Optimizing Local Smoke Alarm Inspections with Federal Data

Jeremy Krinsley
Engineer, Enigma
jak@enigma.io

Brian Abelson
Engineer, Enigma
brian@enigma.io

ABSTRACT

This paper outlines a fully-realized civic tool that predicts municipal blocks least likely to have homes with functioning smoke alarms and most likely to have residents who are at highest risk for fire fatality. Using a novel merge of the American Community Survey (ACS) and the American Housing Survey (AHS), we are able to model these two risk factors at the geography of census block groups, and with the aid of the TIGER Census dataset, return actual street addresses with associated risk scores. This tool represents a potential model for developing reusable civic analytic applications that can serve multiple cities while responding to local particularities.

Keywords

Civic Analytics, Fire Prevention, Census, ACS, AHS, TIGER

1. INTRO

We can expect 20,000 people to be injured or killed by fires in the United States this year. With over 130 million housing units across the country, 4.5 million of them do not have smoke detectors, placing their inhabitants at substantial risk. Driving this number down is the single most important factor for saving lives put at risk by fire.

A broad range of people are trying to address this problem, from local fire departments to the Red Cross. However, they all face the same problem: What door do we knock on first?

A community knows best which blocks and which citizens are at risk for injury or death in a fire. First responders and outreach organizations know the facts of their municipality better than a dataset ever can. Firemen know their districts; Red Cross volunteers and religious organizations know their local residents. These are the instincts that up to now have been the primary, if not the only resources available to identify at-risk homes and residents.

With that in mind, we have gone farther than any previous effort to rank likelihood of fire risk by city block in the hopes that our work can offer more precise, granular leads for those who perhaps had no such prior resources. What we are able to present is a systematic ranking for 209 cities that fall within 40 census-tabulated Metropolitan Statistical Areas. We have furthermore developed a tool that allows these same first responders and active community members to improve the accuracy of local scores by uploading fire data into the model.

Prioritizing outreach efforts for fire prevention can be difficult without access to data and analytics that identify houses at risk for not having a smoke alarm. By drawing together many different public data sets, we have developed a predictive model that helps identify hot spots for homes that are unlikely to have smoke alarms.

Our goal is to provide a tool that helps fire departments and other groups work more efficiently. Local fire department outreach coordinators can combine their on-the-ground knowledge about their areas with trained models scored at a newly enhanced geographic granularity. Before this project, this resolution simply did not exist.

This is only a first step. We are also releasing all of the data, components and algorithms that make this tool work in hopes that others can improve upon what we have begun.

2. COLLABORATORS

In November of 2014 there was a house fire in the Broadmoor neighborhood of New Orleans, killing five people, including three children. The house did not have a smoke alarm in it. Enigma began working with New Orleans' Fire Department and Office of Performance and Accountability to develop a model that identified New Orleans blocks least likely to have smoke alarms, and most likely to experience a fire fatality. This enabled the New Orleans Fire Department to conduct a door-to-door outreach campaign to place smoke alarms in as many at risk homes as possible. Drawing on the learnings from New Orleans, we extended the model to apply to the entire United States, in hopes that more people can use and improve on our insights.

Data Kind /Red Cross

3. COMPONENTS

The risk model and resulting tools were built off of a number of different federal census datasets. Below we outline these datasets and the ways in which we used them.

3.1 AHS/ACS Merge

Data For Good Exchange, 2015

The central crux of our work employs a systematic join of the American Housing Survey (AHS), a resource for nationally-representative, detailed housing characteristics, and the American Community Survey (ACS), the census' most extensive and thorough demographic survey. This merge generates meaningful local data from a Federal dataset that was once used only to describe characteristics normalized at the level of entire cities.

Making the ACS-AHS merge a machine-readable process greatly enhances the value of the data in the American Housing Survey, and by extension, enriches what we can learn about a given few blocks within the biggest metro-areas in the U.S. We are not the first[1] to attempt a method for explaining the relationship between the two datasets. However, to the best of our knowledge, we have gone the furthest in making that process programmatic.

Many questions asked in the AHS directly mirror those asked in the ACS. For instance, both surveys ask about a respondent's age. While the ACS groups these responses into bucketed counts per block group (i.e. ”Males under the age of 5” or ”Females between the ages of 70 and 74”), the AHS simply records the respondent's actual age (4 or 72). While both surveys capture the same concept, they each record this information in different ways. Our merge enables translations between the two surveys by mapping these concepts into a common schema. This is useful as it allows models trained on AHS data to be scored on ACS data. For an example of such an application, check out this project.

We came up with these mappings by scanning the AHS Codebook for questions that were also asked in the ACS. For the most part, these were demographic variables and information about a respondent's household. In the case of the AHS, responses are recorded in categories (i.e. ”married” = 3 and ”divorce” = 4), or continuous numbers (”year house built” = 1964). On the other hand, as mentioned above, the ACS buckets responses by counts for census geographies like block groups and tracts. In order to map the two surveys, we had to similarly bucket AHS responses into binary indicators. Here's an example of what this process looks like for ”Marital Status”:

```
- concept: Marital Status
  ahs:
        type: categorical
        var: hhmar
        map:
            hhmar_married_spouse_present: 1
            hhmar_married_spouse_absent: 2
            hhmar_widowed: 3
            hhmar_divorced: 4
            hhmar_separated: 5
            hhmar_never_married:  6

  acs:
        table: B12001
        total: B12001001
        map:
            hhmar_married_spouse_present:
            ↪  [B12001005,B12001014]
            hhmar_married_spouse_absent:
            ↪  [B12001006, B12001015]
            hhmar_widowed: [B12001009, B12001018]
            hhmar_divorced: [B12001010, B12001019]
            hhmar_separated: [B12001007,
            ↪  B12001016]
            hhmar_never_married: [B12001003,
            ↪  B12001012]
```

3.2 Risk Model

3.3 TIGER Geocoder

We used the Census TIGER Geocoder in order to map our risk scores to specific geographical boundaries within the cities we analyzed. We also used the TIGER data to create a master CSV of all street blocks within these geographies, organized by address ranges, so that we could associate addresses with risk scores. We also are releasing a version of TIGER that makes it easy for anyone to spin up a geocoder[2].

3.3.1 Geographic Join

Each record in the American Community Survey maps to a census block group, the smallest geographical entity deemed to still present valid sample data, which can range from a population of 600 to 3,000 citizens. Of the total 220,740 block groups in the U.S., we generated relevant scores for the NUMBER with a population density over RATIO that fell within one of the NUMBER MSAs we deemed data-rich enough to provide reliable scores.

Do we want to go into detail here about the actual PostGIS queries or logic?

3.3.2 Creating Address Tables

The TIGER data collection includes a comprehensive ADDRFEAT table that, among other properties, lists every block in the U.S., and includes such features as the name of the street and the address range for the left and right sides of the street. Using geospatial querying in PostGIS, we grouped these address ranges by census block groups, and thereby joined every city block and address range to the relevant census block score in our model.

4. TOOLS

4.1 Address/Score CSV

The result of the geospatial joins between our model and street addresses in TIGER are CSVs for every city that falls within the census' designated Metropolitan Statistical Area. Each record in the CSV lists a street name, street 'start' address and 'end' address, a risk score, and the state and county within which the block is located, along with details about the block group associated with the street.

These CSVs were built with the intention that they would provide maximum value to local first responder and outreach communities. The data model was designed so that someone with rudimentary ability to order an Excel spreadsheet could quickly organize a list of the highest risk blocks by street name and address range for their local districts and areas of interest.

[1] http://www.census.gov/content/dam/Census/programs-surveys/ahs/publications/CombiningAHS-ACS.pdf

[2] https://github.com/enigma-io/ansible-tiger-geocoder-playbook

4.2 Interactive Map

Included in out presentation of the risk model is a slippy map that enables a quick bird's eye view of the relative risks of areas in a given community. The map is BRIEF EXPLANATION OF FEATURES

IMAGE OF MAP

4.3 Data Augmentation

Our model is built entirely from federal data. We are able to further enhance our risk scores by understanding where fires have actually occurred in the past. If someone with access to local fire incident uploads data, we will rerun our model and generate fresh scores for the community for which that data applies.

The csv must include:

- a latitude column (?latitude?)
- a longitude column (?longitude?)
- each row must represent a fire incident

Our analysis pipeline disregards any additional features in the CSV beyond these simple requirements, thereby reducing the amount of data munging required to generate enhanced scores.

5. CONCLUSION

Still a work in progress

Built without a training set

Results still coming in

More granular than anything that's come before it