Demographics of Disaster

STAT 259 Final Project
Data Science for 21st Century
NSF National Research Traineeship

Authors:
Laura Alexander
Jenna Baughman
Matthew Kling
Dana Seidel
Carmen Tubbesing
Valeri Vasquez
Christine Wilkinson

University of California, Berkeley
Spring 2016

Contents

List of Figures

1 Defining the Question

1.1 Concept

Our project began with a highly theoretical question: given the multitude of environmental risk factors inherent to different geographic regions of the United States, where should human populations be concentrated? And given the realities of climate change, would such "ideal" locations shift over time?

The original functioning hypothesis was that, as rational beings, people should choose to live in areas that enjoyed relative natural serenity - places with as little chance as possible for wild fires, tornadoes, floods, and other such disasters. Our alternative hypothesis was that people would prioritize geographic regions that featured great natural beauty, regardless of risk. A second alternative hypothesis was that inertia would overtake rationality - namely, that people would tend to stay in one place regardless of the increasing risks that accompany climate change, be it for monetary, emotional, or other reasons.

As our thinking evolved, we began to consider how natural disasters might impact different socioeconomic and racial segments of the population. We initially posited that low-income groups and ethnic minorities would tend to be disproportionately affected by extreme events. We put this proposal forward under the assumption that, due to lesser resources (economic, political, etc.), such groups would be constrained in their abilities to either move away from riskier areas of the country or afford living expenses in safer geographic regions.

While the analytical work we have done for this project stops short of assigning causation, we are able to demonstrate a number of correlations between specific subsets of the population of the United States and their respective levels of exposure to environmental risk. We call this the "demographics of disaster."

1.2 Approach

Once we had modified the research question, the sheer scope of this project required us to carefully define our terms. Namely, how narrowly should we categorize race? What types of natural events should be included in our analysis? By what metric should impacts be assessed?

Our decisions on these fronts were in part dictated by the nature of the topic, and in part by the availability of data. For example, on the question of race, we debated the best means of including people who identify as Hispanic given that this group might be comprised of representatives from any number of racial categories. Ultimately, we opted to omit the explicit inclusion of the "Hispanic" identifier as a racial category, and instead included it as a separate variable in which the only two values are Hispanic and Not Hispanic.

To furnish another example: one pertinent element in any comprehensive calculation of natural risk factors should be the location and volume of flooding in the contiguous United States. However, this information was not publicly available at the level of granularity that we wished to conduct our analyses. Manipulating the data that was available would have distorted our final product. While flooding is not included in our analysis, a major cause of flooding, namely hurricanes, is included.

When it came to assessing impacts, we again found ourselves slightly constrained by the availability of information for certain data sets. Therefore, rather than defining impact as mortality levels or economic damages, we chose to assess the cumulative risk of a given geographic area as the combined probability of a set of severe natural disasters, defined roughly as a disaster of sufficient magnitude to cause damage to whatever infrastructure may be located in its path.

Our terms established, we began formatting our cleaned data for the purpose of establishing the necessary correlations. First, we created a master data set to include all elements of the demographic and environmental information points that we had collected. This enabled us to create a single script that spatially projected

these national-level data sets at the county level.

We split the aforementioned master county data set into two separate, cleaned tables: one for risk and one for social indices. For risk, we examined the distributions, considered a variety of transformations, and ultimately decided on using cumulative risk value as referred to earlier in this report. Using the `Shiny` App available in `R` afforded us the ability to present three options for transformations visually: 1) the raw linear transformation, 2) a log based transformation for use with highly skewed distributions which also drops negative values, and finally 3) a percentile rank transformation which flattens the histogram and turns it into a uniform distribution such that individual counties are ranked by percentile value.

1.3 Goal

Our ultimate goals for this project were to build our programming skill sets and to create a reproducible, interactive final product that would feature our efforts.

We collectively succeeded in honing our data collection and cleaning abilities, as well as in refining our expertise in a number of languages and tools - including `R`, `LaTeX`, and `Shiny`, to name a few. Given additional time, we might have expanded our project goals to include the development of a more definitive analysis. At least one way in which we could improve our final product is highlighted in the "Approach" section of this report; there are doubtless many other visual and analytical enhancements that could be made. For example, a valuable next step that might be taken from the platform of data and correlations that were established by this project could be to isolate a source of identifying variation that allows for the estimation of causal effects. Looking ahead, that is one potential area for expansion.

2 Finding and Cleaning the Data

We conducted a thorough search for demographic and environment data. We then converted these data to comma-separated values format (.csv) and cleaned them using various `R` packages (R Core Team, 2016). Data that were not available at the county level were also converted to the county level in `R`. For example, wildfire risk data were only available in shapefile raster format. The cleaned data were then incorporated into a master list organized by a primary key of state and county Federal Information Processing Standards (FIPS) codes. See Figure 1 for a flow chart outlining the major methods.

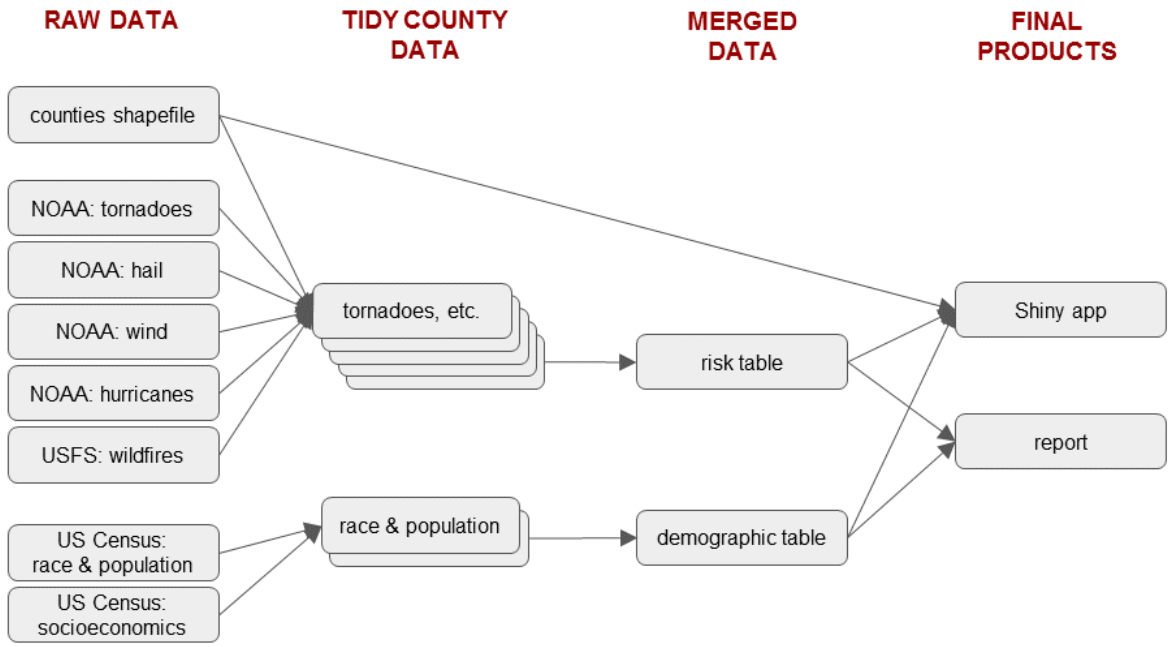

Figure 1: Project workflow from raw data to final products.

2.1 County Demographic Data

Demographic data are from the United States Census Bureau (see https://www.census.gov/popest/data/). Population estimates are available to county level from 1875 through 2014 using decadal census data and population estimates for interim years. The 2014 file included decadal census data and estimates from 2010 through 2014 with age and sex structure.

Before analysis we reduced this file to only include county estimates from 2014 with no sex or age structure. The county and state FIPS are encoded in two separate columns. Races included are African-American/Black, Asian, American Indian, Native Hawaiian/Pacific Islander, and White. People who identify as Hispanic are also considered explicitly in this data set as they intersect with all surveyed races.

2.2 Income & Employment

Income-related information is included as: median household income, in dollars, estimate of people whose income is below the poverty limit, and unemployment rate in 2014.

All three of these data sets were available on a county level from the Economic Research Service (ERS) division of the United States Department of Agriculture (USDA) (direct download link: http://www.ers.usda.gov/data-products/county-level-data-sets/download-data.aspx).

2.3 Natural Amenities

Natural amenities is quantified by the USDA by considering physical characteristics that are typically preferred and commonly understood to enhance a location as a place to live. The composite scale quantifies the following measures as natural amenities for each county: warm and sunny winters, temperate summers, low summer humidity, topographic variation, and available water area.

This data set is available at http://www.ers.usda.gov/data-products/natural-amenities-scale.aspx.

2.4 Fire

The wildland fire risk data are produced by the USDA Forest Service. The dataset consists of a raster of modeled wildfire risk on a 270 meter resolution grid across the entire contiguous United States.n.d.

According to the authors, the goal of this data is to describe the "relative potential for wildfire that would be difficult for suppression resources to contain." Thus, the dataset does not depict low-severity wildfires that pose little risk to structures and human safety, and are easily put out.

Fire risk codes represent the following risk categories:

1. Very Low
2. Low
3. Moderate
4. High
5. Very High
6. Non-Burnable Land
7. Water

In the analysis, the variable of interest is percent of each county that is in either Risk Category 4 or 5 (High or Very High risk of a fire that will escape suppression efforts).

Several data sources were used to create this geospatial dataset: vegetation and fuels (dead vegetation) data from LANDFIRE, a remote sensing effort led by the Forest Service and the US Department of the Interior; actual fire occurrence ca. 1992-2012; and a model called FSim that predicts fire occurrence based largely on weather and fuels (Finney, McHugh, Grenfell, and Riley, 2010).

This publicly available dataset, called the Wildfire Hazard Potential (WHP) can be downloaded from http://www.fs.usda.gov/rds/archive/Product/RDS-2015-0046.

2.5 Hurricane

Our hurricane data come from NOAA's HURDAT2 database for the Atlantic region, and represent the tracks and windspeeds of every hurricane since 1851. Beginning with this raw data, we:

1. Restructure the data into a tidy tabular format.
2. Interpolate the raw 6-hour time steps to a one-hour frequency to better represent continuous storm paths. Geographic coordinates and windspeeds are linearly interpolated.
3. Derive a measure of destructive force by cubing maximum windspeed at each timestep (force is proportional to the cube of velocity).
4. Perform a GIS query, combining storm coordinates with a counties shapefile to sum these force values for all points falling in each US county.
5. Divide this total intensity by county land area, to correct for widely differing county sizes and better represent the likely per capita exposure to hurricanes.

The direct download link for the dataset is http://www.nhc.noaa.gov/data/hurdat/hurdat2-1851-2015-021716.txt.

2.6 Tornado, Wind, and Hail

Data for these three severe weather types come from NOAA's storm prediction center, and share a common format. The raw data represent individual historic instances of storm events since 1950, organized by county. For each of these three weather types, we derive an index of each county's exposure using a procedure very similar to the hurricane procedure outlined above: windspeeds (or hail sizes) are cubed, summed, and then divided by county area. By happenstance, physical destructive force is proportional to the cube of hail diameter and also the cube of wind speed, which allows this approach to work for all three storm types.

These data can be found here: http://www.spc.noaa.gov/wcm/.

3 Conducting the Analysis & Visualizing the Results

Given that our analysis was entirely correlative in nature, the group felt that the optimal means of displaying our findings was to develop a compelling way of visualizing the data that we had collected. A visualization would be ideal for identifying and discussing any emergent demographic patterns in risk exposure at the county level, nationwide.

To best achieve this, we chose to employ a tool that was introduced in class and which we wanted to learn more about: the `Shiny` web application framework for `R` that was developed by `R-Studio`. This application seemed ideal in that it was fairly simple to use, it enabled us to apply a variety of transformations to the data we had in hand, and it provided a public interface for our findings.

In order to best make use of `Shiny`, the master data set referenced in the previous section of this report was divided into two tables for manipulation: (1) a table including all risk indices (fire, wind, hail, tornado) and (2) a table including all social metrics (population by 5 races and Hispanic ethnicity, poverty, aesthetic value, unemployment, and income).

We then developed and ran a second, separate script on the risk table to be able to calculate cumulative risk - the metric that we had decided to use for this project - across the entire contiguous United States. In doing this, distributions of risk factors were considered, and all fields standardized, before summing the cumulative risk metric by county. All correlations and log transformations were run internal to `Shiny`, as previously noted. The working `Shiny` App can be accessed at the following link: https://matthewkling.shinyapps.io/demographics_of_disaster/.

3.1 Notable Findings

Our mapping exercise demonstrated that minority and low income populations tend to bear a slightly higher cumulative risk exposure than their whiter, wealthier peers (Figure 2). This correlation is visible in two ways on each iteration our textttShiny visualization product, as users click through different combinations of natural disaster risk factors and demographic groups: first, through the slope of the fitted line on the chart to the left, and second through the coloration of the map on the right.

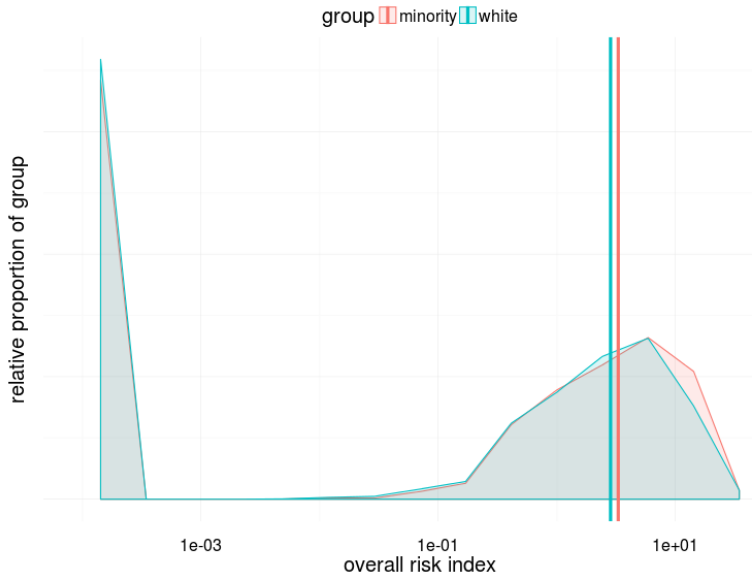

Figure 2: Overall risk index for white and minority social groups made through the `Shiny` app.

Some of the most striking patterns highlighted by our exercise include:

- The following map and its associated plot, with an 0.8 slope, demonstrates the correspondence between the highest hurricane risk and highest concentration of African-American/black people, as indicated by the yellow and other warm colors along the southeastern U.S. coast. Shown in Figure 3.

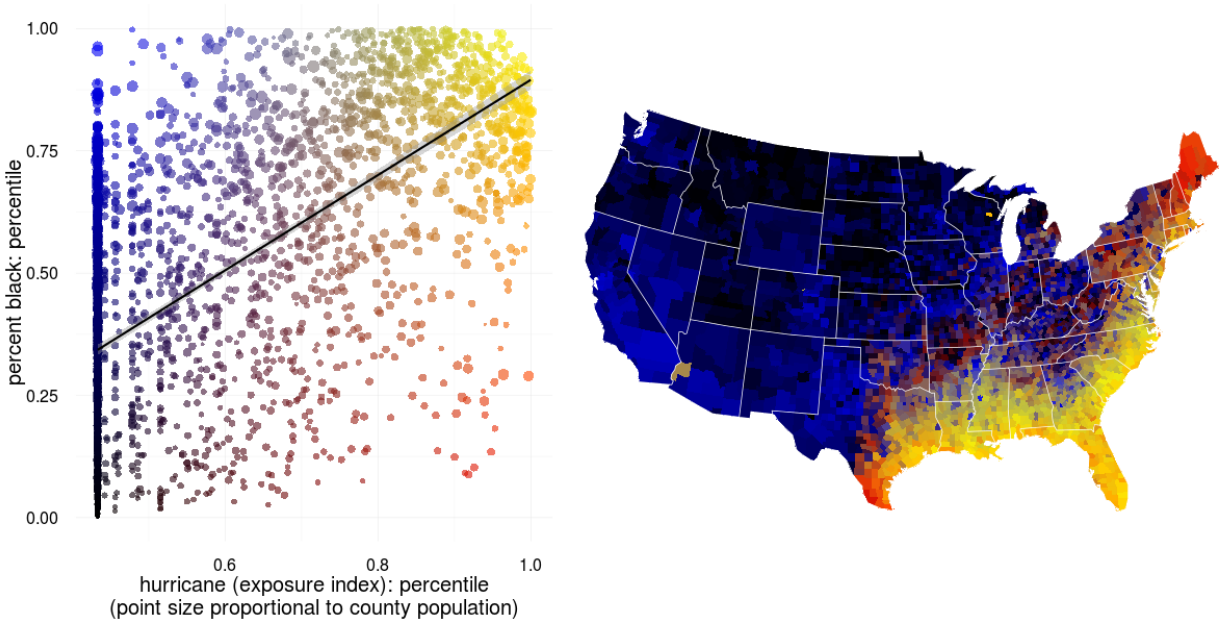

Figure 3: Example correlation and heat map of percent black vs. hurricane exposure risk made through the `Shiny` app.

- The large variation in land area among U.S. counties (Figure 4) and how that might affect comparisons using counties or other geopolitical units.

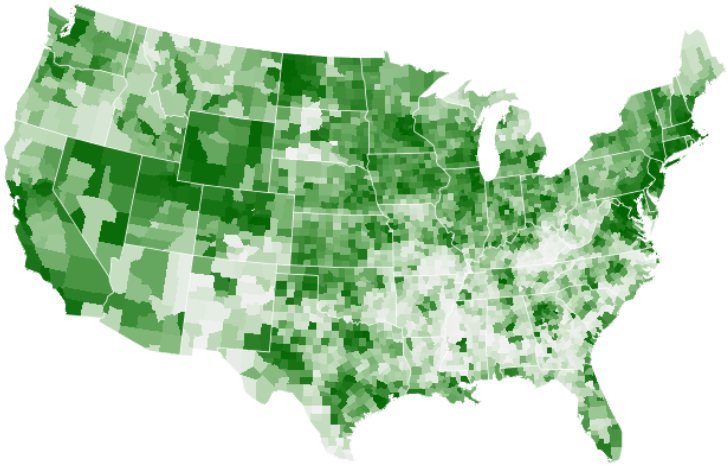

Figure 4: Median household income variation across counties of drastically different land area.

In particular, we were struck by the results of comparing percent American Indian with wildfire risk. While this correlation may not be inherently surprising as we may have predicted higher incidence of both of these variables in the western U.S., the emergent pattern was one that is perhaps not commonly discussed. The image generated by that correlation speaks for itself; shown in Figure 5. This comparison also highlights that there exists another high wildfire risk area in the Southeast in which American Indians are not as highly represented.

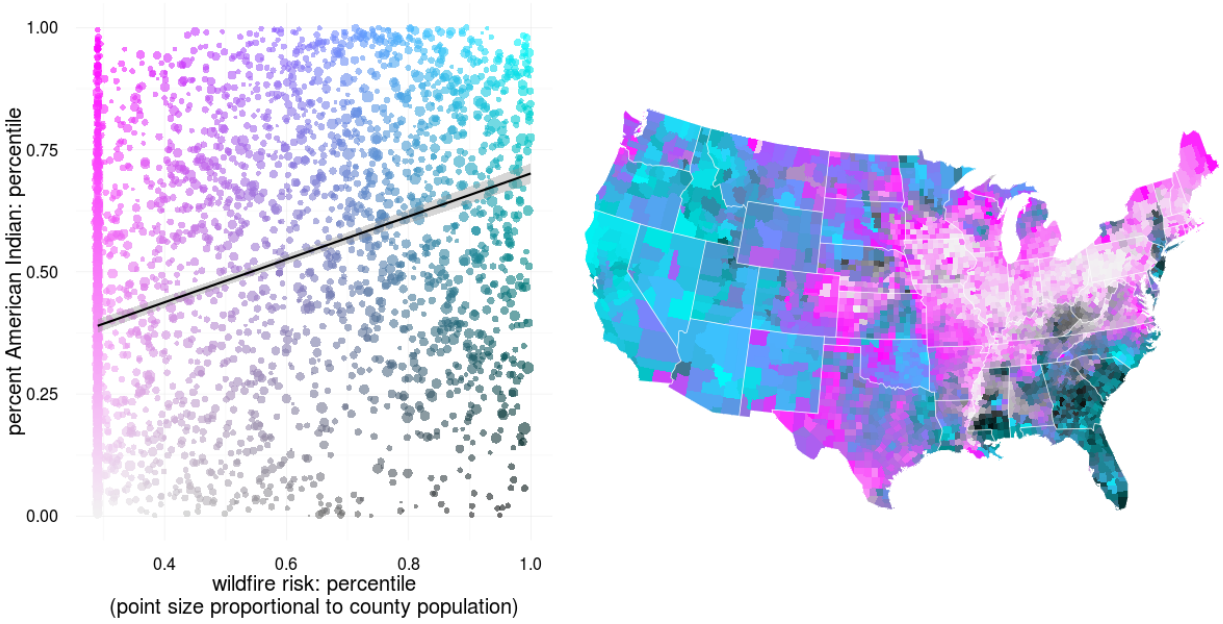

Figure 5: Example correlation and heat map of percent American Indian vs. wildfire risk made through the `Shiny` app.

4 Drawing Conclusions

The DS421 project has provided an opportunity for our group members to learn valuable tools in data science, and in many cases has allowed us to meet personal goals for mastering the basics of a given language, action, or application. It has also taught each of us the importance of reproduciblity - particularly when attempting to work efficiently and effectively with large numbers of people - and afforded us the chance to push ourselves into interdisciplinary approaches and conversations about climate change, race, and geography.

Causation is difficult to ascribe given the complex local and global forcings that influence both social and risk factors. However, a more temporally dynamic analysis - for example, expanding our examination to include the entire century rather than just a single year - would potentially allow us to draw stronger conclusions about correlation between where minority populations live and risk of catastrophe.

In many ways, reviewing the data central to this project has furnished us with the beginnings of questions rather than the answers that might lend themselves to a more satisfying report conclusion. While we have focused here on the correlations between social and natural aspects particular to the United States, a few minutes of toying with the `Shiny` visualization brings to light a number of other potential correlations in the accumulated data - some of which are purely social. For example, the app paints an interesting picture of the correlation between race and income level, as well as between population density and poverty. Further, seeing the data projected on a national map, and then at a state and county level of granularity, clearly demonstrates a limitation that carries implications for any correlative conclusions: there is a huge variation in the size of discrete counties throughout the United States. Because counties are not proscribed by population or area, we can only draw inferences if we ignore this variation at the sub-county level.

5 References

Dillon, G. K. (n.d.). Wildfire Hazard Potential (WHP) for the conterminous United States (270m GRID), version 2014 classified. doi:10.2737/RDS-2015-0046

Finney, M. A., McHugh, C. W., Grenfell, I., & Riley, K. L. (2010). Continental-scale simulation of burn probabilities, flame lengths, and fire size distribution for the united states.

R Core Team. (2016). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org/

6 Acknowledgments

Many thanks to Gaston Sanchez for his dedication to introducing us to the basics of reproducible data science, and for his great patience as we went through the learning process of putting those tools to use. Thanks also to Phil Stark - who offered his wisdom on questions of data interpretation and statistical analysis - as well as David Ackerly, Heather Constable, and the entire DS421 faculty and staff. It was a privilege to participate in this inaugural year of the program; we look forward to the coming semesters!