Memorandum

To: Rudy Uribe

From: Colin Barrett

Date: November 18, 2019

Re: Thomas Fire Study

The purpose of this memo is to describe the findings from an analysis of the Thomas Fire Data and the statistical methods used to find them.

I hope that the information provided within helps answer your research question:

> *"Which characteristics of home construction, landscaping, and terrain might be associated with a higher chance of burning in a wildfire like the Thomas Fire?"*

This memo is organized into five sections:

I. Abstract of Key Findings

Data was gathered on several housing features and several terrain types that were speculated to have a potential impact on burn rates. Of these features that were investigated, fencing (Figure 1) was the one feature that was found to have a significant association with burn rate. For all other housing features and terrain types, we do not have evidence of a significant difference in burn rates associated with that feature or terrain type at this sample size (eg. Landscape Vegetation, shown in Figure 2).

For fencing, the factor we found to be significantly associated with burn rate, 58.21% of all homes with combustible fences burned, while only 38.21% of houses with non-combustible fences burned. In other words, based on our sample we found that having a combustible fence is associated with 2.22 times higher odds of burning in this fire (222% increase) relative to having a non-combustible fence. This difference was found to be significant at a 0.0002 significance level.

All results are discussed in more detail in the Results section further on in the memo (page 7).

Figure 1. Distribution of Burn Rate by Fence Type (Significant, $p < 0.01$)

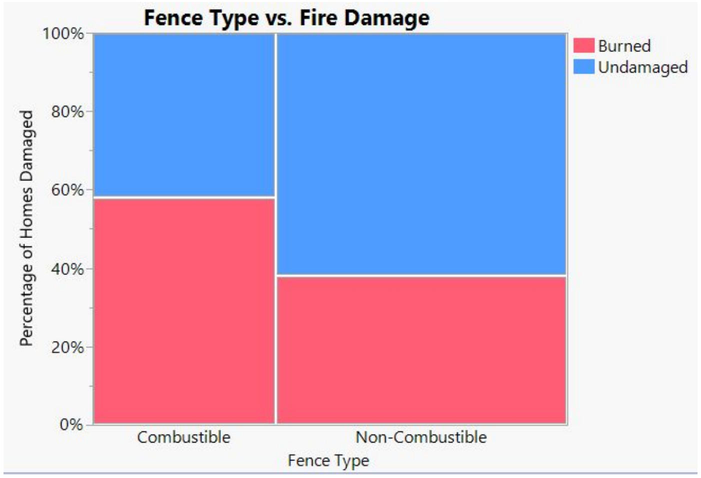
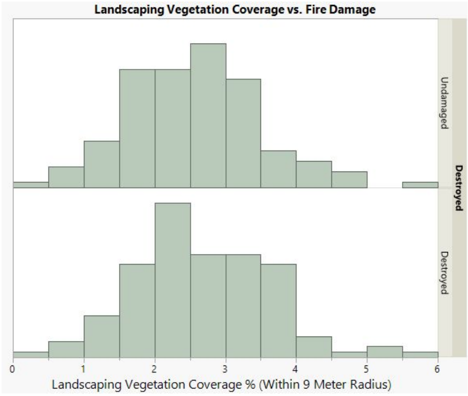


Figure 2. Distribution of Burn Rate by Landscape Vegetation Coverage (Not Significant)



II. Background and Data

The goal of this research as I understand it is to determine which factors (if any) of homes in the areas impacted by the 2017 Thomas Fire might be associated with a higher or lower probability of being burned in the fire. Data were collected on about 450 homes for housing features like fencing, and roof type. Additionally data were collected on and about 250 homes for terrain coverage data. Data were sourced from combining a preexisting government dataset on homes burned in the fire with data collected by Rudy Uribe via Google Street View, as well as terrain data gathered by gis software.

Homes were added to the dataset in pairs to control for confounding variables and to have about an equal amount of burned and undamaged homes in the dataset. The reasoning behind this approach is further detailed in the statistical methods section (page 5).

In table 1 and table 2 below is a summary of all the variables that were investigated initially (but not necessarily included in the final model). NA's and unknowns are generally removed.

Table 1. Quantitative Variables, Including Terrain percentages and Year Built (n = 252)

Factor Name	**Details**	**Min (9m)**	**Median (9m)**	**Max (9m)**
Year Built	Year the home was constructed	1925	1973	2005
Asphalt %	% Terrain within 1.5, 9, 15 meters	0.00%	0.40%	2.46%
Developed %	% Terrain within 1.5, 9, 15 meters	0.00%	0.29%	2.05%
Forest %	% Terrain within 1.5, 9, 15 meters	0.00%	0.02%	2.54%
Herbaceous %	% Terrain within 1.5, 9, 15 meters	0.00%	0.17%	2.708%
Landscape Veg %	% Terrain within 1.5, 9, 15 meters	0.37%	2.53%	5.83%
Lawn %	% Terrain within 1.5, 9, 15 meters	0.00%	0.51%	3.97%
Sidewalk %	% Terrain within 1.5, 9, 15 meters	0.00%	0.98%	3.20%
Wildland %	% Terrain within 1.5, 9, 15 meters	0.00%	0.07%	2.92%

Table 2. Categorical Variables (n = 450)

Factor Name	**Details**	**Levels (Categories)**
House Damage **(Response)**	Whether the home burned	Burned, Unburned
Window Panes	Type of window panes	Multipane, Single, Unknown
Exterior Siding	Type of siding materials	Combustible, Fire Resistant, Unknown
Roof Construction	Type of roof material	Combustible, Fire Resistant
Deck/Porch	Deck/Porch type	Combustible, Fire Resistant (Masonry/wood), Unknown, NA
Eaves	Enclosed/unenclosed eaves	Enclosed, Unenclosed, Unknown, NA
Vent Screen	Had vent screens or not	No, Yes, Unknown
Fence Type	Type of Fence	Combustible, Non-combustible, None, Unknown

All variables included in the tables above were at least considered, but a cursory investigation determined that many factors were unlikely to be significant predictors of burn rate. Therefore these variables were not included in the model building process as they would have weakened our ability to detect any significant factors.

For example, terrain percentages (in figure 3 above) were investigated but determined not to be useful in predicting burn-rates, so these were not considered in the final model that included the Fence type as a significant predictor.

III. Statistical Methods

Before discussing the model theory and specific analysis techniques, I'd like to discuss the reasoning behind the sampling strategy we used to add houses to our dataset since it was somewhat unique.

The sampling plan we recommended for picking houses from this dataset was to create pairs of unburned and burned houses for comparison. The main reason we suggested this approach was to reduce the confounding impact of other factors that were not of interest in the study but may have had an effect. For example, one of these potential confounding effects might have been whether a fire engine was able to reach the house as it burned, and how quickly the engine arrived. By pairing each burned home with an unburned home located very close, the idea is that hopefully, the two homes had similar suppression resources like fire engines available. And the hope is that other potential confounding variables like the effect of wind patterns or local weather and temperature might be controlled for similarly.

An example of a common application of this sort of technique is in clinical drug studies. For example in a drug clinic, the participants in the study will often be paired where one receiving a drug and one will receive a placebo, with the goal being to determine whether the drug has a significant effect. The idea is that each of the two people in the pairing have similar health, age, physical fitness etc. to each other- meaning that any differences we detect in the effectiveness of the treatment will most likely be caused by the drug itself instead of maybe being related to one of those other factors like health or age. So this is similar to how in the Thomas Fire study we would like to rule out factors like location, wind speed, or suppression resources as much as possible, and focus on the factors we care about instead. And while drug studies are an instance of a designed experiment as opposed to our project which is an observational study, we can still apply the idea of this "**matched pairs**" technique to get many of the benefits of this experimental design approach. Note: though we employ this strategy as a sampling technique to potentially reduce variance, we are not intending to use the exact statistical analysis/modeling that might be employed in the drug study as our assumptions differ.

Another reason for recommending this sampling approach over a full census, ie. sampling every single house available is to be efficient with resources. Taking a full census would not be likely to improve our ability to detect significant results very much compared to taking a more selective sample like this paired design, but it would mean a lot more data would need to be collected. And it is quite likely that much of this extra data would end up being unhelpful/uninformative to our final model. More data is always good, but our hope is that this sampling design will be more than adequate for the study goal. However, a full census is still available as an option if the resources are there.

Now I will move on to the model building and analysis. As mentioned earlier, while all variables were considered initially as a potential significant factor, after initial cursory investigation many were determined to be not worth considering for the final model. The variables that were not worth considering further included all the continuous variables for terrain type. However most of the categorical factors were at least considered during the model building process.

From the reduced list of potential variables that might have an associated difference in burn rate, I proceeded with **stepwise selection techniques** to further reduce the model and determine if there were any significant factors. Stepwise selection is a process where we start with a pool of potential factors that might be included in a final model and systematically remove the least helpful factor until we are left with a strong model. In our case, this meant almost all variables were removed as most did not help strengthen our model.

Our final model is an instance of a **logistic regression model,** a statistical model frequently used to analyze the relationship between a binary response (one that can be treated as a yes/no question, like "did this home survive?") and several explanatory variables (in our case, our home design and terrain factors). Specifically, a logistic regression model takes the form of an equation that predicts the odds of an event (odds of a house burning or surviving) based on any combination of explanatory variables. From the model results, we are able to make statements about the size and significance of any of these explanatory variables, such as how the odds of burning in a fire might differ for different homes.

IV. Results and Discussion

The final logistic regression model with calculated coefficients (model created in R) is as follows:

Table 3. Final Model
Model Summary:

Coefficients

Term	Coef	SE Coef	z	P-Value
Constant	-0.331	0.175	-1.892	0.508
Fence Type (Non-Combustible)	0.812	0.225	3.608	<0.001

AIC: 468.13

As an equation, this model can be written as:

Est. Log Odds of Home Surviving = -0.33 + 0.812 * (Non-Combustible)

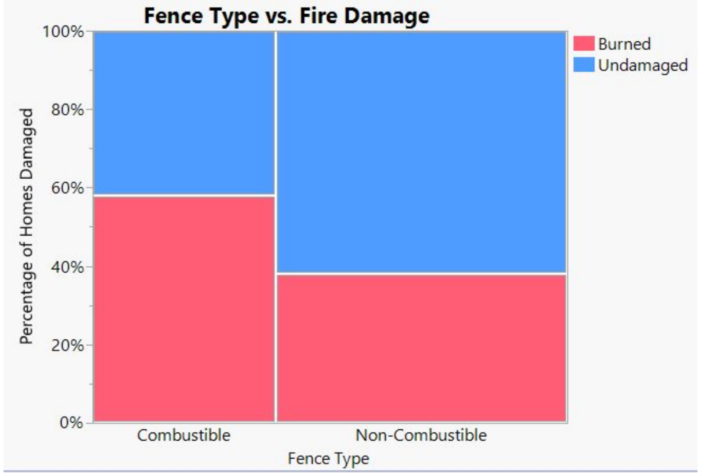
Which can also be written as **odds of home surviving = e^(-0.33 + 0.812 * (Non-Combustible))**

Where ‘Non-Combustible’ takes a value of 1 for homes with non-combustible fencing and 0 for homes with combustible fencing.

If we plug in 1 to this equation and calculate, we can find the estimated odds of a home with a non combustible fence surviving the fire which is 1.62:1, meaning their estimated probability of survival is 0.618 or 61.8%. If we plug 0 into the equation to represent combustible fenced homes, we find their odds of surviving the fire is 0.72: 1, meaning the estimated probability of survival is 0.417 or 41.7%

If we take the ratio of these two odds, we can determine an estimate for how much more likely a home is to burn with a combustible fence than a non combustible fence. So the ratio 1.62/0.72 tells us that having a non-combustible fence is associated with 2.22 times higher odds of surviving in this fire (222% increase in the odds of burning) relative to having a combustible fence. This difference was found to be significant at a 0.0002 significance level (<0.001).

Figure 3. Distribution of Burn Rate by Fence Type (Significant, p < 0.01)



As can be seen from the model (Table 3) and visual (Figure 3), fence was the one factor determined to have a significant association with burn rates in homes like the ones sampled. For all other variables/potential factors, we did not have significant evidence to conclude association with a difference in burn rate.

V. Limitations and Suggestions

There are two primary limitations to our results so far that I would like to discuss. The first is the issue of generalization, ie. determining how broadly our statistically significant results can be applied. The main idea here is that we should not try to apply our results to any homes or scenarios that the sample we took is not representative of. So because sampling was restricted to a few specific Central California communities, our results would only apply to homes that are exposed to similar weather, have similar topography, similar home design, and any other aspect that might affect our results. Likewise, our results only apply to fires that are similar to this particular fire, whatever that might mean (similar temperature, flame size, etc.).

Any differences found that were not statistically significant can still be discussed, but no attempt should be made to generalize to anything beyond this particular sample. So a discussion of a finding that is interesting but not statistically significant might go something like “In this particular fire, here’s what we found…” rather than “In fires similar to this fire, this result might apply…”

The second limitation I’d like to review is our ability to show correlation vs. causation and what that might mean for this study. By definition, as this is an observational study (as opposed to a controlled experiment), we are unable to conclude causation. So that’s why we word things like “combustible fences are associated with a higher burn rate” instead of “combustible fences might cause homes to burn more.”

This also means that any effects we found might be confounded by another effect. For example, it's possible that combustible fences were more common in homes that were out on the edge of the community (ie. homes that are less packed into the suburb might have more land and therefore more incentive to have big wooden fences). This would mean that maybe the homes with combustible fencing burned at a higher rate partly just because they were more exposed to the fire on average, and not entirely due to their fencing.

That is just one example of a potential confounding effect- it might not have really happened that way but it is something to keep in mind for discussing results and for designing any future iterations of the study.

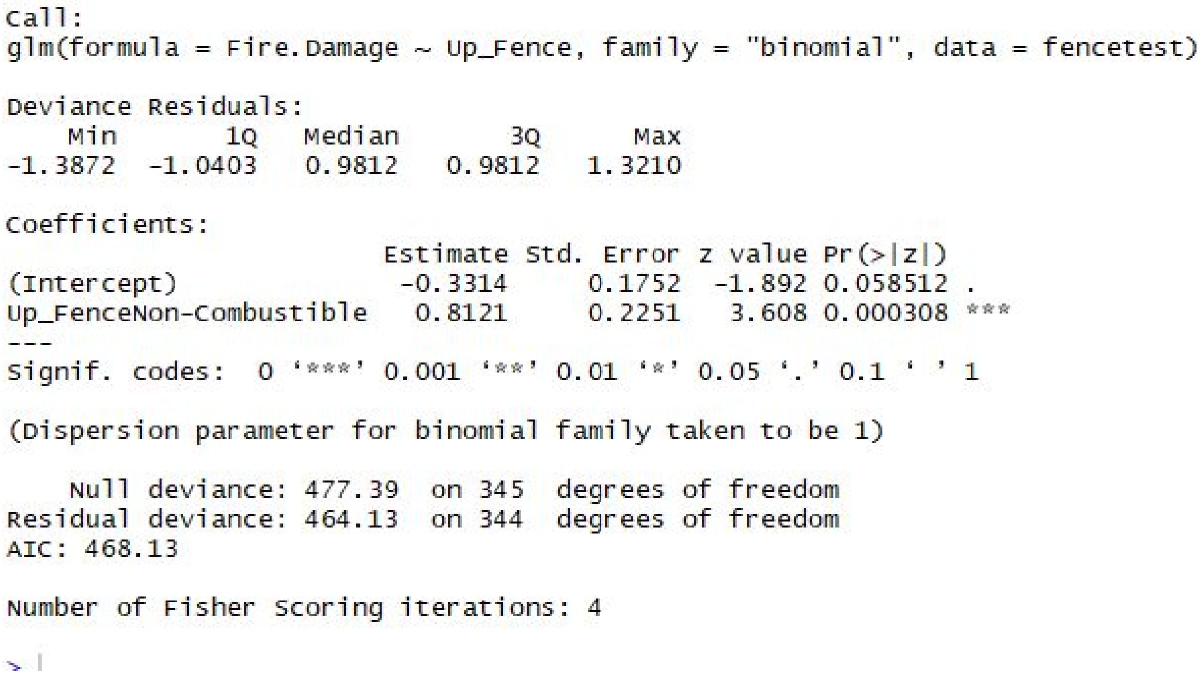
As far as suggestions for improving/expanding on the study, I have a few ideas on some further approaches that could be tried if resources allow. Perhaps most obvious is expanding the study beyond the Thomas Fire. Having more data gives us a stronger ability to detect any effects that might exist, and it also would help us generalize our results more broadly to be sampling from more fires. Note that if this approach was taken, the sampling strategy might have to change to include more randomization.

The next suggestion is specific to the terrain data variables. Rather than strictly testing the percentage of each terrain type, another approach might be to classify homes more generally based on their defensible space. I'd suggest this because it seems to me like the terrain percentages aren't telling us as much as we'd like about the specific terrain features of a particular house. With our current approach, we might have two homes that are both "5% wildland" but one house has all their wildland in a properly constructed defensible space separate from their house, and the other home has wildland randomly strewn all over the place. While both of these homes would be treated the same in our current model, perhaps it makes sense to classify them differently as I'd guess the first home is less likely to burn.

So what this might look like could be manually reviewing each home's terrain (either via streetview or whatever is best) and giving them a rating, say from 1 to 5, with 1 representing "poor defensible space" and 5 representing "excellent defensible space". From there, you might proceed by analyzing whether the proportion of homes burned differed between the different categorizations of defensible space.

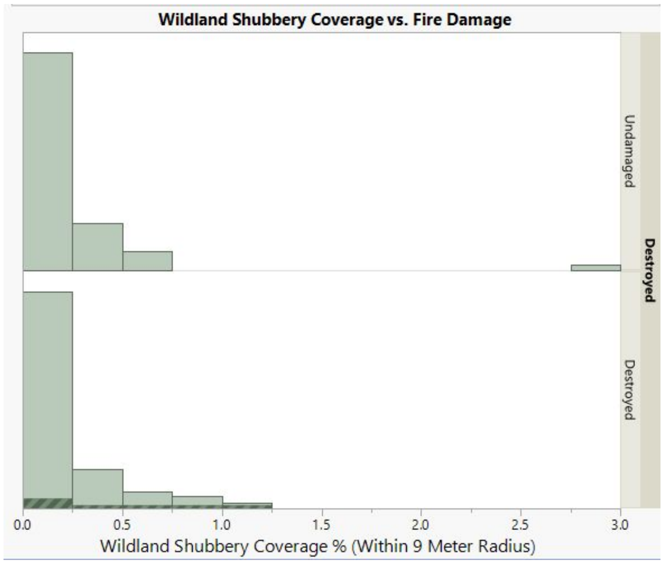
VI. Technical Output/Appendix

Full Model Output (R)



```
Call:
glm(formula = Fire.Damage ~ Up_Fence, family = "binomial", data = fencetest)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.3872  -1.0403   0.9812   0.9812   1.3210  

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)    
(Intercept)              -0.3314     0.1752  -1.892 0.058512 .  
Up_FenceNon-Combustible   0.8121     0.2251   3.608 0.000308 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 477.39  on 345  degrees of freedom
Residual deviance: 464.13  on 344  degrees of freedom
AIC: 468.13

Number of Fisher Scoring iterations: 4

> 
```

Additional Visuals


Wildland Shubbery Coverage vs. Fire Damage
Undamaged
Destroyed
Destroyed
Wildland Shubbery Coverage % (Within 9 Meter Radius)