Quantification of Undermapped Regions in OpenStreetMap and their Associated Risks of Natural Disaster

Doran, Im, O'Donnell, Ofoma
DS6015

Introduction

OpenStreetMap (OSM) has a long history of supporting disaster relief by providing basic geographical data at no cost that can be used by first responders and updated dynamically by volunteers on the ground. OSM's crowd-sourced, collaborative model enables rapid responses to a range of disasters and provides crucial information about roads and structures. Numerous apps have been built with OSM infrastructure and data, and new uses continue to be developed. OSM US is a nonprofit organization helping to support and grow the project in the US. Historically, OSM US's disaster-response mapping efforts have been reactive, rather than proactive, due to the unpredictable nature of natural disasters. This paper aims to detail an illustrative, usable framework and tool that identifies areas in OSM that are undermapped — missing structures, roads, descriptions, etc. — and at risk for natural disasters.

In 2018, the Microsoft Maps team created a deep learning model that uses computer vision to identify all structures in the US from overhead satellite imagery. We propose leveraging this model (MSAI) along with existing OSM data and the Federal Emergency Management Agency (FEMA) National Risk Index (NRI) to generate a metric that identifies areas across the US where preemptive mapping would be most impactful. Successfully predicting undermapped areas in the OSM database will allow volunteers to improve the data prior to a natural disaster, thus facilitating a more effective response by emergency crews and potentially saving lives.

The goal for the project is the efficient and directed identification of at-risk areas for preemptive mapping by OSM US volunteers.

Dataset Backgrounds

OSM Database

The OSM database includes a wide range of geographic features, including infrastructure elements (e.g., roads, bridges, and sidewalks), structural elements (e.g., buildings, sheds, carports), and other elements (e.g., lakes, parks, and neighborhood boundaries). These elements often have semi-structured, non-standard metadata (e.g., hours of operation, existence of traffic lights, and surface type). The database is crowdsourced, and contributors collect data from surveys, trace from aerial imagery and also import from other freely licensed geodata sources. OSM uses its own topology to store geographical features which can then be exported into other GIS file formats. OSM's website also provides an online map, geodata

search engine, and editor. OSM's own API, Overpass, is a comprehensive, read-only API that allows for custom queries of the OSM database. It supports multiple custom query languages and returns data in a format that can be easily used for analysis in python.

MSAI Building Footprints

Microsoft provides a US-wide dataset with nearly 130 million computer-generated building footprints that is available freely for download and use. These footprints were generated from Bing satellite imagery of varying age, with some areas generated from imagery between 2019 and 2020 and the remaining areas generated with imagery from an average source-year of 2012. The impact of this varying data vintage is discussed in both the Methods and Challenges sections below.

To generate the footprints in this dataset, Microsoft created an image segmentation algorithm to distinguish building and non-building pixels using EfficientNet architecture, achieving precision and recall metrics of 94.0% and 95.5% respectively. The algorithm then polygonizes those shapes to create an outline of the structure in geographic space. This polygonization takes into account the angle of the satellite imagery as well as the quality of the image data. To evaluate the polygonization performance, Microsoft utilized intersection over union, shape distance, and dominant angle rotation error with results of 0.86, 0.4, and 2.5 respectively. They also verified the false positive rate by sampling 1,000 buildings polygonized by the algorithm and determined that <1% of identified buildings resulted in a false positive.

Of all the data sources that we reviewed for determining the state of mapping in the US the Microsoft Maps data was by far the most complete and rigorously evaluated, which is why we chose it as our baseline data set for the evaluation of OSM data.

FEMA NRI Dataset

FEMA provides the NRI dataset and tool, which include baseline risk management data at every Census tract level — including the county level — for 18 natural hazards, such as wildfires, hurricanes, and heatwaves (https://hazards.fema.gov/nri/). The dataset is available as a public download and through an interactive dashboard. Risk, as defined by FEMA, is the potential for negative impact as a result of natural disaster. Risk values in the NRI are carefully calculated by FEMA using three variables: expected annual loss, social vulnerability, and community resilience.

Expected annual loss represents the average economic loss in dollars resulting from natural hazards each year. It is calculated for each hazard type and quantifies loss for relevant consequence types: buildings, people, and agriculture. Social vulnerability is the susceptibility of social groups to the adverse impacts of natural hazards, including disproportionate death, injury, loss, or disruption of livelihood. Community resilience is the ability of a community to prepare for anticipated natural hazards, adapt to changing conditions, and withstand and recover rapidly from disruptions. Both social vulnerability and community resilience include socioeconomic factors such as income, age, number of hospitals, road systems, etc. as well as environmental

factors among other contributing factors. Using these three overarching metrics, FEMA implemented k-means clustering to calculate the expected annual loss and risk values.

US Census Boundaries

The US Census Bureau provides cartographic boundary files for varying levels of administrative divisions at several resolutions. For this project, we used data at the most detailed resolution (1:500,000) for the 3143 counties as of 2020. This year was selected because the Microsoft data, as mentioned above, includes imagery only as new as 2020; in addition, Connecticut's counties and county boundaries have shifted since 2020 and will continue to shift due to a proposal approved by the Census Bureau in 2019. The dataset used is a lower-resolution representation of the Bureau's Master Address File/Topologically Integrated Geographic Encoding and Referencing (MAF/TIGER) System. The higher-resolution dataset was not chosen because the additional fidelity would require additional compute and was not needed for this project.

Methods

To develop a statistic for identifying priority areas for mapping in OSM, we take two intermediate steps. First, we define the concept of an "undermapped" area, and then we establish and calculative the quantitative metric of a region's level of mapping status as compared to the baseline dataset from Microsoft. From this definition, we then define a combined statistic that integrates this quantitative undermapped metric with the disaster risk metric from the FEMA NRI. Throughout this project, we use counties as the smallest geographic region using official shapefile boundaries from the US Census Bureau. We choose counties as the distinct geographic unit for two reasons. First, counties are a typical geographic delineator that are used fairly widely in the analysis of statistics like this. Second, counties are a visually intuitive and analytically logical size — they are not so small that the map is cluttered, and not so large that we mask smaller regions within the state.

Defining an Undermapped Metric

To define the concept of an undermapped area, we focus on buildings, which are somewhat standardized within the database. Further supporting this choice, buildings are one of the primary points of interest for first responders regardless of the natural disaster type, and buildings tend to be undermapped at a much higher rate than roads. Buildings in OSM's database have an associated polygon as visible from overhead

Figure 1: Visualization of OSM Polygons

satellite imagery (Figure 1), so we define our quantitative mapping metric to be the area of the polygons for a given geographic region. With this metric in mind, we define the state of mapping completion as the area covered by structures in OSM vs the actual area covered by structures in reality. This area metric provides a quantifiable approach to the definition of undermapped. Given the area metric defined above, we then leverage the MSAI dataset as our ground-truth building data.

Generating the Undermapped Statistic

To generate the undermapped statistic for each county, we calculate the total building footprint area in the county for both the OSM data and the MSAI data. To retrieve OSM data, we use the Overpass API through OSMnx, a python library built specifically to work with OSM data. To read in MSAI data, we download each state's GeoJSON file and use GeoPandas to read in the files. For each dataset, we calculate the area of each building polygon (or multipolygon, which is a set of polygons). Then, to calculate the total building footprint area per polygon, we use the Census shapefiles as polygon masks to only select buildings within a given county. The areas of those buildings are then summed, giving a final area per county per dataset. The undermapped statistic itself is given by the following formula, which we run for each county:

$$Undermapped\ Statistic\ =\ (1\ -\ \frac{Area\ of\ OSM\ Building\ Footprints}{Area\ of\ MSAI\ Building\ Footprints})\ \times\ 100$$

The formula assumes that the MSAI dataset produces an area value that is always greater than the area value from the OSM data. If this assumption holds true, then the statistic should be between 0 (for a fully mapped county) and 100 (for a fully unmapped county).

This assumption does not hold true in all counties, and some county statistics are negative as a result of the OSM data being more up to date than the MSAI data. As briefly discussed in the Dataset Background section, the MSAI data for some areas was derived from satellite imagery collected before 2012. Negative values typically occur in outdated imagery areas that are either heavily populated (where there are a significant number of local contributors) or have had high growth (where new construction is prevalent and only visible in newer imagery). Despite the challenges associated with the MSAI dataset and the accuracy to the current built environment, the dataset still provides an excellent "ground truth" for the state of OSM mapping for a vast amount of the country, as shown in Figure 2, which is drawn using the statistic calculated according to the formula above.

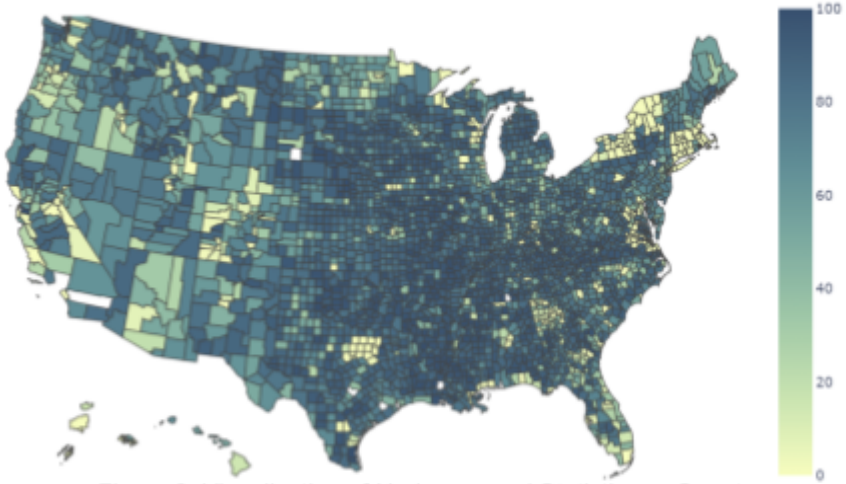


Figure 2: Visualization of Undermapped Statistic per County

Incorporating Risk from Natural Disasters

To incorporate the risk statistics associated with various natural disasters, we leverage FEMA's National Risk Index (NRI). Using tabular data downloaded from the complete 2022 dataset, we augment the dataframe from above, which includes county identifiers (such as name and Federal Information Processing Standards (FIPS) codes), OSM areas, MSAI areas, and the undermapped statistics. We merge the area data with the NRI data by FIPS code to create a full dataset with the necessary components to generate a final priority statistic.

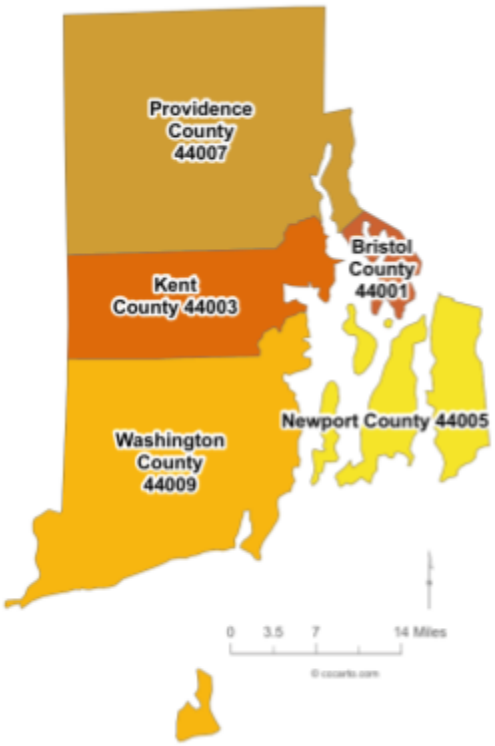


Figure 3: Example of National FIPS Codes

Generating the Priority Statistic

To achieve the project's ultimate goal of a single statistic to determine the priority of an area's need for mapping, we combine the undermapped statistic with the risk statistic with a weighted ratio formula. The single priority statistic serves as a visual aid and representation. The priority statistic is variable based on the disaster of interest, as well as the influence of that disaster's risk statistic. We run the following formula for each county:

$$Priority\ Statistic\ =\ (Risk\ Statistic\ \times\ Ratio)\ +\ (Undermapped\ Statistic\ \times\ (1\ -\ Ratio))\ \times\ 100$$

In this formula, the ratio, which is a value between 0 and 1, determines how much of the priority statistic is attributed to the risk statistic vs. the undermapped statistic. If the ratio is 0, the priority metric is entirely calculated with the undermapped statistic. If the ratio is 1, the priority metric is entirely calculated with the risk statistic. The priority statistic is ideally between 0 and 100, but due to the negative undermapped statistic values of certain counties, the priority statistic can also be negative.

Visualizing the Priority Statistic

Changing the ratio shows the impact of different disasters on different regions of the US. Given that an unexpected number of counties in the US are highly undermapped, a ratio that too heavily favors the undermapped statistic generates a visually uninteresting map with large swaths of the country consistently undermapped, which does not help prioritize counties for mapping. Figure 3 below illustrates an example of the priority statistic generated with five different ratios ranging from 0% risk to 100% risk weighting.

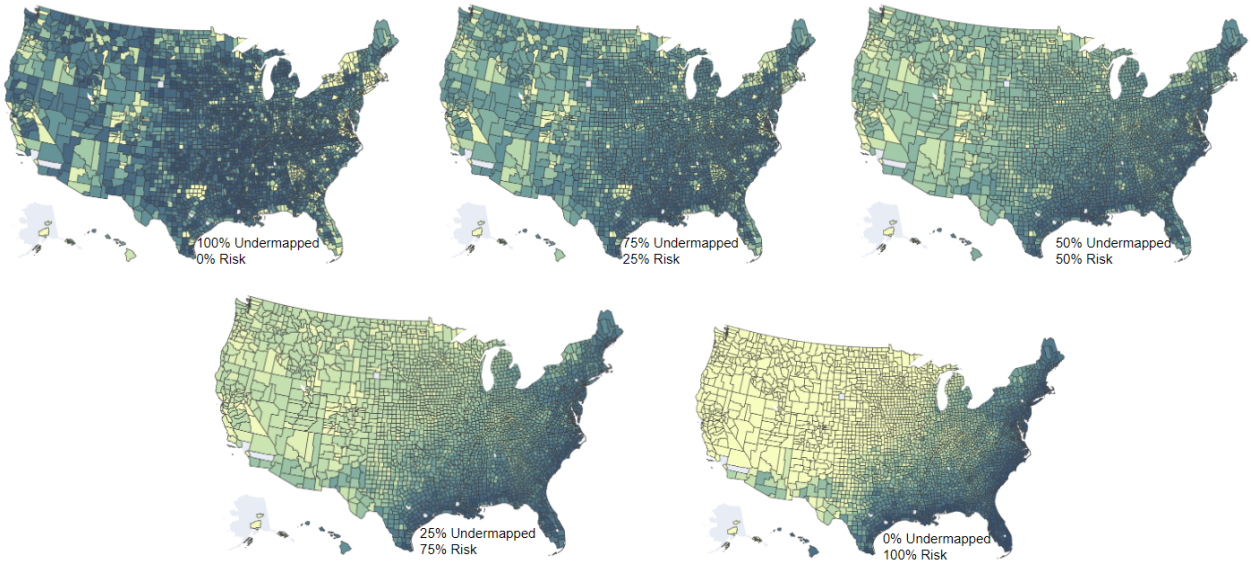


Figure 4: Maps of the Combined Risk Statistic with Ratios from 0% to 100%

Challenges

GIS Data Challenges

This project represented the most in-depth work that any project member had done on GIS data. Despite using many Python libraries that work with GIS data, such as GeoPandas, OSMnx, Shapely, and Fiona, we ran into several hurdles. The most impactful challenge within this topic was learning about and deconflicting the various projections of each database. Any data that uses a certain coordinate reference system (CRS) may represent the absolute locations of objects differently from data that uses a different CRS. In addition, a geographic CRS locates objects in a 3D space, such as the Earth's sphere, whereas a projected CRS locates objects on a 2D plane, such as a map of the Earth's surface. This distinction is critical when calculating areas of polygons or creating masks of areas to filter a dataframe. As part of the initial data wrangling, our code standardizes the CRS of the data, translates the geographic CRS to the appropriate projected CRS for area calculation, then translates CRS back once more. Understanding the correct choices represented a significant time investment in the beginning of the project.

Baseline Metric Challenges

During this project, the most difficult challenge was finding the best ground-truth dataset — or even any ground-truth dataset — to define areas as fully mapped or undermapped. We considered several options that leveraged different aspects of OSM's database, which contains a wide variety of features. After many failed options, we finally considered the option of sourcing satellite imagery and creating a neural network-based computer vision algorithm to identify building footprints. While researching this option, we discovered the MSAI Building Footprints

Program and the associated dataset. The various alternatives we initially considered fall into three broad categories: inferring the level of mapping based on indirect data, utilizing mapping APIs to identify features and structures in a geographic area, or utilizing image data to identify structures and roads.

The Indirect Data Approach

Several papers have explored indirect ways of determining the level of mapping in an area using proxy data, element density, etc. The paper "The world's user-generated road map is more than 80% complete" from Barrington-Leigh (2017) develops a methodology for determining completeness of OSM worldwide by two approaches: first, a sampling of visual data comparing OSM to Google Maps at random locations, and second, parametric time-series road-length models that attempt to identify saturation of feature edits in a way that is sensitive to the difference in contribution rates from country to country. The approaches are visualized in Figure 5 below. Although the approach for this paper is novel and results in effective estimates of the completeness of the maps, it focuses almost entirely on roadways. Although roads are a critical feature for disaster response, we found in our preliminary research that in the US, roads tend to be much more well mapped than structures. The weak correlation between mapped streets and structures made this approach insufficient for this project.

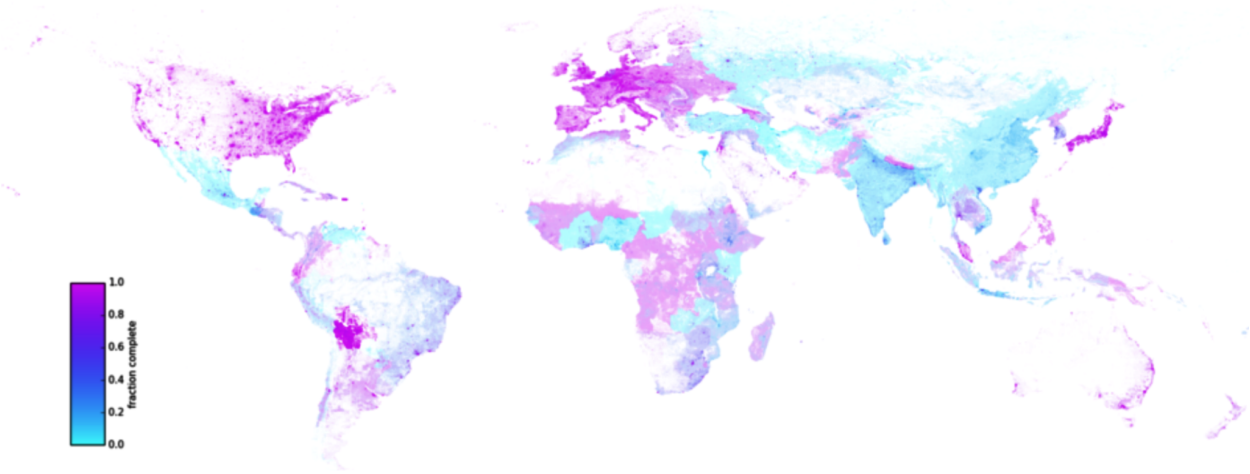


Figure 5: Completeness of the OSM dataset, by grid cell, January 2016 (Barrington-Leigh)

We also considered leveraging points of interest and cultural institutions such as places of worship, schools, 3rd places, and trying to model the relationship between the degree to which those are mapped and the overall level of mapping. This indirect method ran into similar issues, namely the establishment of a baseline.

The Third-Party Map Data Approach

Several companies offer geospatial data — companies like Google Maps, Waze, Apple, and Mapbox all provide APIs or other data endpoints. We investigated the APIs from those companies and eventually pursued Google Maps's Places API as the most comprehensive and

accessible option. We attempted to bulk extract data with the API, first trying to compare specific types of structures (e.g., schools, places of worship, gas stations) between the OSM database and the Google database, but ran into several issues with every attempt. For example, Find Place was the cheapest option and allowed for county-level filtering, but it returned only the top one result. Text Search was more expensive but returned only up to 120 results and could not be strictly filtered based on the county or on the building type. In the end, although we felt that Google Maps itself likely had sufficient data to provide a basis for comparison of OSM data, the constraints of the API made it unworkable for the scope and timeframe of this project.

Image Analysis

The final avenue of investigation was direct analysis of satellite imagery to detect and quantify structures to find those that were missing from OSM's database. There has been extensive research into the use of Convolutional Neural Networks (CNN) for building detection (Hamaguchi; Wei; Femin). The first challenge was finding an open source of satellite imagery to run a CNN model on, and the second challenge was building or finding a suitable CNN model for overhead imagery. Various parameters of overhead imagery (e.g., ground sampling distance, slant angle, type of sensor, cloud cover) can make it difficult to gather enough training data for a model to perform well. While searching for a suitable corpus of imagery, we discovered the MSAI project, which had not only sourced imagery but completed the CNN model and generated building footprints as well. The Rapid OSM editing tool (https://rapideditor.org/), originally developed by Facebook, integrates a base OSM map with additional tooling with data from the Facebook Roads database and the Microsoft Buildings database (Figure 6). As noted above, the MSAI database served as our baseline ground-truth dataset.

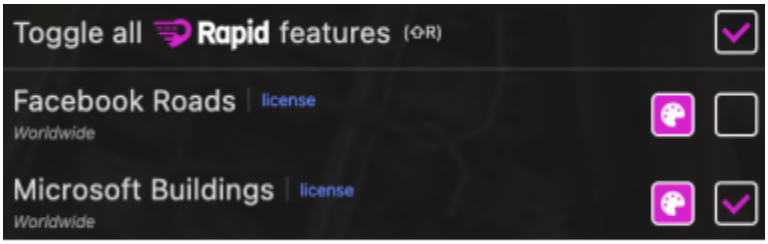


Figure 6: Rapid Editing Tool User Interface

MSAI Challenges and Limitations

Though the MSAI dataset was thorough, using the dataset still presented challenges. The most unfortunate limitation of the dataset was its inconsistent data vintage. For large swaths of the country, as shown in Figure 7, the base satellite imagery is out of date. The orange regions indicate areas that were recalculated with imagery from 2019-2020; areas outside of those regions were calculated with older Bing imagery with an average imagery source year of 2012.

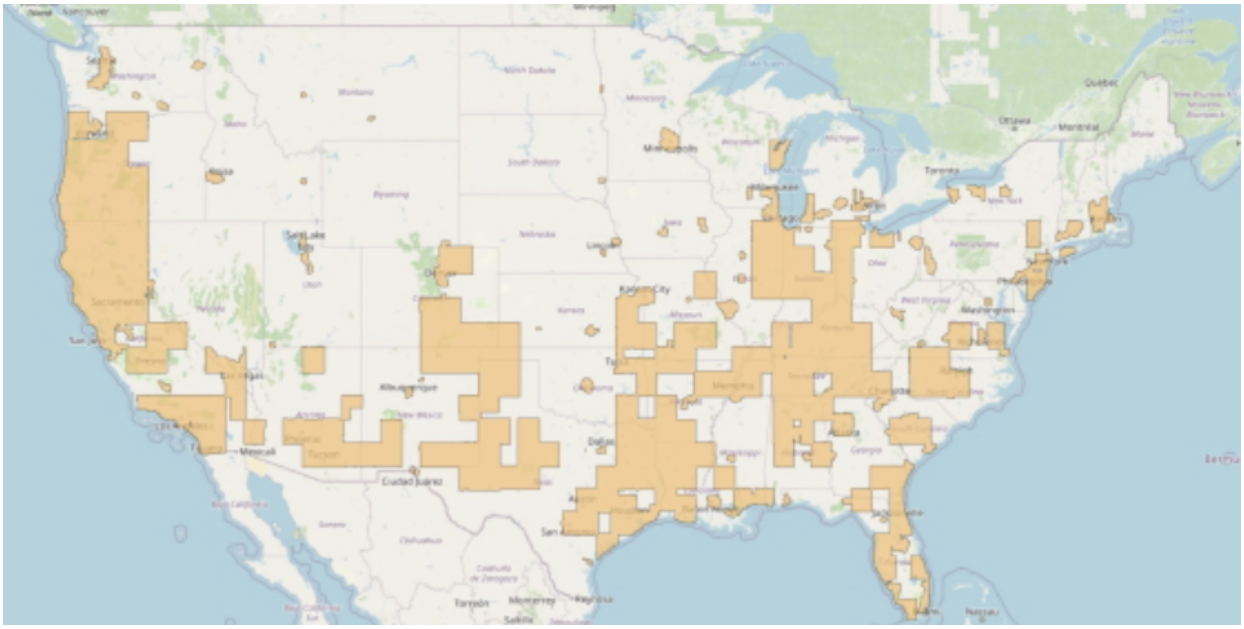


Figure 7: Map of Most Recent Satellite Data

The inconsistency of the data vintage impacts the accuracy of the undermapped statistic, which assumes that the MSAI data is the ground truth. For areas with old enough imagery, the OSM database will have more up-to-date information than the MSAI data. For more rural, heavily undermapped areas, the statistic is still fairly accurate, but for more urban areas with a greater OSM contributor presence, the undermapped statistic is inaccurate and sometimes negative, indicating more information in OSM than in the MSAI dataset.

Another limitation of the MSAI data accuracy is that certain shapes are simplified during the segmentation process. For example, structures with courtyards are covered in one simple polygon within MSAI, but more accurately segmented in OSMA. As shown in Figure 8, the OSM footprints show up as blue, Microsoft AI as red, and areas of overlap are purple. This level of detail can cause the MSAI to overinflate the area of a building footprint.

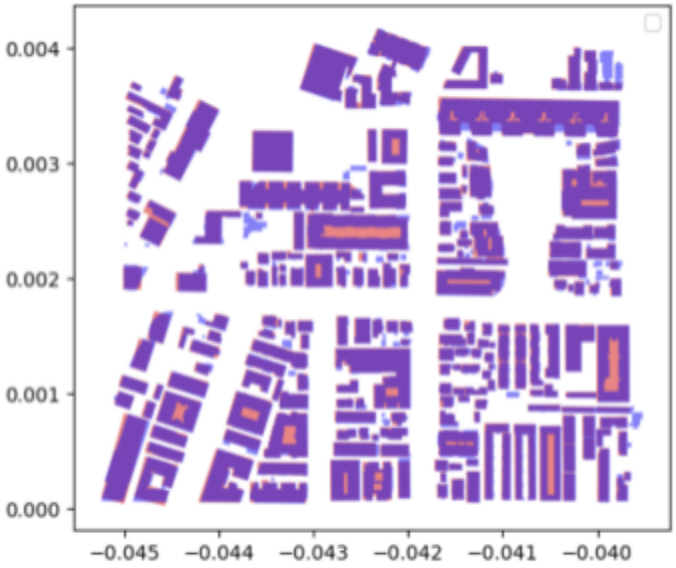


Figure 8: Overlay of OSM and MSAI Polygons

Finally, as mentioned previously, the Microsoft model has a recall of 94.0%, which does indicate that some structures are missed. As shown in Figure 9, several houses in Shaker Heights, OH, are unidentified (not marked with a purple polygon), despite the vast majority of houses being successfully marked.

Figure 9: Unidentified Structures in the Microsoft Data Set

Results

Our final dataset includes the raw areas from each dataset, the undermapped statistic, and the risk statistics for several natural disasters at varying ratios. This enables further customization of statistical visualization, where each statistic can be pulled to gain a more tailored understanding of the characteristics of different regions of the US and of different natural disasters. The Appendix contains a series of these visualizations that can be utilized without needing to run any code and provide the basic info needed to assess any of the primary risk types.

The visualizations follow explainable patterns for disasters. For example, in Figure 10, we still see the underlying distribution of risk for avalanche in the Rockies, or for coastal flooding along the east and west coasts. Although seemingly obvious, these maps provide another tool for visualizing these trends. Additional ratios for these two disaster types are available in the Appendix.

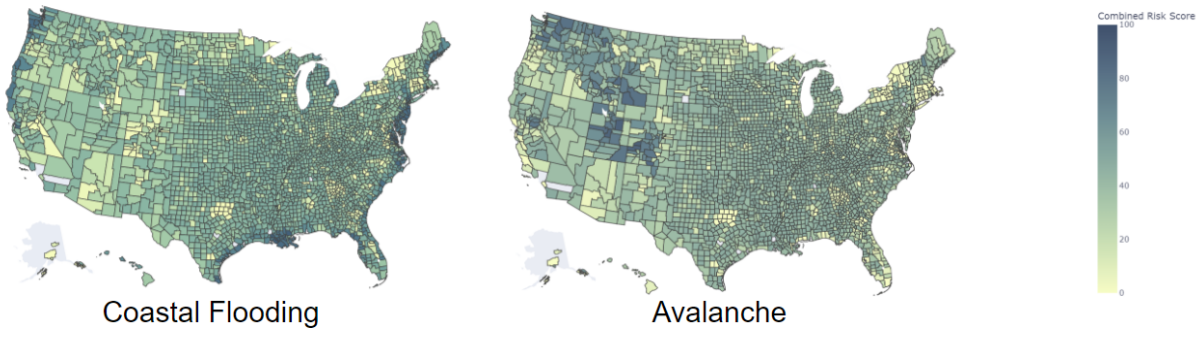


Figure 10: Explainable Disaster Patterns for Coastal Flooding and Avalanches

OSM can leverage this data to identify and assign mapping tasks within the OSM contributor community. The OSM Tasking Manager is a tool that connects mappers with critical needs, typically and often including urgent mapping tasks following natural disasters. The output from our analysis identifies and targets high risk, undermapped counties that are ripe for preemptive mapping tasks. The identified counties would then be vetted by subject matter experts and support the determination of preemptive mapping projects.

Discussion

This project has demonstrated the feasibility of combining human-generated building footprints (OSM) and computer-generated building footprints (MSAI) with a use-case dataset (FEMA NRI) to generate an intuitive yet effective metric for understanding a region's combined level of mapping within OSM and level of risk from a given natural disaster. The data generated from this project can be used as a tabular dataset or in an interactive visualization to determine at-risk regions and allocate mapping resources preemptively, before a natural disaster strikes. This project is particularly relevant today as the threats from natural disasters increase in both their destructive power and frequency. This puts an even heavier burden on the OSM network to respond proactively, rather than reactively. The ability to identify and map regions prior to a disaster has the ability to save lives and conserve the resources of first responders.

One unanticipated outcome of the project was that the visualization quickly showed the degree to which more rural areas of the US tend to be undermapped. Areas with higher population density also tend to have higher numbers of contributors on OSM, leading to better mapping of those areas. Though this was held to be true anecdotally, the visualization provides another tool to convey this information.

We recommend that future studies address the limitations faced in this project to greatly improve accuracy and reduce dependence on external datasets with unclear data vintage. Improving the existing model or training a new custom computer vision model on up-to-date imagery would provide the most accurate information on current building footprints for a better baseline dataset. Another approach is to determine the areas in the MSAI dataset that are most up to date, then combine the data from those areas with external data, such as Census data (e.g., population, average age, household income), to train a machine learning model to then predict the baseline area for out-of-date areas. Future studies could entirely redefine the undermapped metric and choose a different baseline dataset — such as social media activity, nighttime light intensity, or USPS address data — to ensure that gaps in the current approach could be mitigated.

Despite some of the challenges related to the underlying data set and the code we have created, we believe this will be a valuable tool that is accurate enough to produce positive results for the OSM community and directly affect emergency response after a natural disaster.

References

Barrington-Leigh, C., & Millard-Ball, A. (2017, August). The world's user-generated road map is more than 80% complete. PLOS ONE, 12, e0180698. https://doi.org/10.1371/journal.pone.0180698

Federal Emergency Management Agency (n.d.). National Risk Index. Retrieved from https://hazards.fema.gov/nri/

Femin, A., & Biju, K. S. (2020). Accurate Detection of Buildings from Satellite Images using CNN. In 2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE) (pp. 1-5). https://doi.org/10.1109/ICECCE49384.2020.9179232

Hamaguchi, R., & Hikosaka, S. (2018). Building Detection From Satellite Imagery Using Ensemble of Size-Specific Detectors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (pp. 1-5). June 2018.

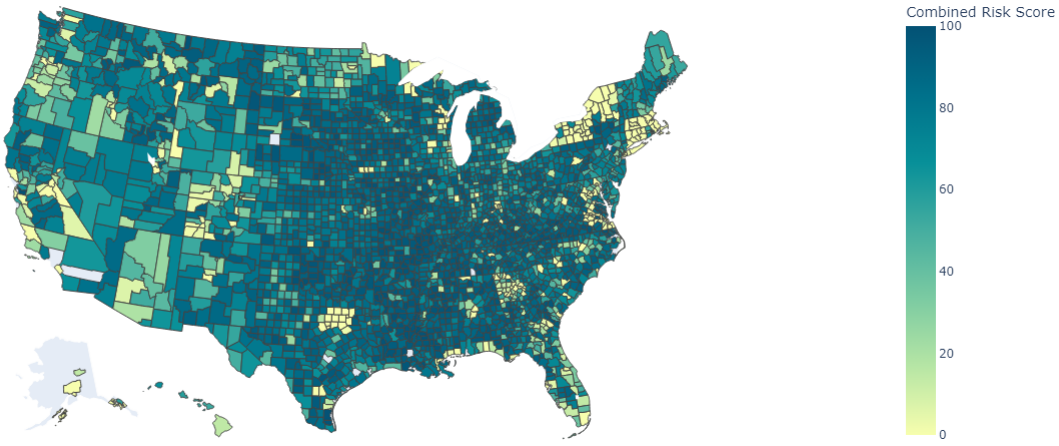
U.S. Census Bureau. (n.d.). ANSI Codes. Retrieved from https://www.census.gov/library/reference/code-lists/ansi.html

Wei, S., Ji, S., & Lu, M. (2020). Toward Automatic Building Footprint Delineation From Aerial Images Using CNN and Regularization. IEEE Transactions on Geoscience and Remote Sensing, 58(3), 2178-2189. https://doi.org/10.1109/TGRS.2019.2954461

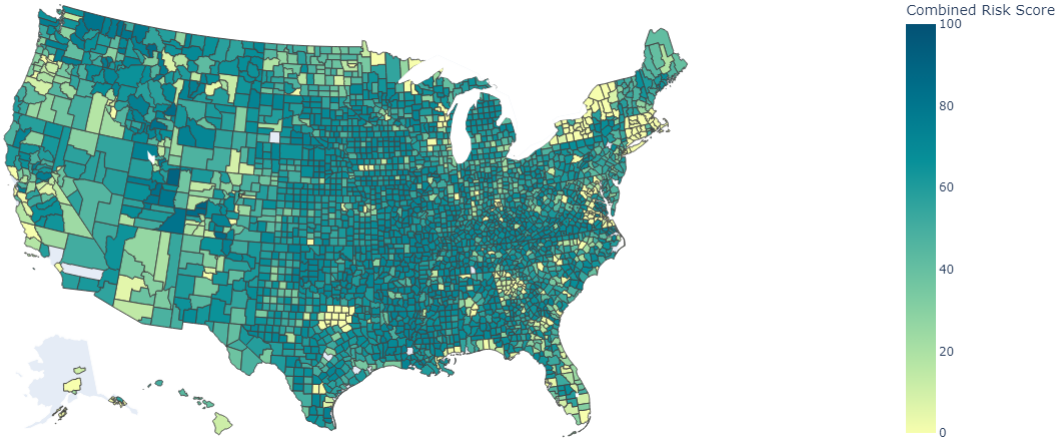
Appendix:

The following maps show the range of Priority Statistics for Avalanches and Coastal Flooding from 0% Risk, 100% Undermapped, to 100% Risk and 0% Undermapped in increments of 0%, 25%, 50%, 75%, and 100%.

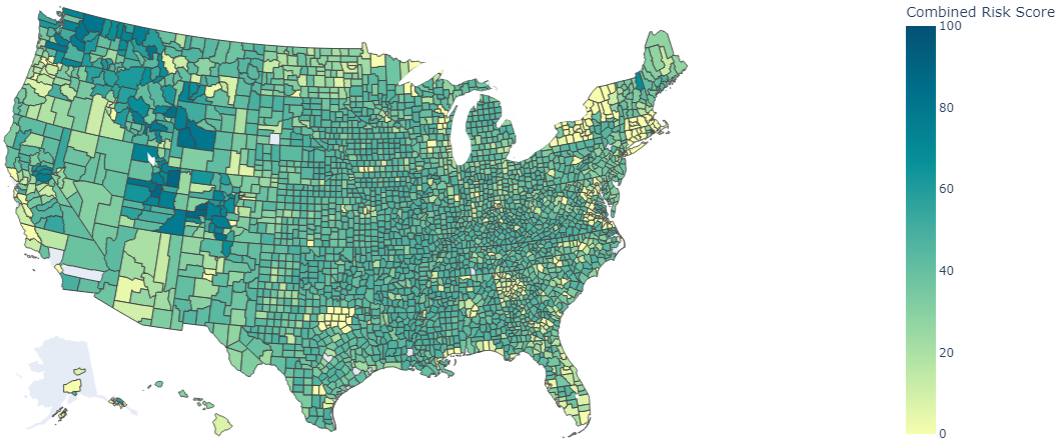
Avalanche



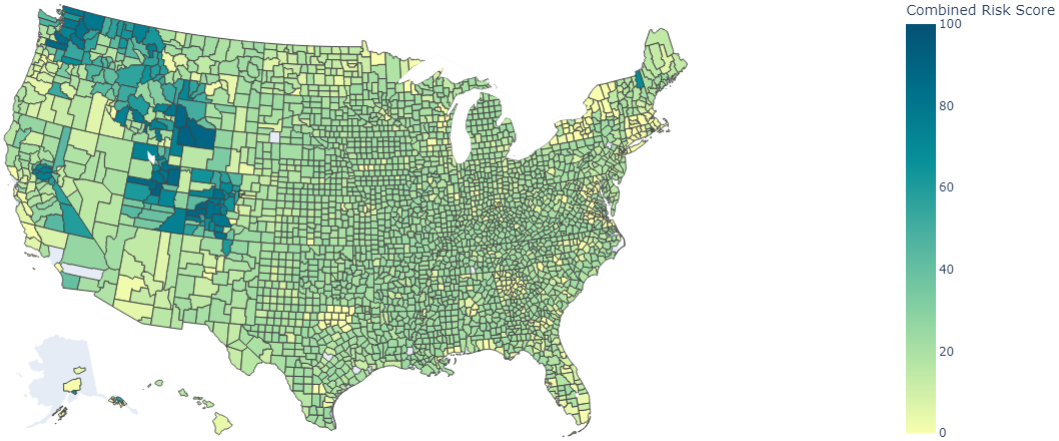
100% Undermapped - 0% Risk



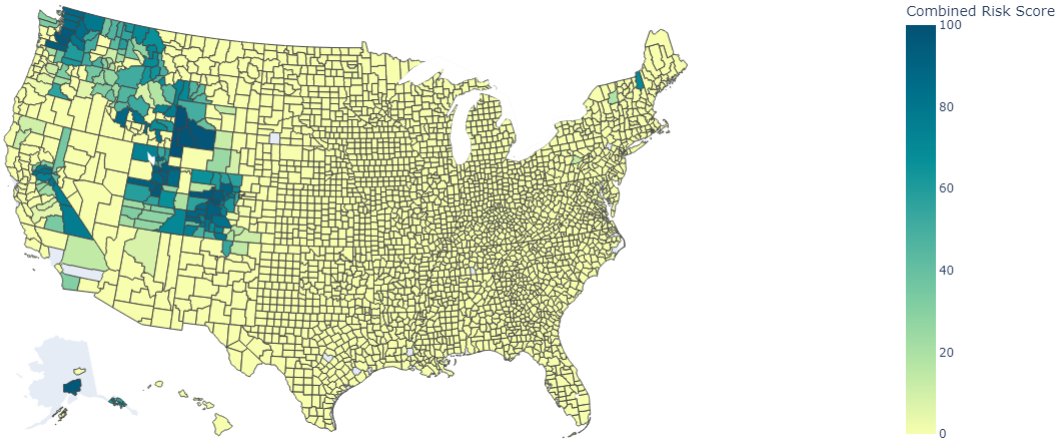
75% Undermapped - 25% Risk



50% Undermapped - 50% Risk

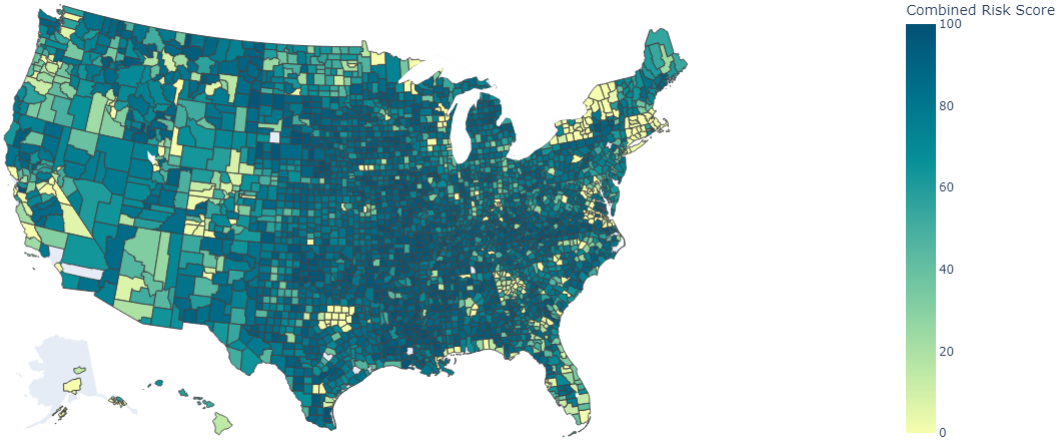


25% Undermapped - 75% Risk

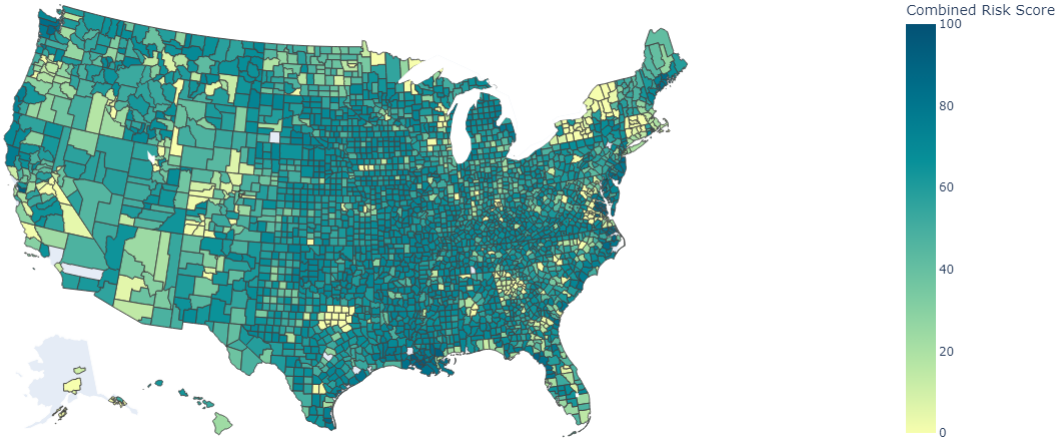


0% Undermapped - 100% Risk

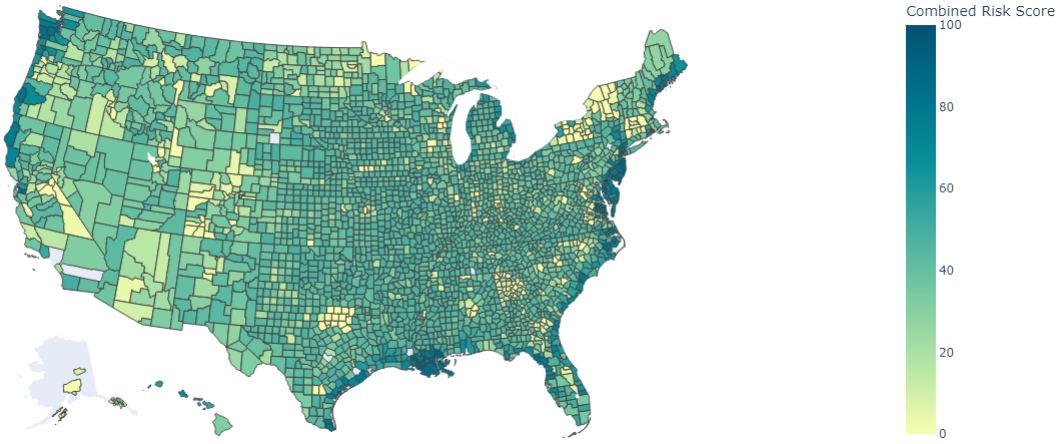
Coastal Flooding



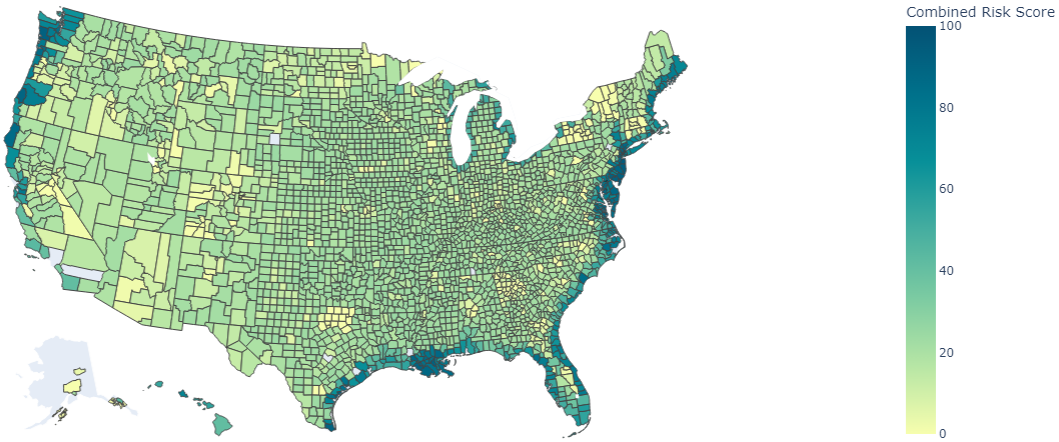
100% Undermapped - 0% Risk



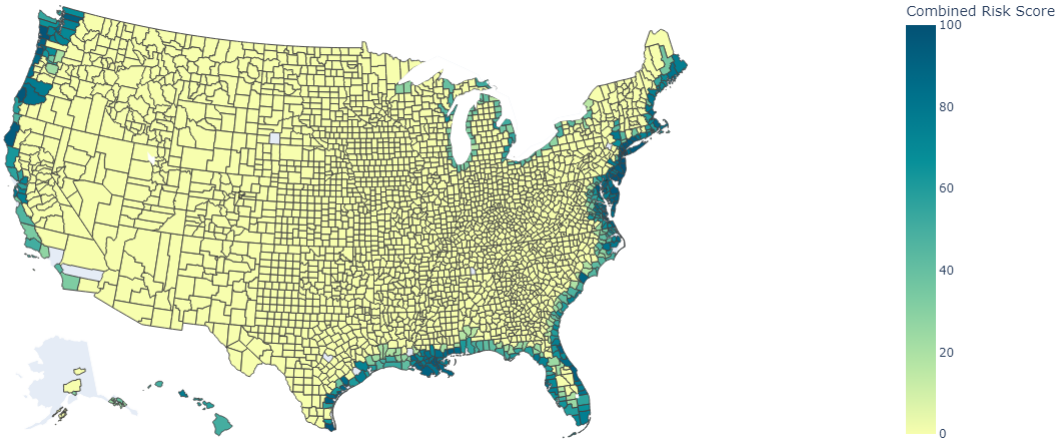
75% Undermapped - 25% Risk



50% Undermapped - 50% Risk



25% Undermapped - 75% Risk



0% Undermapped - 100% Risk