Non-confined Residential Structure Fire Incident Modeling:

An Overview of the Model and Instructions for Use

FEBRUARY 1, 2019

PRESENTED BY:
Evan Herring-Nathan, Statistician, Statistics and Methodology
Julie Banks, Associate Director, Statistics and Methodology
Michael Meit, Senior Fellow, Public Health
Megan Heffernan, Principal Research Analyst, Public Health
Molly Powers, Research Analyst, Public Health
NORC at the University of Chicago

NORC at the UNIVERSITY of CHICAGO

TABLE OF CONTENTS

Overview

This model estimates the number of non-confined residential structure fires in one year for each county of interest. Only counties for which model estimates are desired should be input into the model. Counties for which NFIRS data quality is believed to be high may be better estimated by scaling up NFIRS totals using the method outlined in the methodology report (see section "Scaling high-reporting counties to national totals").

Estimates from all U.S. counties and territories can then be aggregated to different jurisdictions as needed provided that they are mutually exclusive with respect to county. If the jurisdictions are not mutually exclusive the estimate from the county can be allocated proportionally to its component jurisdictions. For example, county fires can be aggregated up to Red Cross divisions relatively easily because each county belongs to one and only one division with one exception.[1] Aggregating county fires to Red Cross regions however requires slightly more complexity because many counties belong to multiple regions (e.g. Kern County, California).

The model estimates can be calculated by inputting publicly available data at the county level into the model formula. Use the information below to understand where to source this data and how to apply the model.

Background

The final estimates from the developed process are derived by following four steps:

1. Identifying counties with high-levels of fire reporting (high-quality) in the National Fire Incident Reporting System (NFIRS) and scaling up their county-level totals to be consistent with national reported totals.[2]
2. Using these high-quality counties to build a generalized linear model and a simple linear regression model from which to predict fire incidence for the remaining counties (with assumed lower levels of reporting, low-quality) and territories
3. Applying the generalized linear model to the low-quality counties
4. Applying the simple linear regression model for U.S. territories.

For optimal results, this same process should be used when developing estimates in the future.
The NFIRS data were instrumental in developing this model. While NFIRS data are very useful in understanding fire incidence across the nation, they do not perfectly and comprehensively represent all fires. It is understood that fire departments have variable reporting rates and accuracy regarding the data elements that are reported. Our data quality assessment included custom methods to determine counties where under-reporting was most severe. There were also counties that did not report any fires at all. In most cases, especially in the counties with higher residential population, it is believed that the lack of data in NFIRS is due to non-reporting rather than a true 0% fire incidence rate.

The model was built using only the ~2,200 counties for which fire incidence was believed to be reasonably and consistently reported. While these counties are believed to have good reporting they still suffer from some degree of under-reporting. Their degree of under-reporting is not near the levels seen

[1] Christian County, Kentucky (FIPS 21047) has one zip that belongs to the Southeast and Caribbean division while the other zip codes belong to the Crossroads division.

[2] Methods used to assess reporting quality for this study are discussed in the methodology report.

with the counties with no or inconsistent reporting; however, it is still necessary to scale up their totals to be in line with estimates based on national rates. This scaling was performed using national estimates of fire incidence published by the National Fire Protection Association (NFPA) as well as known population totals for the counties and the country as a whole. For more details refer to the methodology report.
The 2,200 high-quality counties were used to build the model, first by scaling their fire incidents to be in line with expected levels. Appropriate statistical methods were used to build and validate the model using these counties. Then the resulting model was applied to the remaining ~900 counties for which NFIRS data were not accurate so that estimates could be realized. U.S. territories were then also modeled using a simple population model because other data elements required were not available, or in the case of Puerto Rico, resulting estimates were believed to be more reliable using this model.

Data preparation

To apply the formula several data elements are necessary. First, NFIRS reported totals of non-confined residential structure fires for each county are required. These counties are then split out into two groups; counties with high-reporting rates, and counties with low-reporting or no reporting. High-reporting county estimates are scaled up to be consistent with national levels. Counties with low or no reporting will be modeled later using a generalized linear model, or in the case of territories, with a simple linear regression model.

Predictive data elements are then collected for all counties, as are population totals for U.S. territories. Several data manipulation steps are required to make the data ready for use in the model code. For instructions on sourcing and deriving required data elements please refer to the “Model formula” section below.

When all necessary data has been collected the model code can be applied to counties for which model estimates are desired. Typically this would be counties for which reporting is low or non-existent as well as U.S. territories.

Model formula

The following formula is used to create county-level fire estimates for the 50 states and the District of Columbia. (Note that coefficients have been rounded to the nearest thousandth below, but all significant digits are included in the SAS code.)

fire_estimate = exp(-**8.795+** hu_ln + (**0.126** * region_2) + (**0.152** * region_3) + (-**0.567** * region_4) + (-**0.014** * months_bel_60max_imp) +(-**0.039** * popdensity_ln) + (**3.548** * perc_disability_ambulatory) + (**1.367** * perc_snap) +(**1.472** * perc_occ) + (**1.004** * perc_black) + (**3.329** * smoke_perc) + (**0.085** * region_4 * months_bel_60max_imp) + (**0.015** * ruca_wgt_cat))

Explanation of predictors (data dictionary)

The variables used in the model are described in Table 1.

Table 1. Full model predictors and descriptions

Variable name	Description	Source	Special Notes
fire_estimate	Estimated number of fires in the county (annual)	N/A (This is the outcome variable that the model creates. Once all necessary data is input into the model, the fire_estimate is created.)	
hu_ln	Natural log of the number of housing units in the county	American Community Survey (5-year)	Download the number of housing units from Census and then take the natural log
region_2	An indicator for belonging to U.S. Census region 2, Midwest (1 if yes, 2 if no)	Census Region Data (see attached SAS datasets)	See derivation process in "Derived Variables" section
region_3	An indicator for belonging to U.S. Census region 3, South (1 if yes, 2 if no)	Census Region Data (see attached SAS datasets)	See derivation process in "Derived Variables" section
region_4	An indicator for belonging to U.S. Census region 4, West (1 if yes, 2 if no)	Census Region Data (see attached SAS datasets)	See derivation process in "Derived Variables" section
months_bel_60max_imp	Number of months for which the monthly average maximum daily temperature was below 60° F	CDC Wonder - North America Land Data Assimilation System Daily Air Temperatures and Heat Index (year 2011 only)	See derivation process in "Derived Variables" section
popdensity_ln	Natural log of the population density for the county	American Community Survey (5-year)	Download the population density from Census and then take the natural log
perc_disability_ambulatory	Percentage of the county's civilian non-institutionalized population with an ambulatory difficulty	American Community Survey (5-year)	
perc_occ	Percentage of the county's housing units that are occupied	American Community Survey (5-year)	
perc_black	Percentage of the county's population who are African-American	American Community Survey (5-year)	
smoke_perc	Percentage of adults in the county population who are regular smokers	County Health Rankings & Roadmaps (Robert Wood Johnson Foundation)	
region_4*months_bel_60max_imp	The interaction of the variables indicated	NA	This is simply the interaction effect of the previously mentioned variables
ruca_wgt_cat	A derived variable describing the county's ruralness	United States Department of Agriculture Rural Urban Commuting Area (RUCA) codes	See derivation process in "Derived Variables" section
perc_snap	Percentage of households receiving SNAP benefits	American Community Survey (5-year) SNAP statistic	

Evaluation of predictors

All variables in the model, with the exception of the standalone *months_bel_60max_imp* variable*,* were statistically significant (α = .05)[3]. A variable can be positively or negatively correlated with the number of fires. A positive correlation means that higher values of the variable correspond to higher values in the number of fires when holding other predictors constant. Similarly, a negative correlation means lower values in the variable correspond to lower numbers of fires. Variables may also interact with each other whereby the effect of one variable on the outcome depends on the value(s) of one or more different variables. The associative relationships discovered in the model are shown in Table 2.

Table 2. Full model predictors and associative relationships

Variable Name	Association	Interpretation (holding all other variables constant)
hu_ln	Positive	Counties with more housing units experience more fires than counties with fewer housing units
region_2	Positive	Counties in Census region 2 experience more fires than those in region 1
region_3	Positive	Counties in Census region 3 experience more fires than those in region 1
region_4	Qualified by interaction term	Qualified by interaction term
months_bel_60max_imp	Qualified by interaction term	Qualified by interaction term
popdensity_ln	Negative	Counties with higher population density experience fewer fires than other counties
perc_disability_ambulatory	Positive	Counties with a higher percentage of their population with an ambulatory difficulty experience more fires than other counties
perc_occ	Positive	Counties with a higher percentage of their housing units occupied experience more fires than other counties
perc_black	Positive	Counties with a higher African-American population percentage experience more fires than other counties
smoke_perc	Positive	Counties with higher rates of regular smoking experience more fires than other counties
region_4*months_bel_60max_imp	Positive (qualified by component standalone coefficients)	Counties in region 4 with colder weather patterns (more months with average daily maximum temperatures below 60 F) experience more fires than other counties
ruca_wgt_cat	Positive	Counties that are considered more rural experience more fires than other counties
perc_snap	Positive	Counties with a higher percentage of their households receiving SNAP benefits experience more fires than other counties

Estimates for U.S. territories

[3] Because *months_bel_60max_imp* is significant at the .05 level when used in an interaction term with *region_4* the inclusion of this variable is justified.

None of the predictive data is available for U.S. territories save for Puerto Rico. Additionally, the NFIRS public data release does not include reported fires for these territories. So it was necessary to develop a separate territory model that would provide reasonable estimates despite the lack of available information. Because of the high correlation of population and fire incidence a linear regression model using population as the sole predictor was considered. The high predictive power of this model justifies its use on the territories and it can be applied with the following formula for the territories (American Samoa, Guam, Northern Mariana Islands, Puerto Rico, and U.S. Virgin Islands). Table 3 describes the *acspop* variable which is the only variable used in the territory population model.

Fire_estimate=**14.701 + 0.000455** * acspop

Table 3. Population model predictors and descriptions

Variable name	Description	Source	Special Notes
acspop	Population of the territory (or county for Puerto Rico)	For Puerto Rico: American Community Survey (5-year) For others: United Nations data or other reputable source	For Puerto Rico the model creates an estimate for each county separately so county-level population is required. All other territories have one population statistics and accordingly one fire estimate.

However, because there was no NFIRS data for the territories it was not possible to validate this model. Therefore, some degree of caution is advised when using estimates derived from it. For future research, special emphasis on deriving a more sophisticated model for the territories could be explored if appropriate data can be found. In the current study, some effort was put into researching more complex ways to estimate the territories but ultimately the necessary data to do this could not be found.

Derived Variables

Several data elements were transformed during the modeling process in order to better utilize the information they contained. As a result there is some data preparation that must be performed prior to running the model program code.[4]

1. *Census Regions.* Three indicator variables are necessary for each county in order for the model to determine which Census Region the county belongs to. The SAS data set provided ("reg.sas7bdat") contains these identifiers according to the latest Census definitions. It is read in by the model code and then appended to the dataset for which estimates are desired. All counties should have either a 0 or 1 for each of these three variables after the "reg.sas7bdat" file has been merged onto the dataset. Appending these variables requires that the target dataset contains the 2-digit state_fips variable in numeric format. If Census region definitions have changed this dataset should be updated.

2. *Climate Data.* The "months_bel_60max_imp" variable is derived monthly climate data for each county. To compute this variable it is necessary to have the average maximum temperature for each month for the most current year. The variable then equals the number of months that had a

[4] All derived variables are included in the datasets provided. However, the code to read and compute these variables is not included due to potential errors if and when data source formats, coding, and content changes. Written explanations of the process to derive these variables are therefore sufficient to enable the user to derive them.

maximum temperature below 60° F. The range of possible values is therefore between 0 and 12 and all counties should have a value. If imputation is required the user can consider neighboring counties or other data sources for possible values. The SAS data set provided ("climate.sas7bdat") contains this variable according to the most recent climate data available. It is read in by the model code and then appended to the dataset for which estimates are desired. If climate patterns have changed this dataset should be updated.

3. *Housing Units and Population Density.* These two variables can be downloaded from Census online products. However, for use in the model the natural log must be applied.

   Hu_ln=log(housing_units)

   Popdensity_ln=log(population_density)

4. *RUCA Codes.* To calculate the "ruca_wgt_cat" it is necessary to first download the tract-level RUCA codes from the USDA as well as tract-level population codes from the Census Bureau. Then for each county a weighted average of the primary RUCA code, weighted by tract population, is computed and the result is rounded to the nearest integer. The SAS data set provided ("ruca.sas7bdat") contains this variable according to the most recent data available. It is read in by the model code and then appended to the dataset for which estimates are desired. If commuting patterns have changed this dataset should be updated. After this data set has been merged in all counties should have a value between 1 and 10.

   As an example, the ruca_wgt_cat is computed in the following fashion for county "i" using tract-level information for each tract "j" within the county:
   $\text{ruca\_wgt\_cat}_i=\sum_j(\text{tract\_population}_j*\text{primary\_RUCA\_code}_j)/ \sum_j (\text{tract\_population}_j)$

NFIRS Scaled Estimates

Counties for which data quality is determined to be high will be best estimated with a direct approach using scaled totals from the NFIRS data rather than the full model estimates. Two variables, *exclude* and *nfirs_scaled_est*, are used to populate NFIRS scaled estimates in place of full model estimates. These variables are described in Table 4, are included in the *for_prediciton* template dataset, and referenced in the provided SAS code. Counties for which data quality is low will be estimated with the full model described in the preceding sections.

Table 4. Additional variables for high-quality counties and scaled NFIRS estimation method

Variable name	Description	Source

exclude	Describes the quality of the county's NFIRS data. 0=high-quality, use NFIRS scaled estimate 1=low-quality, use full model estimate missing=no fires reported in NFIRS, use full model estimate	Derived from NFIRS based on data quality assessment This variable is derived and is based on data quality evaluation metrics. Refer to the methodology report for steps undertaken to derive this variable. Alternative data quality assessment methods may also be considered at the time a new NFIRS dataset is evaluated
nfirs_scaled_est	For high-quality counties, this is the number of annual fires estimated by using the NFIRS scaling technique (described in the methodology report)	Derived but informed by NFIRS county-level data and other data sources (see methodology report)

Using the model estimates

We have provided estimates for the total number of estimated fires for 3 different geographic areas of interest (U.S. counties, Red Cross regions, and Red Cross divisions). These estimates reflect the total number of non-confined residential structure fires (using standardized terminology from NFIRS). The estimates can be used as inputs into the decision making process for determining data-driven targets for Red Cross response. However, the estimates are totals and should not necessarily be treated as the Red Cross targets themselves.

How to apply the model for future use

Using the current estimates

If the relationships between the predictive variables and the number of fires remain relatively constant, the model predictions could be carried forward as is. However, changes in any number of factors (e.g. changes in county composition especially in regards to the predictive measures; shifting demographics, housing characteristics and climate patterns; increased fire department resources and efficiency; improved preventative initiatives and fire education) could render the model predictions out-of-date and therefore less than ideal. Therefore, it is recommended that continued maintenance be performed on this model.

SAS code

The required SAS code and datasets to generate estimates for counties and territories have been provided. The user is advised to read this report before attempting to run the code. Once sufficient background knowledge is acquired the user may run the "fire_estimate_code.sas" code. The user is advised to run the code step-by-step as only some parts of the process are able to be easily automated. Generally, the user is advised to follow the example provided, leaving macro variable names unchanged, and structuring their datasets in a way consistent with the code. Comments throughout the code guide the user in the correct application of the model. The complete listing of provided code and datasets is as follows:

- *fire_estimate_code.sas* The main code that starts the estimation process. This is where the user should start.

- *climate.sas7bdat*: Input dataset containing the climate variables for each county
- *ruca.sas7bdat:* Input dataset containing the RUCA variables for each county
- *reg.sas7bdat:* Input dataset containing the Census region indicator variables for each county
- *county_formula.sas:* The code that estimates fire incidence for each county of interest
- *rc_reg.sas7bdat:* Input dataset containing the Red Cross region for each county
- *rc_div.sas7bdat:* Input dataset containing the Red Cross region for each county
- *for_prediction.sas7bdat:* Input dataset containing all counties and territories for which estimates are desired. This particular version of the dataset is for demonstration purposes only and is provided to show the user the correct structuring of the data.

Imputation

In some cases imputation may be necessary for counties with missing data elements. In the current effort it was only necessary to impute climate data for Alaska and 9 other counties due to either changes in county FIPS codes or unavailability of data. In these cases, neighboring counties were used to infer plausible values. If the number of missing items is very high then consideration should be given to the cause of the missingness. In some cases inconsistencies in county FIPS codes across data sources can introduce missingness. If imputation is required the user can consider using values from similar counties (e.g. neighboring counties in the case of climate data). Similarity can also be defined in different ways (e.g. statistical matching) so that the imputation results in a reasonable value.

Rebuilding and refreshing the model

Continued evaluation of existing data sources and model performance is strongly recommended. Specifically, NORC advises the annual review and rebuilding of the model to ensure that the most recent data is being used and so that model estimates are as accurate as possible. There are several reasons why this is necessary.

Associative relationships in the current model may not hold in the future due to changes in fire risk and response characteristics across the county. Policy and educational initiatives may have implications for the way in which risk factors influence fire incidence. It is also possible that new risk factors become more relevant as these changes influence the operating environment. The availability of new data sources that are particularly predictive of fire incidence, or advances in the understanding of modeling fire incidence, could also warrant a rebuilding of the model.

Furthermore, shifting demographics and county characteristics may make current county-level estimates obsolete. Even though changes in these characteristics and associative relationships may be small or incremental, it is advised that this regular maintenance step is built into the fire estimation process. Ideally, the model would be constructed anew every year as new NFIRS and predictive data become available. The data exploration process should be performed in light of existing knowledge but also with an eye toward new and powerful data sources and variables.

Alternatively, although less preferable, if resources are limited a model refresh can be performed by collecting the most current data sources then running the current model formula to produce new estimates.

Refer to the “Explanation of predictors” section above to understand what data to collect to perform the refresh. Also refer to the “Model formula” section to understand how to apply the formula to these new data.

Predictive power of the model

Multiple methods were used to evaluate the predictive power of the final model. Special emphasis was placed on the comparison of the final model (also referred to as the “full model”), the “housing unit model” (which uses the housing unit variable as its only predictor), and a “simple model” which uses the same estimate for each county (the average number of scaled fires which is about 60). Model fit statistics such as the AIC and BIC all suggested that the models in order of highest predictive power are as follows: the full model, the housing unit, and the simple model. Therefore, the full model was selected.

The full model was evaluated in terms of its accuracy for both the dataset used to construct it as well as holdout datasets in order to ensure generalizability to counties not used to build the model. Residuals were used to determine the degree to which the model accurately predicted the NFIRS scaled estimates. Cross-validation was also performed to determine the reliability of the model and its predictive power for all counties. Finally, estimates at the region- and division-level were explored to evaluate consistency with Red Cross 2018 fire incidence.

Predictions to NFIRS scaled values (scatterplots for build and validation, housing unit model, intercept only model)

Generally, model estimates for the data used to build the model, or “building data”, follow the scaled NFIRS values well (see Figure 1). There is high correlation between the estimates and the scaled NFIRS values, and residuals hug the green line (which represents perfect prediction). There is some heteroskedasticity noted for high levels of actual fires but model fit appears to be acceptable.

Figure 1. Scatterplot of full model predictions and NFIRS scaled estimates (building dataset)

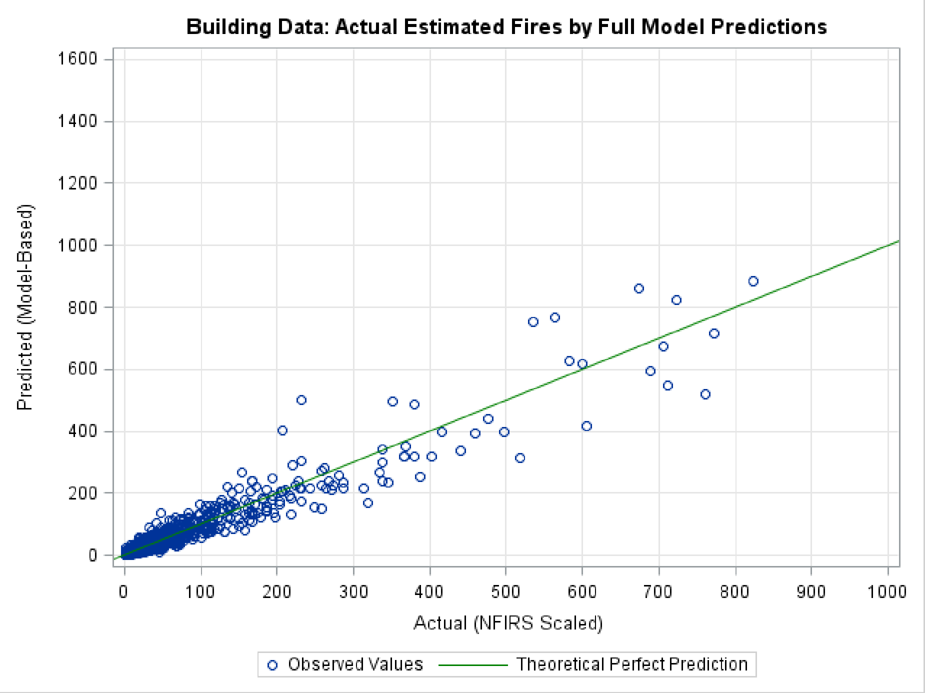
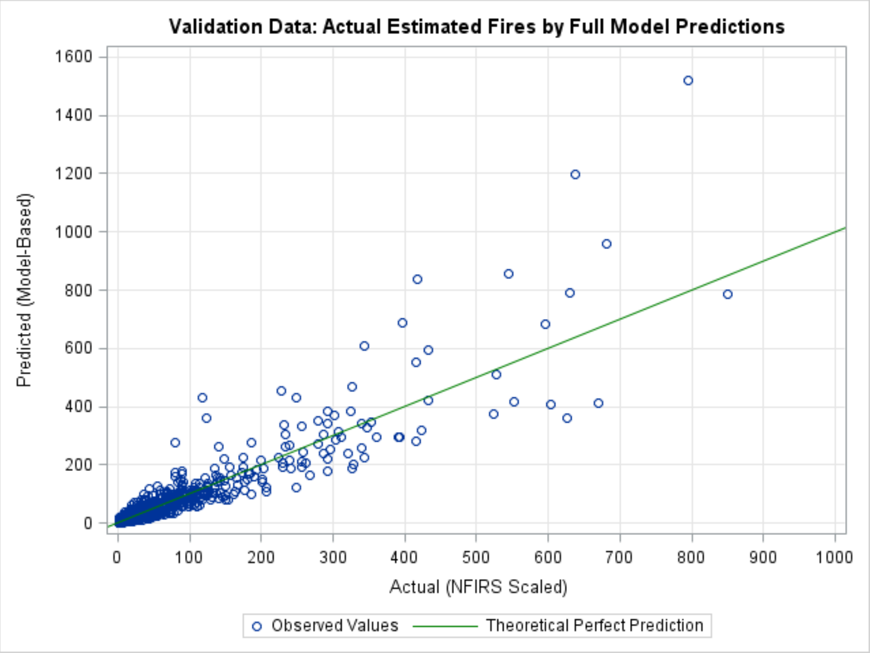


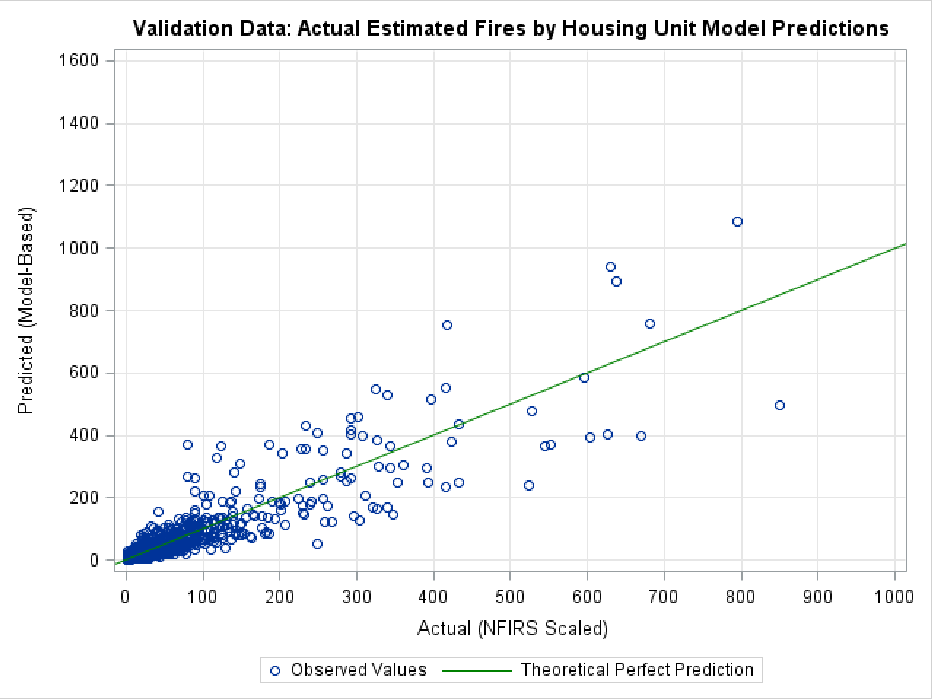
Figure 2 shows that model fit for the validation data follows a similar pattern. It appears that there are several counties for which the fit is not as good as it was for the building data, but generally the predictions are acceptable and in line with what was seen on the building dataset.

Figure 2. Scatterplot of full model predictions and NFIRS scaled estimates (validation dataset)



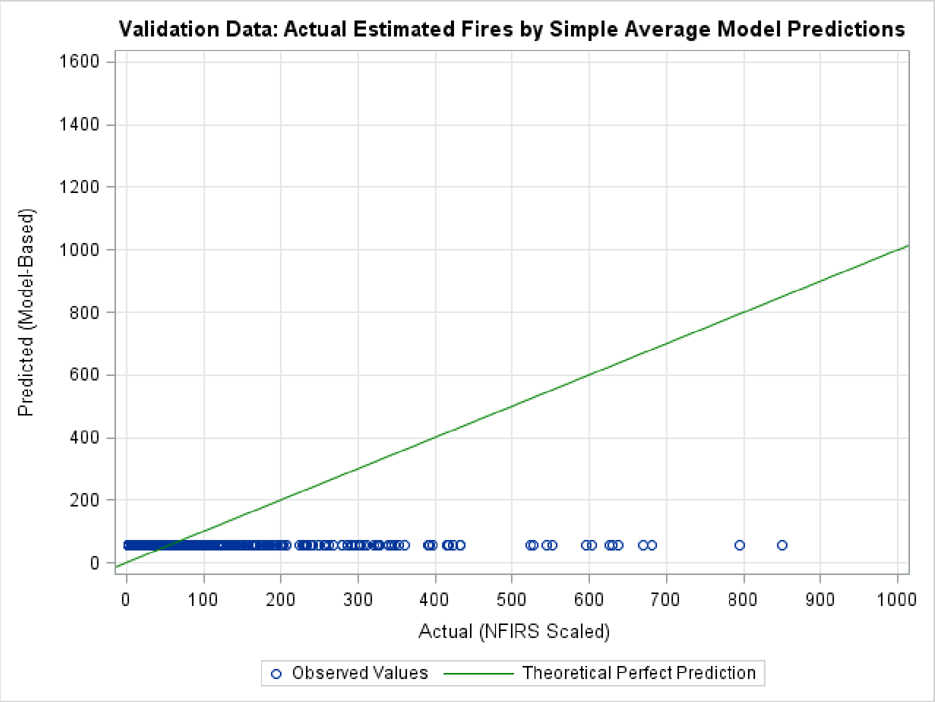
By comparison, a model consisting of housing units as the only predictor variable demonstrates the high predictive power of this variable (see Figure 3). However, the fit is not as good as that of the full model. Further evidence of this can be seen with a closer look at the distribution of the residuals for both models.

Figure 3. Scatterplot of housing unit model predictions and NFIRS scaled estimates



A model using the national average of 60 fires per county clearly demonstrates that a more complex model is necessary to accurately predict fire occurrence (see Figure 4).

Figure 4. Scatterplot of simple model predictions and NFIRS scaled estimates



Residual density (build/validate, 3 models)

In Figure 5, we see that the distributions of the full model residuals for the building and validation datasets are extremely similar. The slightly higher peak near 0 for the building set is not surprising and the similarity is a demonstration of the viability of the model. One or two counties in the validations set with more extreme negative residuals cause a long tail in the density plot.

Figure 5. Density curve of full model residuals (building and validation datasets)

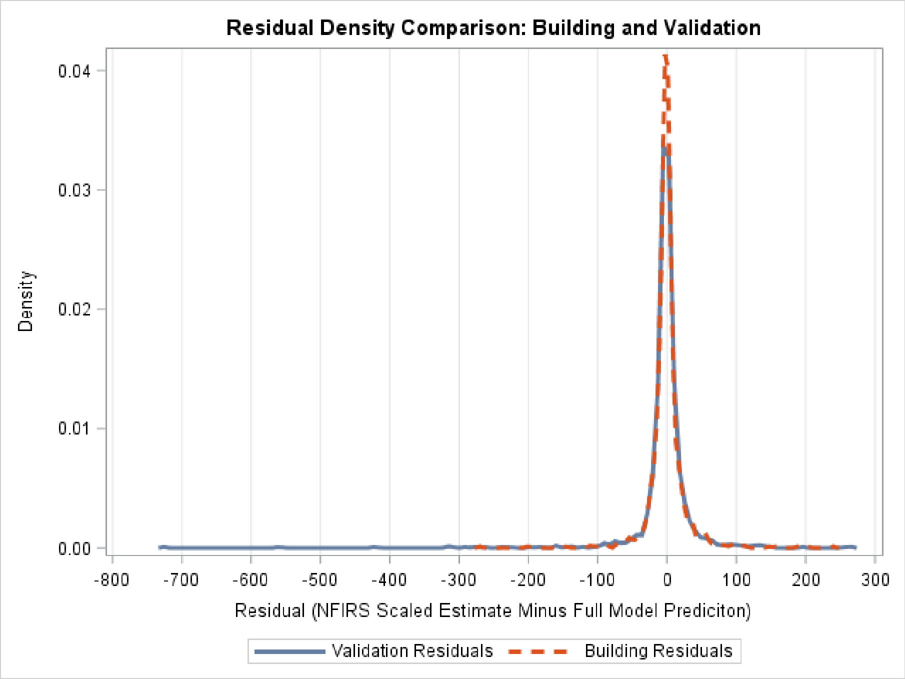


Figure 6 shows that the residual distributions for the full and housing unit models are similar. The zoomed in view in Figure 7 removes the obfuscation caused by the large scale, and helps to illustrate the differences between the two models. The housing unit model over-predicts fire incidence more often than the full model as indicated by higher density in the negative numbers. The shape of the density curve for the full model also shows that there is more concentration near 0 for that model indicating better fit.

Figure 6. Density curve of full model and housing unit model residuals

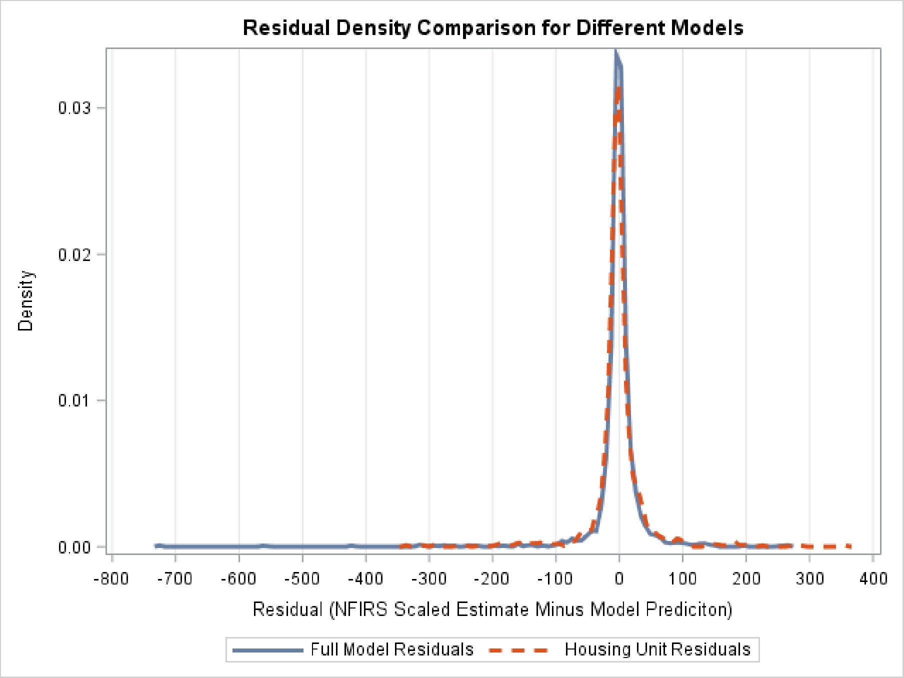
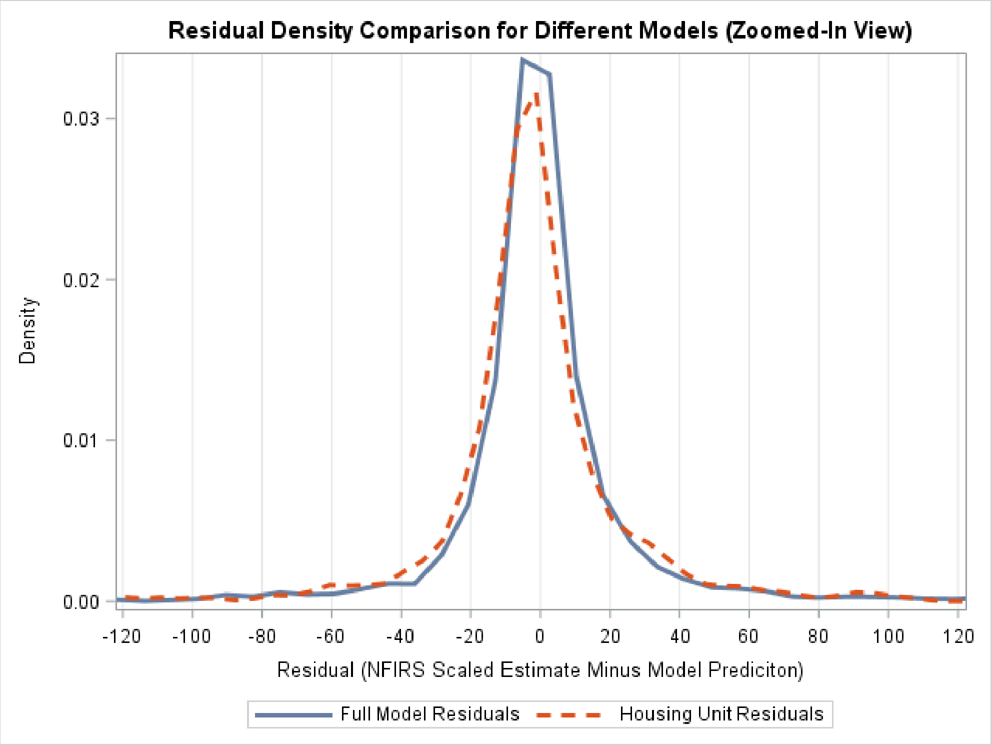
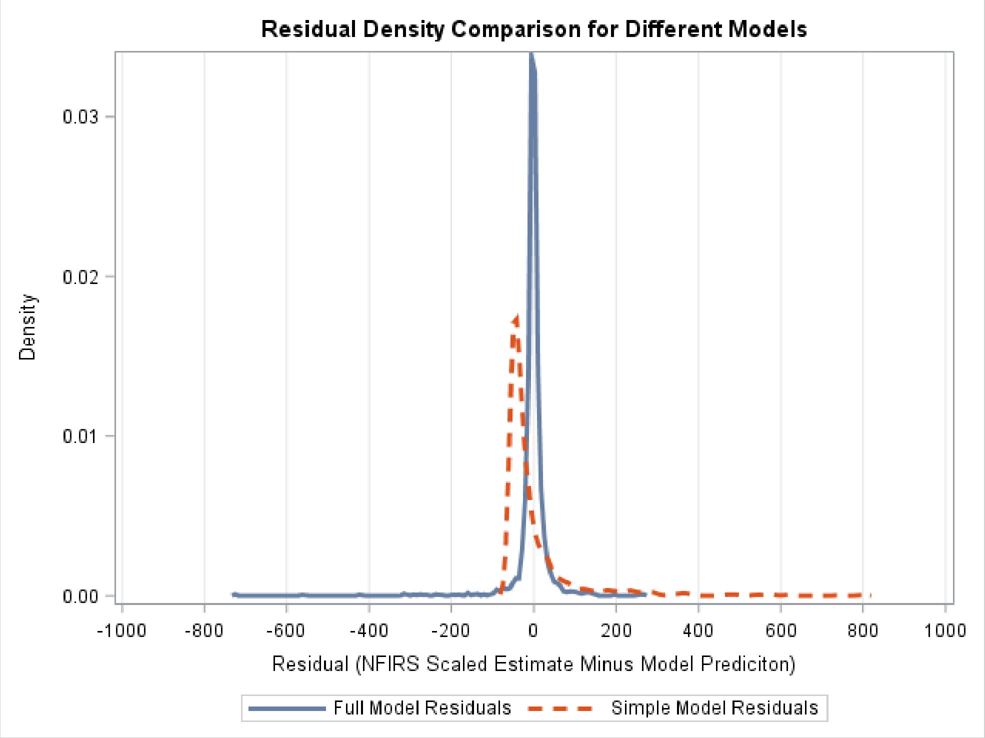


Figure 7. Density curve of full model and housing unit model residuals (zoomed-in)



For context, it is interesting to compare the residuals for the full and simple models (see Figure 8). The full model is clearly preferable as it has a peak near 0 and residuals are spread evenly across the range.

Figure 8. Density curve of full model and simple model residuals



Residual quantiles (build/validate, 3 models)

A tabular analysis (see Figure 9) highlights the similarities between the full model residuals for the building and validation data sets. The similarity between them, especially between the 10th and 95th percentile illustrate the viability of the model. It should be noted that one county in the validation data set causes the minimum residual to be very extreme (-726) and it's possible that to some degree the model may over-predict counties with certain characteristics. It's possible that this slight tendency for over-prediction for some counties could be explained by under-reporting in NFIRS rather than model deficiency although the true cause is not known at this time.

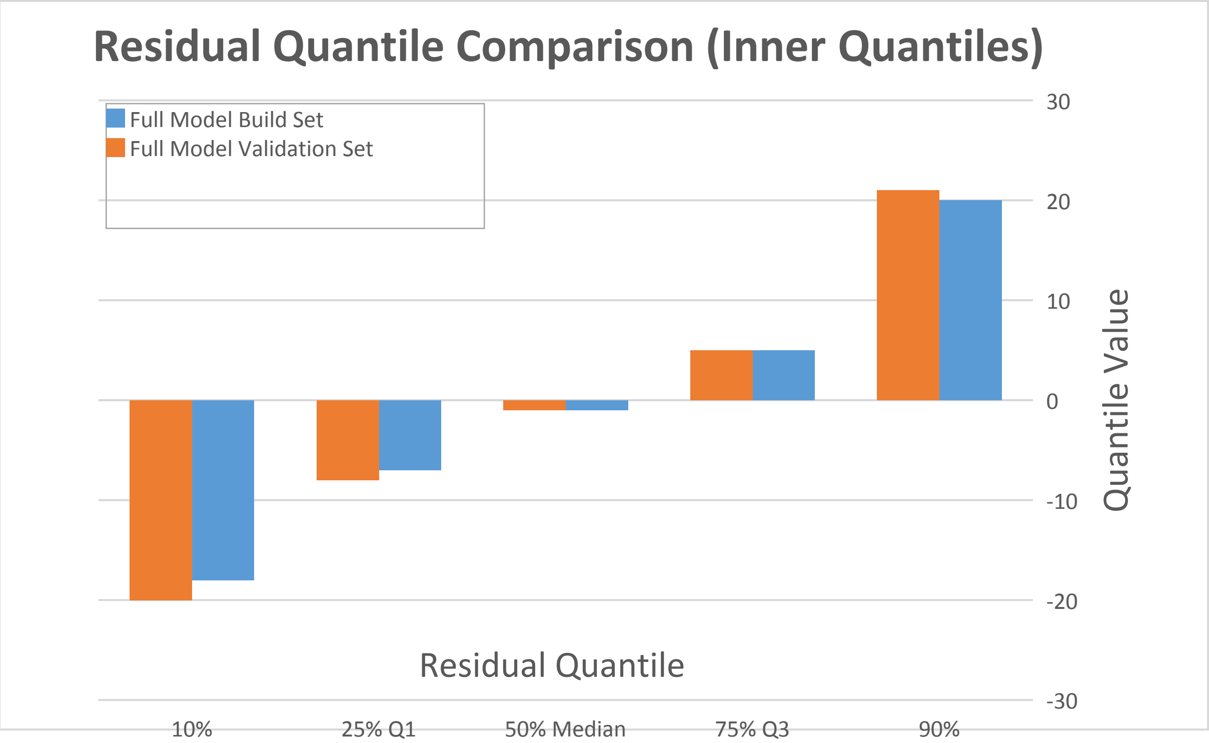
Figure 9. Quantiles of full model residuals (building and validation datasets)

Quantile	Full Model Build Set	Full Model Validation Set

100% Max	240	266
99%	98	112
95%	35	37
90%	20	21
75% Q3	5	5
50% Median	(1)	(1)
25% Q1	(7)	(8)
10%	(18)	(20)
5%	(29)	(39)
1%	(72)	(196)
0% Min	(270)	(726)

In Figure 10 the similarities are made even more apparent with the grouped histograms.

Figure 10. Grouped histogram of full model residual quantiles (building and validation datasets)

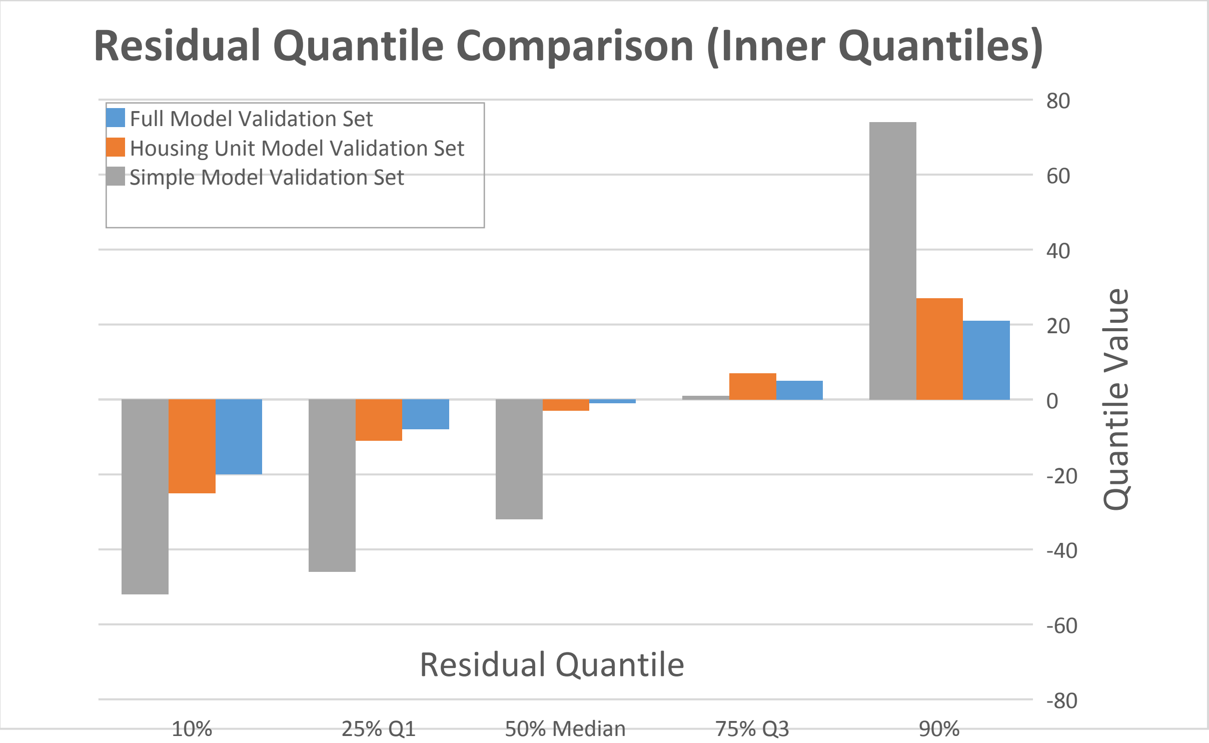


The 3 models' residuals are compared in Figures 11 and 12. Residuals are most extreme for the simple model. The housing unit model provides significant improvement. And finally the full model provides the most accurate predictions. The simple model is not capable of producing extreme negative residuals because the most extreme value taken by Y- is -60.

Figure 11. Quantiles of full, housing unit, and simple model residuals

Quantile	Full Model Validation Set	Housing Unit Model Validation Set	Simple Model Validation Set
100% Max	266	356	793
99%	112	179	487
95%	37	49	192
90%	21	27	74
75% Q3	5	7	1
50% Median	(1)	(3)	(32)
25% Q1	(8)	(11)	(46)
10%	(20)	(25)	(52)
5%	(39)	(49)	(55)
1%	(196)	(187)	(55)
0% Min	(726)	(336)	(55)

Figure 12. Grouped histogram of full, housing unit, and simple model residual quantiles



Cross-validation

Cross-validation was performed by first splitting the high-reporting counties randomly into 3 groups 100 times. In each iteration, 50% of the counties were allocated to a building dataset (used for constructing the model), 25% were allocated to the validation dataset (used to test model performance), and the final 25%

were allocated to the testing dataset (also used to test model performance). Then, the model was fit on the building dataset, and statistics describing the model's fit were output for all three datasets. Results from these 100 iterations showed that significance for the chosen predictors remained consistent across the datasets. Figure 13 shows that average squared error and average absolute error are very similar across the datasets as well. This is suggestive of the reliability of the model for all counties.

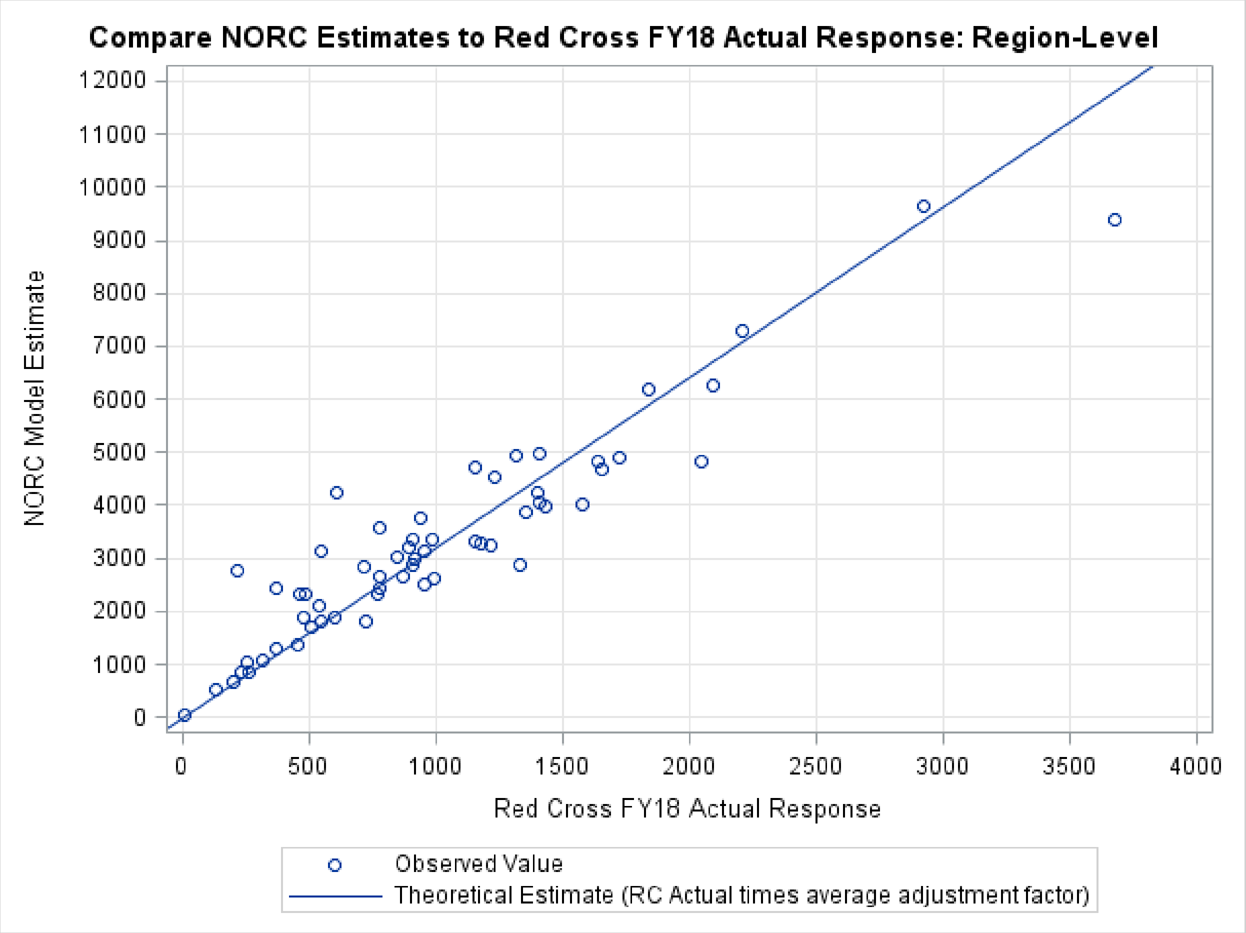
Figure 13. Cross-validation average error results

Dataset	Average Squared Error Over 100 Cross Validations	Average Absolute Error Over 100 Cross Validations
Building	1,432	16
Validation	1,537	17
Testing	1,482	16

Model compared to Red Cross response and target data for regions and divisions

Red Cross 2018 regional actual fire response correlates very well with the model estimates (see Figure 14). The data points generally follow the blue line which represents the number of fires Red Cross responded to times the overall scaling factor of ~3.2. The overall scaling factor corresponds to the ratio of the NFPA national fire estimate (190,300 for the 50 states and the District of Columbia, excluding U.S. territories) to the Red Cross 2018 national response number (59,283). While Red Cross would not necessarily respond to all of these 190,300 fires, ideally response would be provided to a large percentage of them.

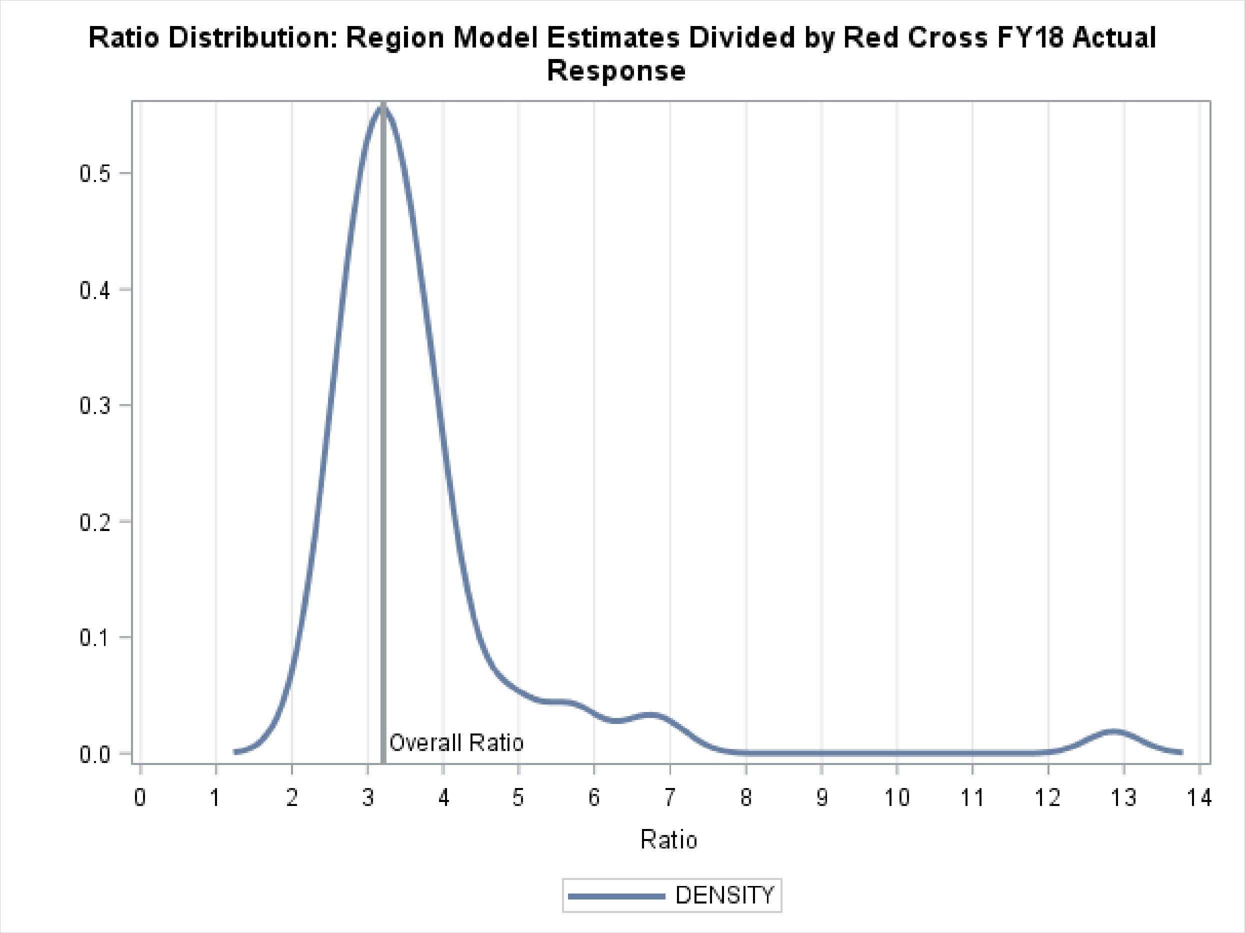
Figure 14. Scatterplot of region estimates and Red Cross FY18 response data



In Figure 15 we see that the distribution of the ratios for the regions (the estimated region totals divided by the Red Cross 2018 response totals) is concentrated in the 2-to-5 interval. Optimal increase in Red Cross response for these regions would be proportional to the ratio, and only regions above 5 would expect to see the most drastic increases in response numbers. Note that the highest ratio observed was for the Puerto Rico region (12.9) and the next highest was for the Massachusetts region (7). The lowest ratio observed was for the Oklahoma region (2.1)[5]. The high ratios observed for Puerto Rico and the Virgin Islands may be indicative of both difficulty in estimating fire incidence for territories as well as sub-optimal Red Cross response.

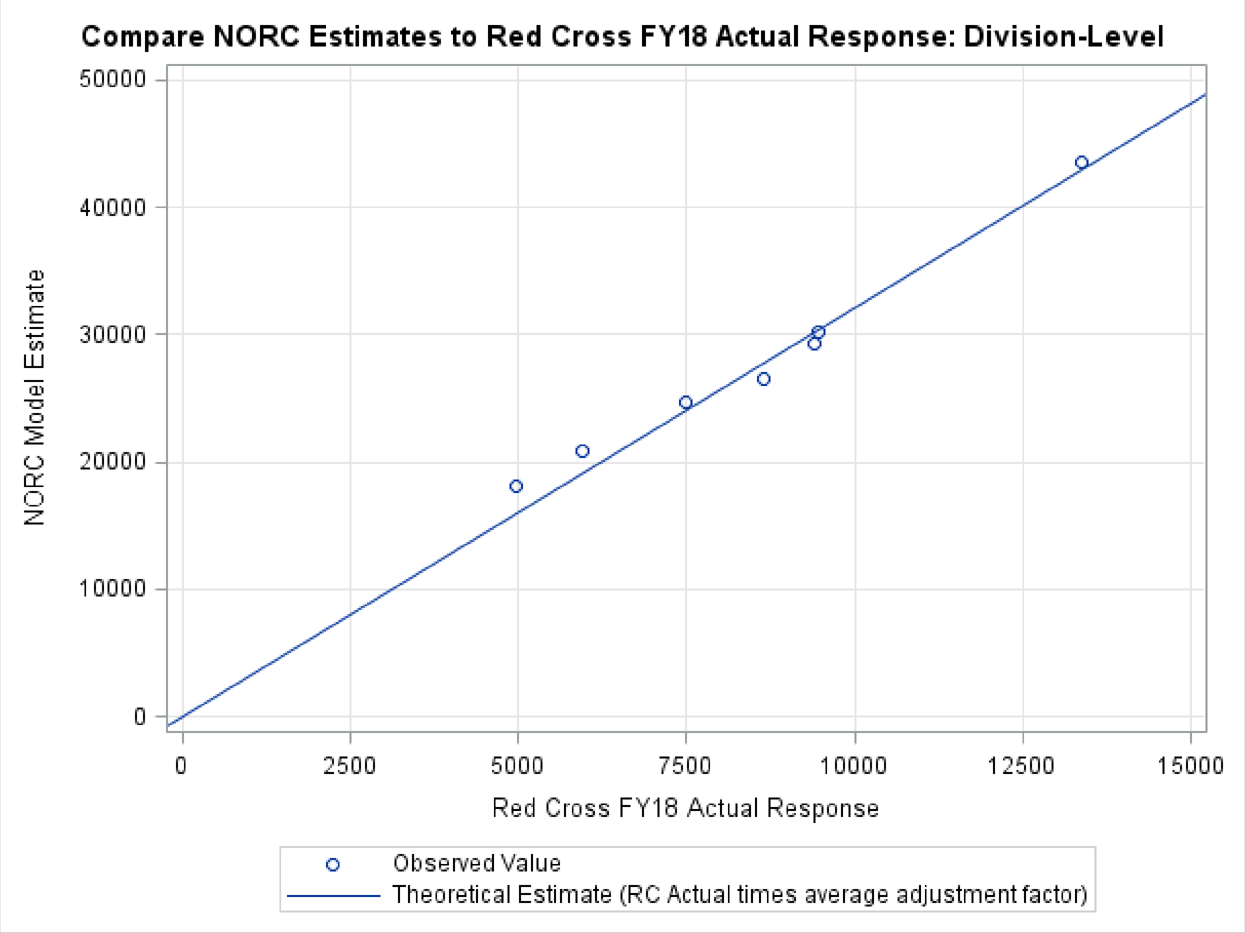
Figure 15. Density curve of region estimates and Red Cross FY18 response ratios

[5] The density curve shown in Figure 15 averages and smoothes results according to a mathematical model in order to better visualize the trend of these data. While the lowest ratio observed was 2.1 (Oklahoma region) the density curve appears to suggest density below 2, when in fact no region had a ratio below 2. A similar observation is noted in the upper tail of the density curve as well. This is not an unexpected result.



The full set of statistics for the 59 regions can be found in the methodology report. The division estimates shown in Figure 16 also show high correlation with actual Red Cross 2018 response data. No one division is singled out as being especially problematic, but rather all divisions could potentially benefit from increased Red Cross response.

Figure 16. Scatterplot of division estimates and Red Cross FY18 response data



The full set of statistics for the 7 divisions can also be found in the methodology report.

Conclusion

Having a data-driven understanding of fire incidence in the United States and its territories is an important goal for the Red Cross. The challenges resulting from a lack of accurate and comprehensive databases of fire incidents make this goal more difficult to realize. Advanced statistical and data analysis methods, along with review of the existing literature and discussion with subject matter experts have all been utilized to manage this knowledge gap. These methods have enabled the realization of estimates at the county-level, and importantly, for the Red Cross regions and divisions.

There was a high degree of consistency between what was discussed with fire experts and in the literature and the results from the model. The model evaluation shows that the final model is reliable and powerful. The estimates derived from the model will help Red Cross in developing data-driven targets to effectively deploy available resources. Continued maintenance of the model is highly recommended to facilitate the most accurate understanding of fire incidence into the future.