Fall 2025 CAMCOS Wildfire project

Meeting with Wilkin Lab

2025-09-11

Data Issues

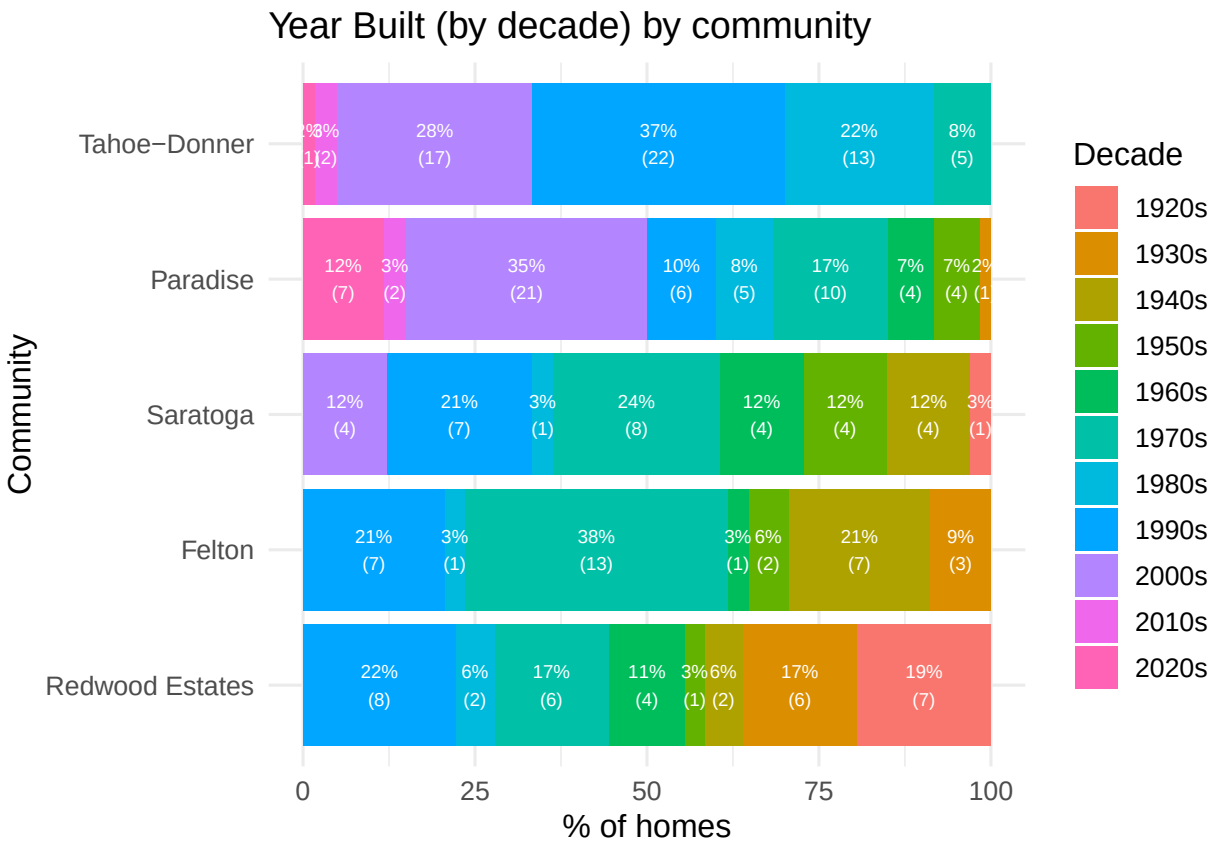
We took a closer look at the data and noticed a couple of issues that were mentioned in the email from Martina to Kate and Basanth on 9/2/25. We should talk about these and resolve questions (if they haven't already been addressed.)

The questions were:

- What is Parcel ID? An alternative enumeration of properties?
- We noticed that several properties (the ones that appear in the bottom rows (176 - 222) of the original file (tab: "Field Data") have duplicates among them. That is, the same "UID" is repeated across several rows. The results in these rows differ (different built years, different roof types etc.) For example, property "727e093140" appears in rows 200 - 204 with two shingle roofs and three metal roofs. These rows all have no "year surveyed" in column F. What's up with these properties? Are these consecutive visits to the same property over time? Do you want us to use these data? Or only the most recent observation for each house? Using repeated observations on the same house would make the model to be fitted quite a bit more complicated.
- There is information on the "most recent Cal Fire" inspection. Is there also information on when that most recent inspection took place?
- Do we know how the slope value was calculated? There seems to be one property with a slope of ~55degrees. That's unbelievably steep.
- Some columns (for instance "attic Vents gap size" and "attic Vent materia;") have the entry "unknown" in all rows. I have some memory of you talking about measuring vent gap sizes with a golf tee. That does not make sense if the data set has no known measurements. Are we sure that there is no information about these two variables at all?

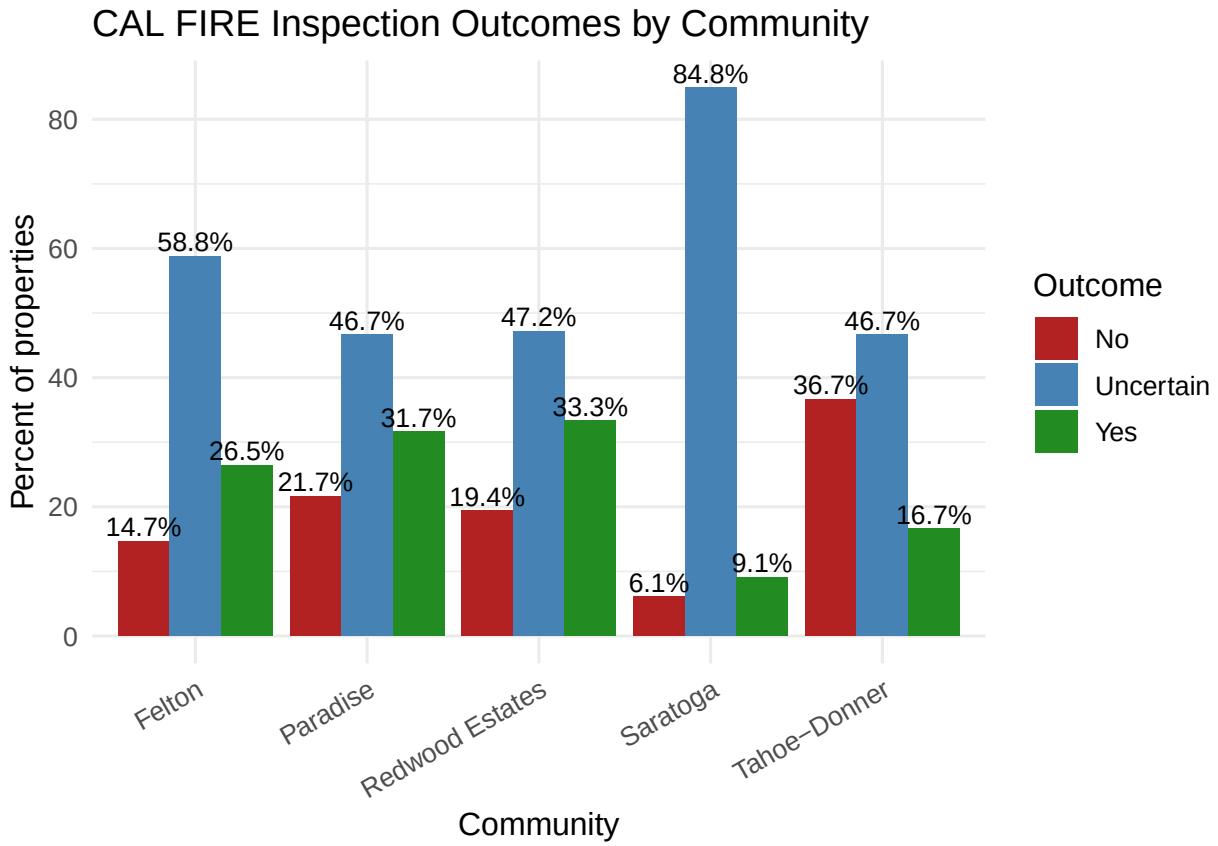
Some Descriptive Statistics

The CAMCOS team has spent a week summarizing parts of the data to look for patterns that can help us find a good model to fit later on. Here are some of the graphs that have found interesting so far.



The age of homes varies widely across communities, reflecting different growth patterns. Tahoe-Donner is dominated by homes built between the 1980s and 2000s (over 85%), making it one of the newer housing areas. Paradise also shows a strong spike in the 2000s (35%) with a secondary rise in the 2020s, suggesting recent rebuilding activity. Felton's stock is older, with nearly 75% of homes built before 1980, peaking in the 1970s. Redwood Estates stands out as the oldest community, with over a third of homes built before 1950. Saratoga has a more even mix, with houses spread from the 1940s through the 2000s, reflecting steady development over time.

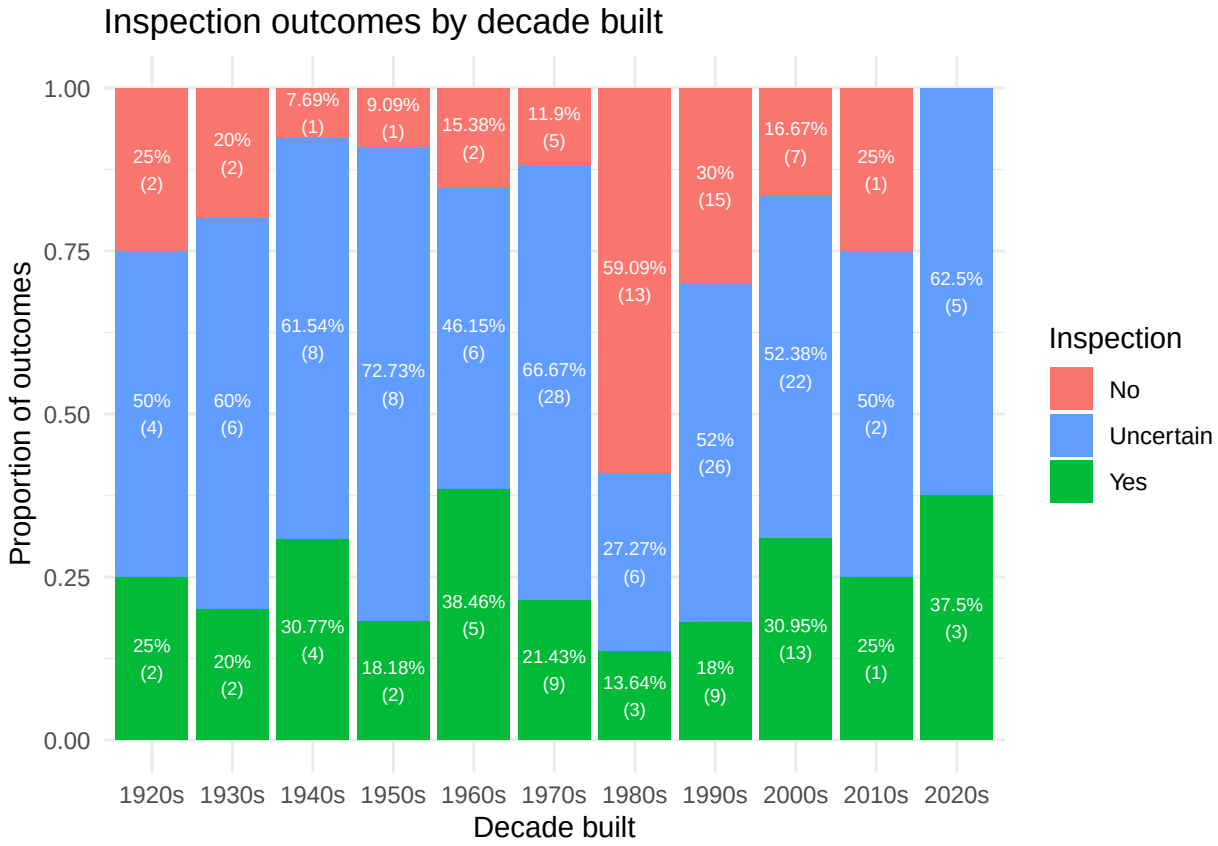
Overall, this highlights that some communities like Tahoe-Donner and Paradise are largely modern, while others such as Redwood Estates and Felton still have older homes that may face higher wildfire vulnerability without renovation.



Across all communities, a majority of properties fall into the "Uncertain" category, making it difficult to draw strong conclusions on compliance.

- Saratoga leads with about 85% uncertain outcomes and very few clear passes or fails, suggesting incomplete or inconsistent reporting.
- Felton follows with nearly 59% uncertain. Paradise and Redwood Estates each have about one-third of homes passing, but nearly half uncertain.
- By contrast, Tahoe-Donner stands out with the highest failure rate (37%) and a low pass rate (17%), showing more properties in need of fire safety improvements.

Overall, the dominance of uncertain results limits clarity, but Tahoe-Donner reveals the most definite challenges.



Inspection outcomes vary noticeably across decades, but one consistent trend is that uncertain results dominate across almost every decade, generally above 50%.

The 1950s take the lead with over 70% uncertain, while the 1980s stand out as the exception, with only 28% uncertain but the lowest pass rate (13.6%) and the highest failure rate (60%).

In terms of passes, the 1960s (38.5%) and 2020s (37.5%) show the strongest results, with 2020 homes showing no failures at all. Interestingly, the 1920s, 1930s, and 2010s all show equal pass and fail rates (25%, 20%, and 25% respectively). The 1980s and 1990s lean negative with fewer passes than fails, and in most other decades, failure rates are more than double the pass rates.

Overall, while uncertainty clouds the picture, the 1980s clearly emerge as the weakest-performing decade, while newer homes are trending toward better outcomes.

Looking ahead

It seems as if the eventual goal is to model hazard status as a function of year built (or perhaps time between built and inspection?) and location. With location as a random factor and time as fixed. We need to discuss what exactly to use as the hazard status. Here are some options:

- The individual columns in binary form (well prepared vs anything else)
  - Potential problems: There are many columns where the sample size for "well prepared" is very small. That will lead to low power.
- The individual columns as multivariate responses (with categories yes, no, unknown, for instance)
  - Which comparisons are we interested in at the end of the day? "Well prepared" against "not well prepared"? "Well prepared" against "unknown"? Something else?
- One could create a hazard status (perhaps individually by category - roof, deck, attic vents etc.) by combining information from various columns. How to combine the columns is somewhat subjective - need subject matter expert input.

Different statistical models that could be used depend on the response variable "hazard status". If the response is quantitative (some sort of hazard score), we could use a mixed effects linear model. If the response is binary, we could use a mixed-effects logistic regression model. If the response is multinomial, we could use a multinomial GLM.