A Fire Risk Index for Boston Residential Buildings

Kit Miller

Table of Contents

A Fire Risk Index for Boston Residential Buildings

Fires occur more frequently in certain types of buildings and in certain areas of Boston. What are the patterns underlying these variations and how can we apply them to identify higher risk buildings and higher risk populations? A citywide fire risk index for residential buildings could be extremely useful for city planning, fire department resource allocation, and outreach efforts aimed at mitigating risk. To that end, this paper puts forth a proposed fire risk index for Boston residential properties and census tracts.

Methods

Datasets

I utilized five datasets for my analysis of fire risk (Table 1). The first is the City of Boston's 2015 Tax Assessor's Database. This dataset contains 168,000 records for every parcel of property in Boston. With this dataset, I created a subset of the 79,420 residential properties. The second dataset contained records for 2,947 fire incidents in Boston from 2010 to November 2015. The fire incident data contained address information that I was able to match algorithmically to the third dataset, the BARI addresses database, which contains all of the unique addresses locations for every unique combination of address, zip, and X and Y in the tax assessor's database. To match fire locations to the Addresses database, I parsed Street Numbers and names to match to the Address and Zip code fields in the Addresses database. Using this method, I was able to match a location for 1,466 out of the 2,947 fire incidents. Once I matched with the Addresses database, I was then able to match the fire incidents with my subset of residential buildings. After filtered out non-residential fires, I had a confident match showing that there had been a fire in 1,227 residential buildings among the 79,420 residential buildings in Boston between 2010 and November 2015.

The fourth dataset is the CRM Database Call Records that contains a record of all calls to the Boston CRM system from 2010 to 2015. This dataset I was able to merge easily with a subset of residential properties from the tax assessor's database based on a common LocationID. The final dataset I used was the Boston Area Research Networks census tract dataset containing aggregated demographic data for 178 census tracts in Boston.

Table 1: Datasets used

Name / Source	Content	Records	Key Variables
2015 Tax Assessor's Database – BARI	Property Parcel Data for all Boston Properties	168,000	Building Style, Number of Kitchens, Number of Floors
2010 – 2015 Fire Incident Data City of Boston	Fire Incidents for Residential and Commercial Buildings	2,947	Address, Property Loss
Addresses – BARI	Unique addresses in Boston derived from City of Boston's Master Address List.	119,000	LocationID (unique)
CRM Database Call records - BARI	Records of calls to the Boston Constituent Relationship Management System	968,000	Housing, Uncivil Use
Boston Census Tract Data - BARI	Aggregated demographic data for 178 census tracts in Boston	178	Various demographic variables

New Measures

In order to calculate fire risk for residential properties, I used measures from the CRM database and created two new measures from data in the Tax Assessor's Database (Table 2). The first measurement was a simple 0 or 1 indicator of whether a building could be matched by LocationID with one of the 1,227 fires for which I had a unique LocationID. Because some preliminary analysis demonstrated that three-level "triple decker" style multi-family houses show a higher risk of fires, I created a zero or 1-based variable to indicate whether a building is a triple decker.

In addition, by merging with the CRM dataset, I was able to bring in certain measures that indicated housing issues with a building. I chose the highest level variable "Private Neglect" that encompasses 3 sub-categories. The first sub-category is the "Housing" designation, which includes issues referring to poor household maintenance (e.g., poor heating, chronic dampness) and the presence of pests (e.g., bedbugs). The second is the "Uncivil Use" designation which indicates items that reflect private actions that can negatively impact the public sphere (e.g., illegal rooming house, poor conditions of property). The third is the "Big Building" designation that indicates issues related to the upkeep of big condo and apartment buildings.

Table 2: New and "Borrowed" Measures

Name	Description	Computation
Had Fire	Indicates whether a building could be matched with a fire from the Fire Incident dataset.	Convert to integer (0 or 1).
Is Triple Decker	R_BLDG_STYLE from Tax Assessor's Database (building style for residential properties) and S_BLDG_STYLE (building style for condo buildings).	Mark a 0 or 1 if either of these columns showed a style of triple decker ("DK").
Private Neglect	Total for the number of CRM complaints indicating Big Building Issues, Housing Issues, or Uncivil Use issues.	Aggregated to the LocationID and totaled.
Not Owner Occupied	Where OWN_OCC equals "N" from the Tax Assessor's database.	Convert to integer (0 or 1).

Results

I ran logistic regressions on a number of different variables, including some other variables from the CRM database. Those additional variables included the lower level Housing, Big Building, and Uncivil Use designations mentioned above. Three variables stood out with high correlations (Table 3 and Figure 1): Big Building Issues, Triple Decker, and Owner Occupancy. And to a lesser extent, there were correlations with Uncivil Use, Number of Floors, and Housing issues. Notably, there is a very high negative correlation between owner occupancy and a very high positive correlation with a building style of triple decker. Also notable is how the number of floors and the age of a building have very little correlation with the occurrence of fire. I had also expected the number of kitchens to have a high positive correlation with the occurrence of fire, but this was not the case. Overall, when I ran logistic regressions combining Triple Deckers with Number of Kitchens, Number of Floors, and Year Built, the correlation coefficient of fire incidence with Triple Deckers was virtually unchanged, suggesting a very strong relationship with this variable.

Table 3: Relationships between Residential Building Variables and the Occurrence of Fire

Variables	Estimate	Std. Error	Z-Value	P Value
Triple Decker	0.94071	0.07082	13.28	<0.0000000000000002
Not Owner Occupied	0.80017	0.05984	13.37	<0.0000000000000002
Big Building Issue	0.6357	0.1031	6.163	0.000000000713
Uncivil Use	0.17923	0.02148	8.345	<0.0000000000000002
Housing Issues	0.142570	0. 008704	16.38	<0.0000000000000002
Private Neglect	0.134229	0.007534	17.82	0.0000000000000002
Number of Floors	0.1424	0.0112	12.71	<0.0000000000000002
Number of Kitchens	0.01290	0. 00187	6.901	0.0000000000518
Year Built	-0.0043707	0.0008861	-4.933	0.000000812

Figure 1: Coefficient Plot for Individual Residential Variables

A reasonable multivariate model predictive of Fire Risk for residential properties then would have to include (1) Whether a property is a triple decker, (2) Whether a property is unoccupied by its owner, and (3) One of the CRM designations. Although the "Big Building" category shows a high positive correlation coefficient, I chose the higher order "Private Neglect" designation because it does not skew the applicability of the Fire Risk Index towards mostly larger buildings, and it neatly encapsulates all private property-related issues. Table 4 shows coefficients for a simplified model composed of "Is Triple Decker" and "Not Owner Occupied." Table 5 shows how the addition of "Private Neglect" does not significantly diminish the strength of the other variables. In the end, the best model combined the all three variables to show a coefficient of 0.14 (Table 6).

Table 4: Coefficients for variables in Multivariate Model 1

Variables	Estimate	Std. Error	Z-Value	P Value
(Intercept)	-4.69079	-0.04919	-95.35	<0.0000000000000002
Is Triple Decker	0.86405	0.07115	12.14	<0.0000000000000002
Not Owner Occupied	0.75568	0.06008	12.58	<0.0000000000000002

Table 5: Coefficients for variables in Multivariate Model 2

Variables	Estimate	Std. Error	Z-Value	P Value
(Intercept)	-4.709628	-0.049212	-95.70	<0.0000000000000002
Is Triple Decker	0.802378	0.071695	11.19	<0.0000000000000002
Not Owner Occupied	0.654865	0.061093	10.72	<0.0000000000000002
Private Neglect	0.112588	0.007525	14.96	<0.0000000000000002

Table 6: Coefficients for Final Model chosen from Model 2

Variables	Estimate	Std. Error	Z-Value	P Value
(Intercept)	-4.348475	0.031837	-135.58	<0.0000000000000002
FIRE_RISK	0.143684	0.007073	20.32	<0.0000000000000002

Although it is difficult to visualize the strength of this relationship for a binomial variable, Figure 2 illustrates how the percentage of buildings that had fires in the last five years for each building category steadily rises according to this newly developed Fire Risk variable. Although the number of buildings with higher Fire Risk steadily decreases to less than one hundred in the highest risk category, the fire rate for those highest risk buildings increases in a dramatic, although not quite exponential manner. Although there were still two hundred eighty fires that occurred in buildings with a fire risk computed at zero, as you can see from Table 7, that number represents less than one percent of all such buildings that show minimal fire risk.

Figure 2: Relationship between Building Fire Risk and Rate of Fire in Buildings

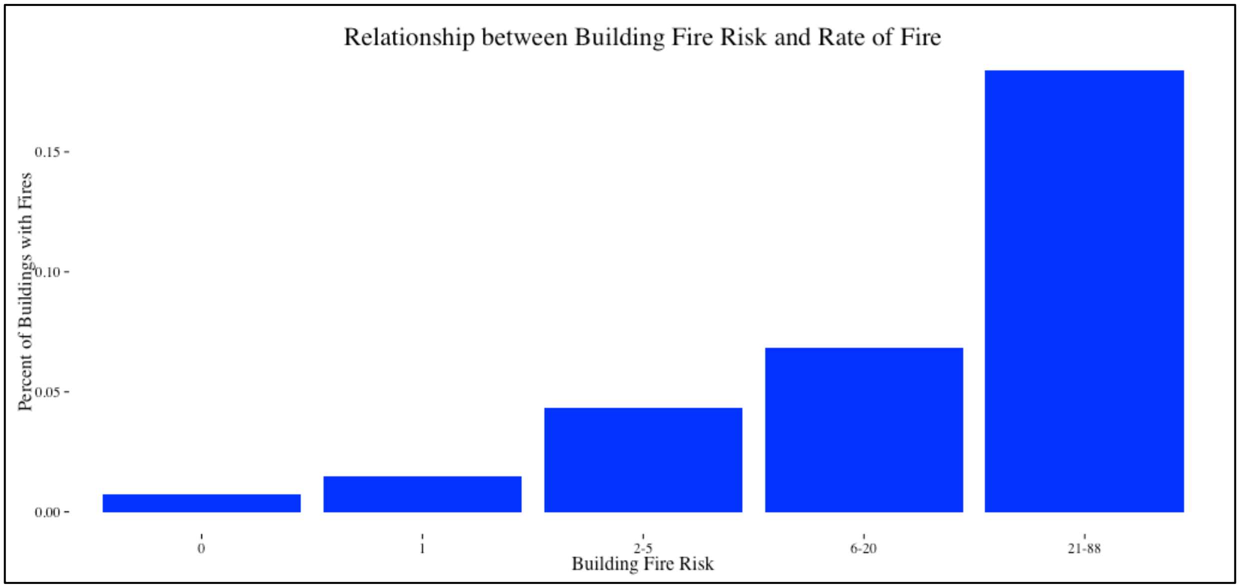
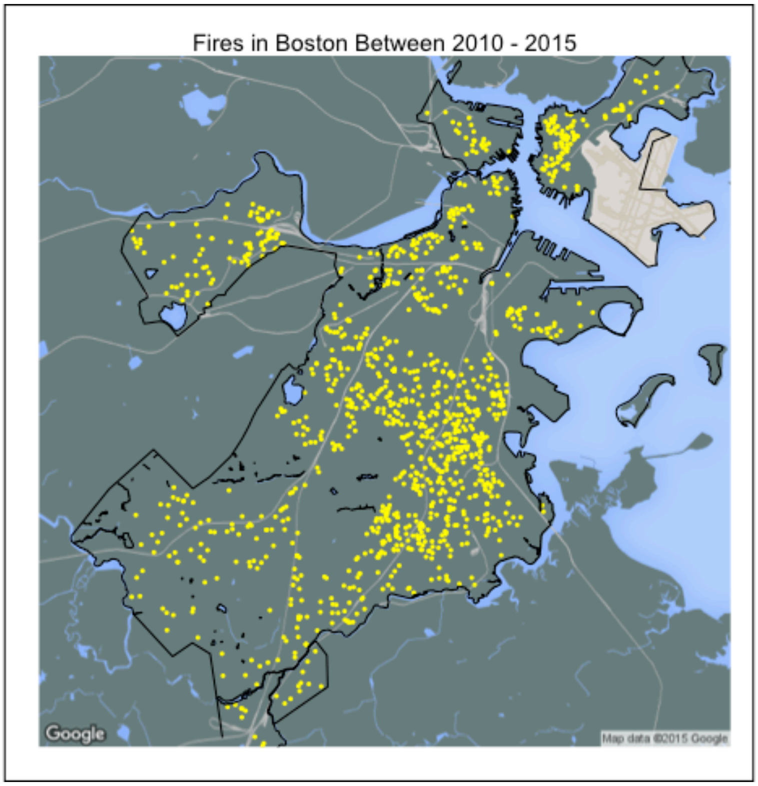


Table 4: Fire Risk Categories

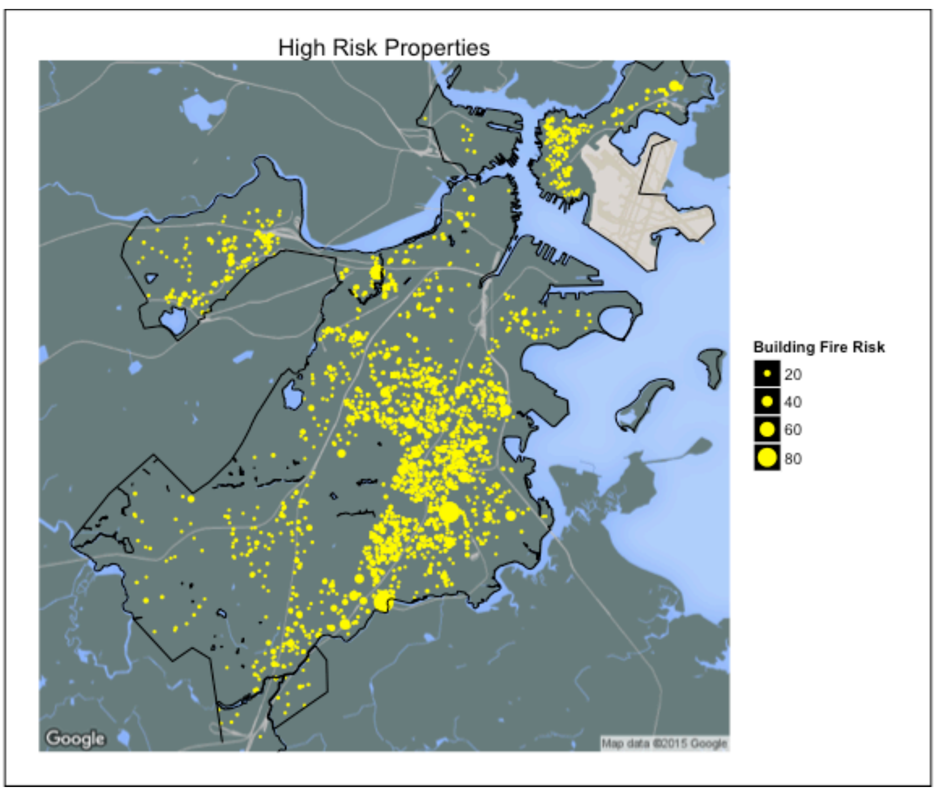
Fire Risk	Number of Buildings	Number of Fires	Percent of Buildings with Fires
0	37,517	280	0.7%
1	28,699	430	1.5%
2-5	10,616	392	4.4%
6-20	1,672	110	6.8%
21-88	83	15	18.3%

Where are these high risk residential properties in Boston? As would be suggested by the strong relationship between the Fire Risk measurement and the occurrence of fire, they are largely clustered in the areas where actual fires are clustered: Roxbury, Dorchester, East Boston, and to a lesser extent, Allston Brighton (Figure 3).

Figure 3: Residential Fires in Boston 2010 to 2015 (a) compared to (3b) Locations of High Risk properties



3a.



3b.

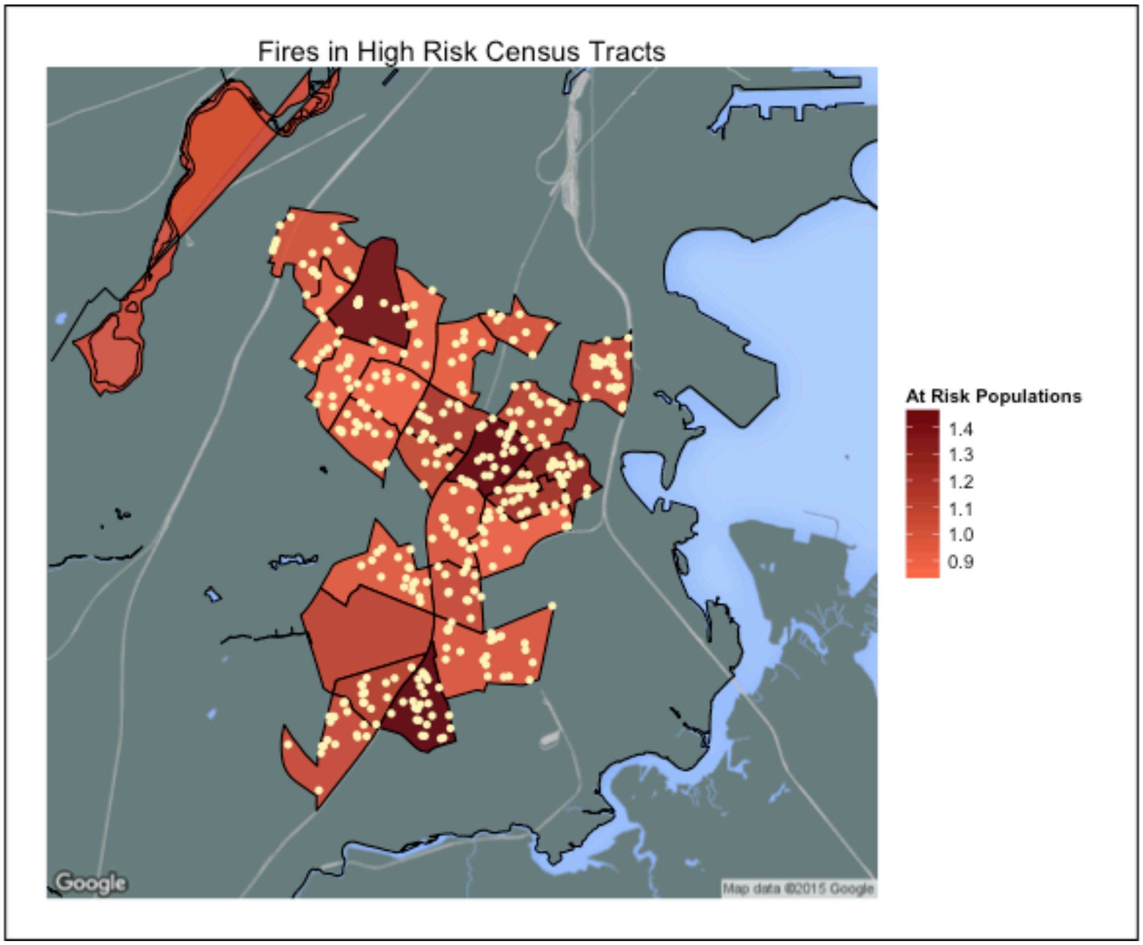
Discussion

This analysis is merely a starting point in what could develop into very meaningful research on the nature of fire risk in Boston and in cities in general. Remaining questions revolve around the relationship between human, building, and environmental factors. Are renter-occupied triple decker homes at a higher likelihood of fire due to something intrinsic in the buildings themselves or does the fire risk of those properties relate more to the behavior of the residents of those buildings? The answer likely involves some nexus of these socioeconomic, building-related, and environmental factors. However, because triple-decker homes are concentrated more heavily in poorer African American and immigrant neighborhoods in Dorchester, Roxbury and East Boston, the relationship between fire risk and this type of house needs to be tested further, possibly using data sources like field surveys that can better link between those residents who experience a house fire and the buildings they inhabit.

The relationship between 311 complaints and the incidence of fire in residential structures is perhaps the most promising finding to come out of this analysis. Such a relationship points toward the possibility of refining more highly predictive models that could be used to mark a subset of “problem properties” in the city and single them out for special inspection and monitoring to ensure fire code compliance, adequate fire escapes, and functioning smoke detectors.

The analysis clearly demonstrates how fire risk is not distributed evenly across properties in Boston. Using such analytic techniques, it would be possible to target outreach programs on fire safety to at risk populations of the elderly, poor, and families with young children who live in areas and structures with higher fire risk. Figure 4 shows one such cluster of high risk census tracts –those with a combination of high child, elderly, and poor residents—and actual fire incidents that occurred between 2010 and 2015. It may be possible in the near future to generate similar maps that show high risk "properties" in such neighborhoods and, with proper monitoring, perhaps prevent the high risk dots on the map from turning into fires on the map.

Figure 4: 2010 to 2015 Residential Fires in neighborhoods with high percentages of poor, elderly, and children in Roxbury and South Dorchester



Appendix: Code

```
# Setup
options(scipen=999)
setwd("Final Project/")

############################################################ Record Level Data ###
### Read in tax data ###
taxdata<-read.csv("Record_Level.csv", stringsAsFactors = FALSE)

### Read fire data ###
library(xlsx)
fires<-read.xlsx2('Fires_2010-1015.xlsx',1,stringsAsFactors=FALSE)
# Format and Correct some columns
fires$Date<-as.Date(as.numeric(fires$Date)-25569,origin="1970-01-01")
fires$PropLoss<-as.numeric(fires$PropLoss)
fires$Year<-as.integer(substring(as.character(fires$Date),1,4))

### Read in the 311 Data ###
crmData<-read.csv("./Main Database 2010-2015___As of May 19, 2015 no IDs.csv",
stringsAsFactors = FALSE)

# subset the taxdata so we only have residential buildings -- actual buildings (no condo units)
resBuildings<-taxdata[taxdata$LU %in% c('A','CM','R1','R2','R3','R4','RC'),]

# remove some columns to make things neater
resBuildings<-
resBuildings[c("PID","ST_NUM","ST_NAME","ST_NAME_SUF","PTYPE","LU","OWN_OC
C","YR_BUILT","YR_REMOD","LIVING_AREA","NUM_FLOORS","STRUCTURE_CLAS
S","R_BLDG_STYL","R_KITCH","R_HEAT_TYP","S_BLDG_STYL","S_UNIT_RES","S_U
NIT_COM","Blk_ID","BG_ID_10","CT_ID_10","X","Y","LocationID","BRA_PD","NSA_NA
ME")]

# pull in the csv of addresses with their locationIds that I was able to match with fire addresses
# in some instances where we found an ambiguous match between fire address and the address in
the BARI Addresses DB
# this CSV started with multiple rows per address and was manually corrected to one single
address and locationID
firesLocMatch<-read.csv("fires_locMatch.csv", stringsAsFactors = FALSE)

### Match fire data to addresses
resBuildingsWithFires<-sqldf("select distinct r.*
                    from resBuildings r
                    where r.LocationID in (select LocationID from firesLocMatch)")
```

```
# Mark a 0 or 1 for a locationid in our residential subset that had a fire
resBuildings$HadFire<-as.integer(resBuildings$LocationID %in%
resBuildingsWithFires$LocationID)

### Aggregate the crm data by locationID
crmByLocID<-sqldf("select c.LocationID,
        sum(c.PrivateNeglect) as PrivateNeglect
        from crmData c
        group by c.LocationID")

# bring in our aggregated CRM data
resBuildings<-sqldf("select r.*,
       c.PrivateNeglect
       from resBuildings r
       left outer join crmByLocID c
       on r.LocationID=c.LocationID")



# Building style can show up in R_BLDG_STYL or S_BLDG_STYL -- merge this together into
# a new
# field called just BLDG_STYL
resBuildings$BLDG_STYL<-resBuildings$R_BLDG_STYL
resBuildings$BLDG_STYL<-
ifelse(is.na(resBuildings$BLDG_STYL),resBuildings$S_BLDG_STYL,
resBuildings$BLDG_STYL)
resBuildings$BLDG_STYL<-ifelse(nchar(resBuildings$BLDG_STYL)>0,
resBuildings$BLDG_STYL, NA)

# Add a new variable for whether the propery is triple decker
resBuildings$IsTripleDecker<-as.integer(resBuildings$BLDG_STYL=="DK")

# add variable for not owner occupied
resBuildings$NotOwnerOccupied<-as.integer(resBuildings$OWN_OCC=="N")

# turn NAs to zeros for our key variables
resBuildings$PrivateNeglect<-
ifelse(is.na(resBuildings$PrivateNeglect),0,resBuildings$PrivateNeglect)
resBuildings$IsTripleDecker<-
ifelse(is.na(resBuildings$IsTripleDecker),0,resBuildings$IsTripleDecker)
resBuildings$NotOwnerOccupied<-
ifelse(is.na(resBuildings$NotOwnerOccupied),0,resBuildings$NotOwnerOccupied)

# Finally, set our fire risk measurment
```

```
resBuildings$FIRE_RISK<-
resBuildings$NotOwnerOccupied+resBuildings$PrivateNeglect+resBuildings$IsTripleDecker

# merge the fire risk into original taxdata dataframe
taxdata<-sqldf("select distinct t.*,
            r.FIRE_RISK,
            r.HadFire as HAD_FIRE
            from taxdata t
            left outer join resBuildings r on t.PID=r.PID")

############################################################### Aggregate Level Data
###
####### aggregating by Census Tract
# get subset of residential buildings
# subset the taxdata so we only have residential buildings -- actual buildings (no condo units)

# get unique fires from taxdata to avoid double counting -- there are some duplicates with PID
uniqueSet<-sqldf("select distinct CT_ID_10, PID, sum(HAD_FIRE) as HAD_FIRE
              from taxdata group by CT_ID_10, PID")

uniqueSet$HAD_FIRE<-ifelse(is.na(uniqueSet$HAD_FIRE)==FALSE &
uniqueSet$HAD_FIRE>0, 1, 0)

# now aggregate
tracts<-aggregate(HAD_FIRE~CT_ID_10, sum, na.rm=TRUE, data=uniqueSet)

# read in demographic data
tracts1=read.table('../data/Tracts/Tracts_Boston_2015_BARI_CSV_1.tab', sep="\t",
header=TRUE)
tracts2=read.table('../data/Tracts/Tracts_Boston_2015_BARI_CSV_2.tab', sep="\t",
header=TRUE)

# Pull together a fire rate by 1000 variable for census tracts
tracts_all<-merge(tracts, tracts1, by="CT_ID_10")
tracts_all<-merge(tracts_all, tracts2,
by=c("CT_ID_10","POP100","HU100","Type","Res","BRA_PD_ID","BRA_PD","City_Counc"
,"WARD","ISD_Nbhd","Police_Dis","Fire_Distr","PWD"), all=TRUE)
tracts_all$FIRES_2010_2015<-tracts_all$HAD_FIRE
tracts_all$FIRES_PER_1000<-1.0 * tracts_all$FIRES_2010_2015/(tracts_all$POP100/1000)

# rescale the FiresPer1000 so we can create a fire risk variable
library(scales)
tracts_all$FIRES_PER_1000_FIXED<-rescale(tracts_all$FIRES_PER_1000)

# High risk populations are where there are more children, elderly, poor, and where there are
more fires
```

```
# Set the fire risk at the tract level this way
tracts_all$FIRE_RISK<-
tracts_all$punder18+tracts_all$p65older+tracts_all$ppubassis+tracts_all$FIRES_PER_1000_FI
XED

# bring this into a new taxdata_agg frame
taxdata_agg<-tracts_all[c("CT_ID_10","FIRE_RISK")]
```