurring in close range.

Table 1: Probability of a significant earthquake in the next 5 years within 20 miles

address	counts	prob	lonlat
U.S. Embassy Tirana, Rruga e Elbasanit, No. 103, Tirana, ALBANIA	4	31.93	19.83245, 41.31776
U.S. Embassy in Algiers, 05 Chemin Cheikh Bachir Ibrahimi, El-Biar 16030, Alger Algerie	3	25.06	3.041788, 36.754758
U.S. Embassy Quito, E12-170 Avigiras Ave. and Eloy Alfaro Ave., Quito, Ecuador	3	25.06	-78.4678238, -0.1383838
U.S. Embassy in San Jose, Calle 98 Vía 104, Pavas, San José, Costa Rica	2	17.49	-121.88525, 37.33874
U.S. Embassy in Cairo, 5 Tawfik Diab Street, Garden City, Cairo, Egypt	2	17.49	31.23329, 30.04125
U.S. Embassy in San Salvador, Final Boulevard Santa Elena, Antiguo Cuscatlán, La Libertad	2	17.49	-89.25803, 13.66437
U.S. Embassy in Athens, 91 Vasilisis Sophias Avenue, 10160 Athens, Greece	2	17.49	23.74824, 37.97657
U.S. Consulate in Thessaloniki, 43 Tsimiski, 7th Floor, 546 23 Thessaloniki GREECE	2	17.49	22.94245, 40.63330
U.S. Embassy in New Delhi, Shantipath, Chanakyapuri, New Delhi – 110021	2	17.49	77.18809, 28.59751
U.S. Embassy in Skopje, Str. “Samoilova” Nr.21, 1000 Skopje, Republic of North Macedonia	2	17.49	21.43346, 42.00532

Estimated probability of a significant earthquake within a 100-mile radius of an embassy/consulate within the next five years

The ten highest-risk embassies/consulates for a significant earthquake event in the next 5 years within a 100mi radius of their locations are listed in **Table 2**. There is significant overlap between these highest-risk locations and the locations described in **Table 1**. Roughly half of embassies/consulates also had no estimated risk ($P = 0$) of a significant earthquake event in the next five years within a 100mi radius (N = 108; 51.7%).

Table 2: Probability of a significant earthquake in the next 5 years within 100 miles

address	counts	prob	lonlat
U.S. Consular Agency Bali, Jl. Hayam Wuruk 310, Denpasar, Bali, Indonesia 80235, Phone: (62) (361) 233-605	22	87.94	115.244452, -8.672247
U.S. Embassy Tirana, Rruga e Elbasanit, No. 103, Tirana, ALBANIA	17	80.50	19.83245, 41.31776
U.S. Embassy in San Salvador, Final Boulevard Santa Elena, Antiguo Cuscatlán, La Libertad	16	78.53	-89.25803, 13.66437
U.S. Embassy Rome, via Vittorio Veneto 121, 00187 Roma	16	78.53	12.49078, 41.90659
U.S. Embassy in Podgorica, Dzona Dzeksona 2, 81000 Podgorica, Montenegro, Embassy Switchboard: +382 (0)20 410 500	16	78.53	19.25119, 42.43692
U.S. Mission Rome, via Boncompagni 2, Rome, Italy	16	78.53	12.49206, 41.90785
U.S. Embassy in Skopje, Str. "Samoilova" Nr.21, 1000 Skopje, Republic of North Macedonia	15	76.36	21.43346, 42.00532
U.S. Consulate General Florence, Lungarno Vespucci, 38, 50123 Firenze	13	71.35	11.24052, 43.77380
U.S. Embassy in Pristina, Arberia/Dragodan, Nazim Hikmet 30, Pristina, Kosovo	13	71.35	21.13761, 42.66248
U.S. Embassy in San Marino, Lungarno Amerigo Vespucci, 38, 50123 Florence, Italy	13	71.35	11.24056, 43.77386

Embassies/consulates located along the borders of tectonic plates where there is significant earthquake activity are at elevated risk - this is visually apparent in the Leaflet map that was constructed as part of this assessment, which is linked on **page 1** of this report. Earthquake points are clustered along the outlines of high tectonic plate activity - e.g. at the western border of South America. The findings in this study concur with this general pattern, with many of the highest-risk embassies/consulates in these probability calculations being located along areas of high tectonic plate activity (e.g. Bali, Greece).

ii. Data quality

Embassy/consulate locations

One limitation of the webscraping approach is that it may have yielded an incomplete list of embassy/consulate locations and physical addresses - it would be better to use an internal list of embassy/consulate addresses that has been verified and is complete. Additionally, it appears that some of the webpages for these embassies may have been out-of-date - the html code in the Afghanistan page, for instance, indicated that the page was last updated in June 2021. Some of the physical addresses provided were PO boxes, which may not reflect the actual physical location of the embassy/consulate but is reasonably expected to be close by - it may be for security reasons that the physical addresses were not published publicly online. A secure, internal list of embassy/consulate addresses would better guarantee accuracy of results.

Earthquakes

The NOAA dataset contained very complete information for earthquake latitudes, longitudes, and magnitudes - this is why these variables were most heavily used in these analyses. While it would have been interesting to construct models including outcomes such as injuries and fatalities from earthquakes events and models that included the Mercalli index, these data were less complete - for example, more than half of the Mercalli index fields were missing (n=1380; 62.6%).

iii. Data pre-processing

Earthquake data

Data from the excel file derived from the NOAA dataset were imported as a csv and inspected for completeness, ranges, and missing values (see data quality notes above). Earthquake data were handled using the `tidyverse` package in R - only variables of interest were selected for each earthquake event (latitude, longitude, and magnitude) and included in the analyses. Coordinates were converted into a shapefile in R using the `sf` package.

Embassy/consulate location data

Embassy/consulate location data were obtained as described in **Section I.iii** above. Embassy/consulate coordinates were converted into a shapefile in R using the `sf` package, and counts for earthquakes occurring within both 20mi and 100mi of each embassy were calculated using the `nngeo` package for nearest-neighbor counts of points located within a given distance - these counts were then used to calculate Poisson probabilities.

iv. Analytic tools used

All data were imported, processed, cleaned, and analyzed in RStudio version 4.1.2. The R packages used for data processing and cleaning were `tidyverse` and `janitor`. For scraping text addresses of U.S. embassy/consulate locations, the `rvest` package was used and the Selector Gadget browser plug-in was used to identify places in embassy/consulate webpages' html code where address text was located. Embassy/consulate addresses were geocoded using the `ggmap` and `tidygeocoder` packages that call the Google Maps API for longitude and latitude information. Nearest-neighbor calculations were completed using the spatial packages `sf` and `nngeo`. The final map product assembled for dissemination purposes was created with the `leaflet` package. This report was created using R Markdown.

v. Future directions

With more time and resources, it would be valuable to add a data field for each embassy/consulate location that specifies the number of U.S. embassy/consulate staff working there - this would be helpful for informing evacuation planning. Other data fields that would be good to add would be ones that reflect other underlying risk characteristics in countries that might exacerbate the severity of an earthquake event - e.g. hospital capacity or history of civil unrest. It would also be interesting to find ways to have the map tool make calls to an API where earthquake events are constantly being reported and data are regularly being updated in real time, instead of using a static dataset to calculate probabilities.

Another exciting future direction for this work would be developing automated reports in R Markdown where a country name could be inputted by a user and a 1-2 page pdf report summarizing earthquake hazards and information for that country's embassies would be automatically generated, including the embassy/consulate location in that country, estimated probability/risk, and other country-specific info.

III. QUESTION 3: Communication

i. Communication strategy

My strategy for communicating this information to internal stakeholders would include:

1) a brief **executive summary** (<500 words) of key findings (please refer to the summary at the beginning of this report) and
2) a **mapping tool** where stakeholders can interactively click through risk percentages described in this paper using a Leaflet map: https://rpubs.com/aruth3/921267