Weijian (James) Han

November 8, 2012

Distribution of Losses Due to Structural Fires in Berkeley, CA

Insurance companies often need to study the distributions of fire losses in order to set appropriate rates for premiums and deductibles. Here, we hope to use the available data of 282 individual cases of structural fires in the city of Berkeley, California, occurring between August 2004 and September 2012, to construct a model in order to predict the amount of loss in future fires.

Part I: Distribution of Total Losses

When looking at only the monetary losses of our 282 cases, the majority of these cases suffered losses less than $100,000, which is only 2% of the largest loss of $5,000,000 (Figure 1.1). However, when categorizing the monetary losses logarithmically, we can see that it follows a bell-shape (Figure 1.2), and if we graph the log of the monetary loss of each individual case, we can see that the majority of the cases suffered a loss between $$10^4$ and $$10^5$ (Figure 1.3), corresponding to the tallest bar in Figure 1.2.

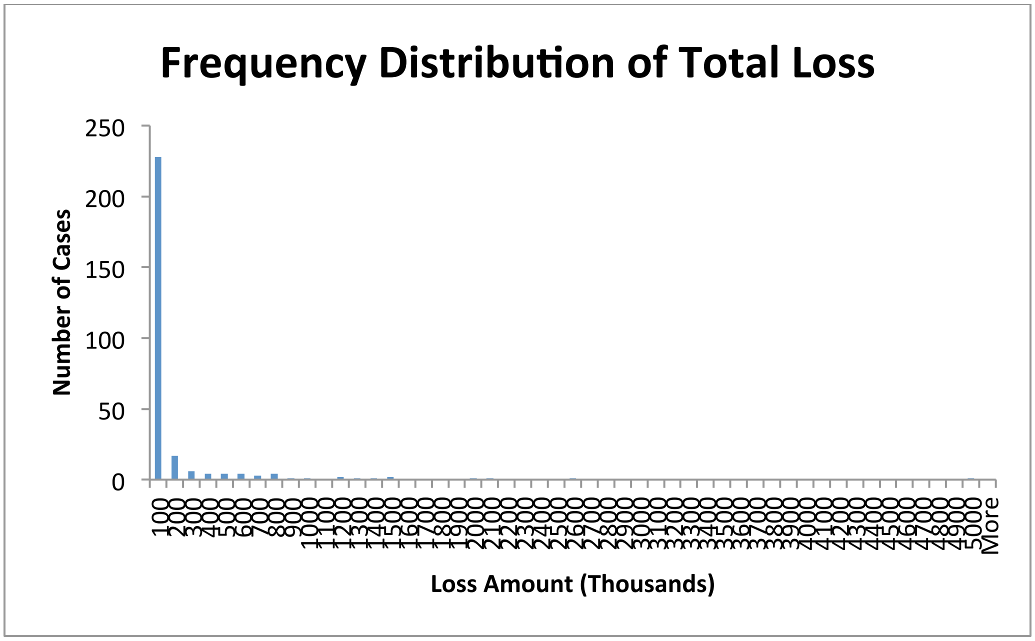


Figure 1.1

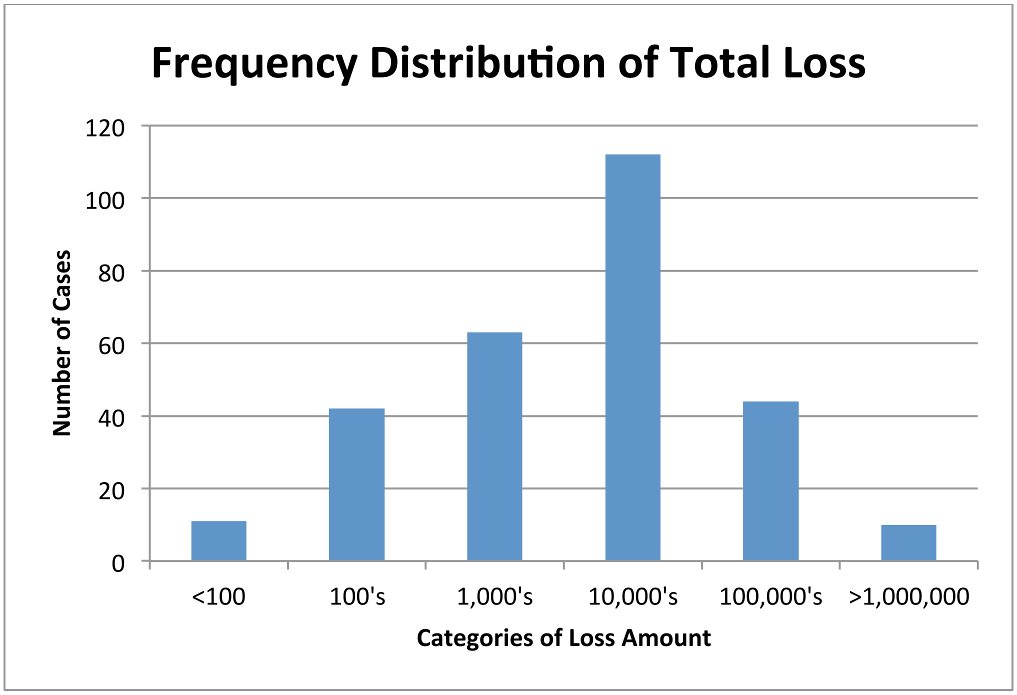


Figure 1.2

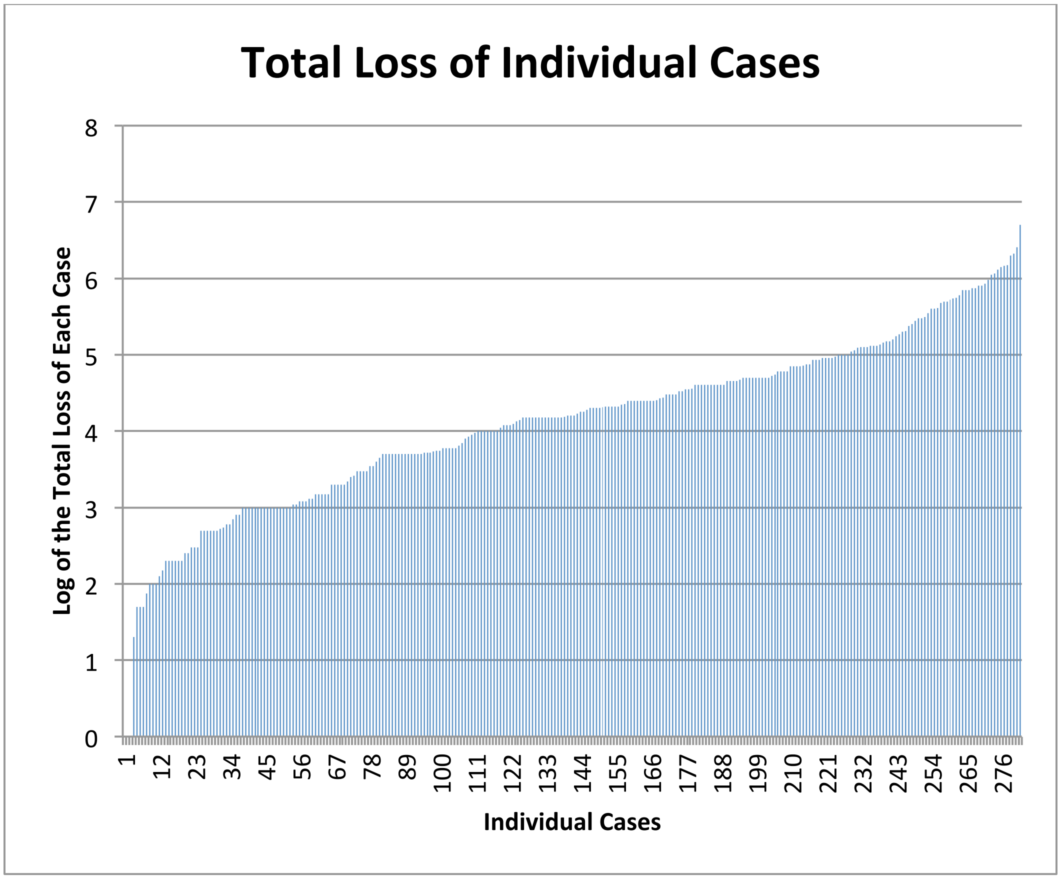


Figure 1.3

Part II: Relationship between Property Loss and Content Loss

The total loss mentioned above consists of property loss and content loss. Since it is easier to calculate an accurate value for the property loss of a fire, it would be useful to determine a relationship between property loss and content loss. When plotting content loss against property loss of the 282 cases (Figure 2.1), we can see that most cases cluster around the origin, on which there is no loss whatsoever, and a few outliers could greatly influence the correlation. In fact, when we take out the four outliers and generate the same graph (Figure 2.2), a linear association seems much more obvious.

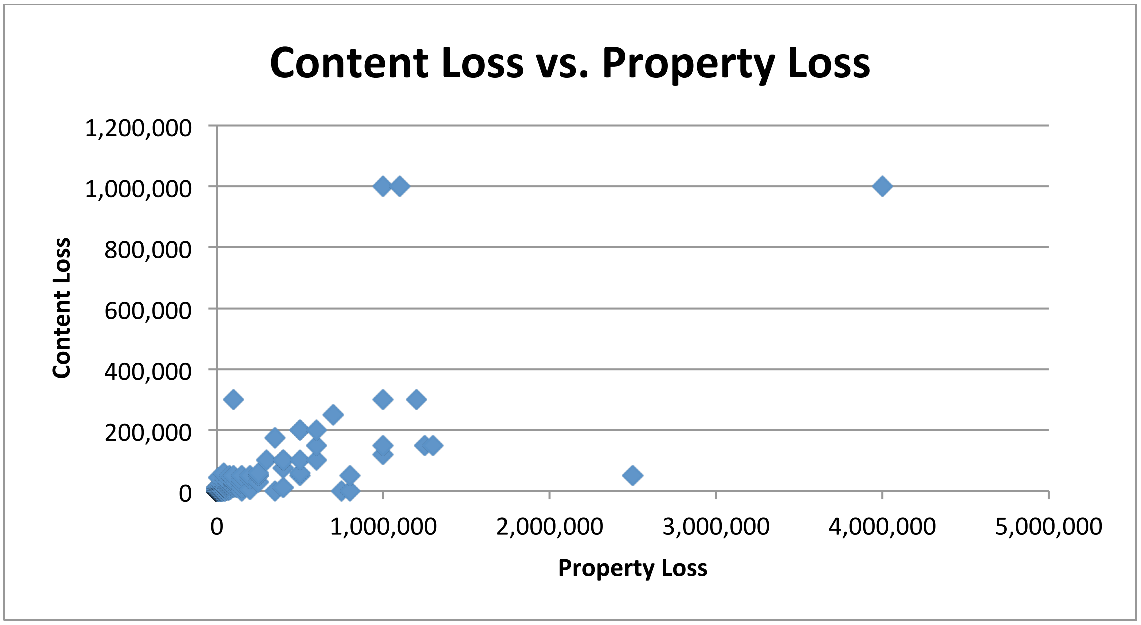


Figure 2.1

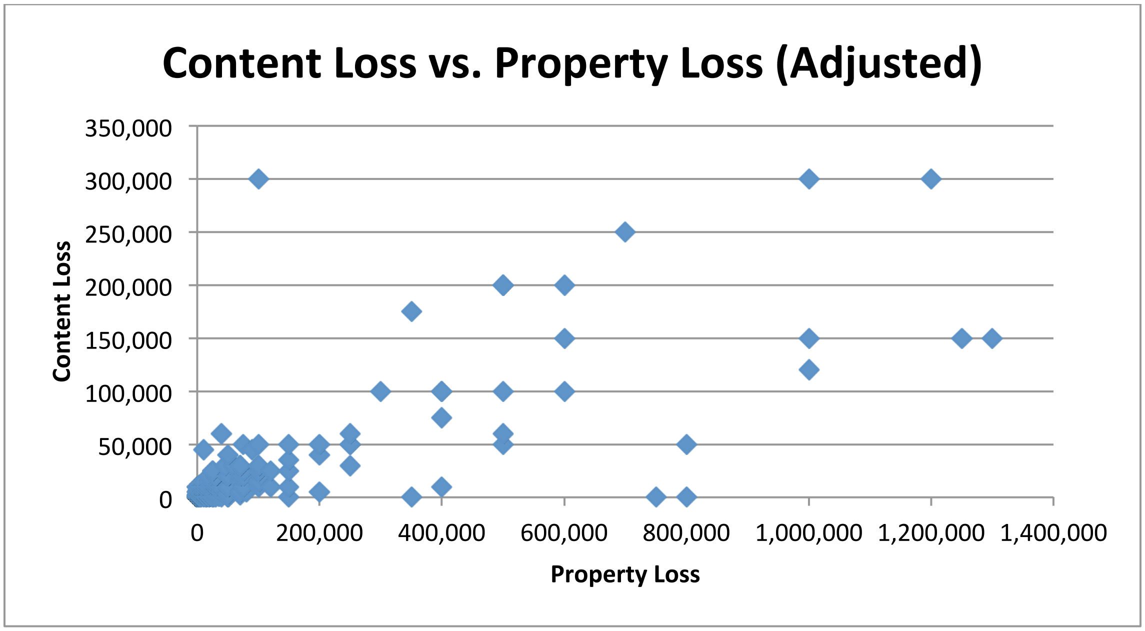


Figure 2.2

By taking out the four outliers, the correlation of the linear association between content loss and property loss increases from 0.707 to 0.757. Calculated with 278 data points, the following equation can be used to predict content loss with property loss:

$$\hat{y} = 0.169x + 4607$$

Part III: Predicting Loss Using Initial Property Value

If content loss could be predicted using property loss, then what can we use to predict property loss? The most logical choice would be using the structure's initial value. However, when we plot property loss against the initial property value* (Figure 3.1), it is obvious that such a prediction will be very inaccurate. Even when we take out some outliers, we could see that the graph still follows the same pattern (Figure 3.2), with the property loss spanning anywhere between 0 and the initial property value.

*Because the Berkeley Fire Department only recorded the initial property values for 65 of the 282 cases, these are the only cases we could work with.

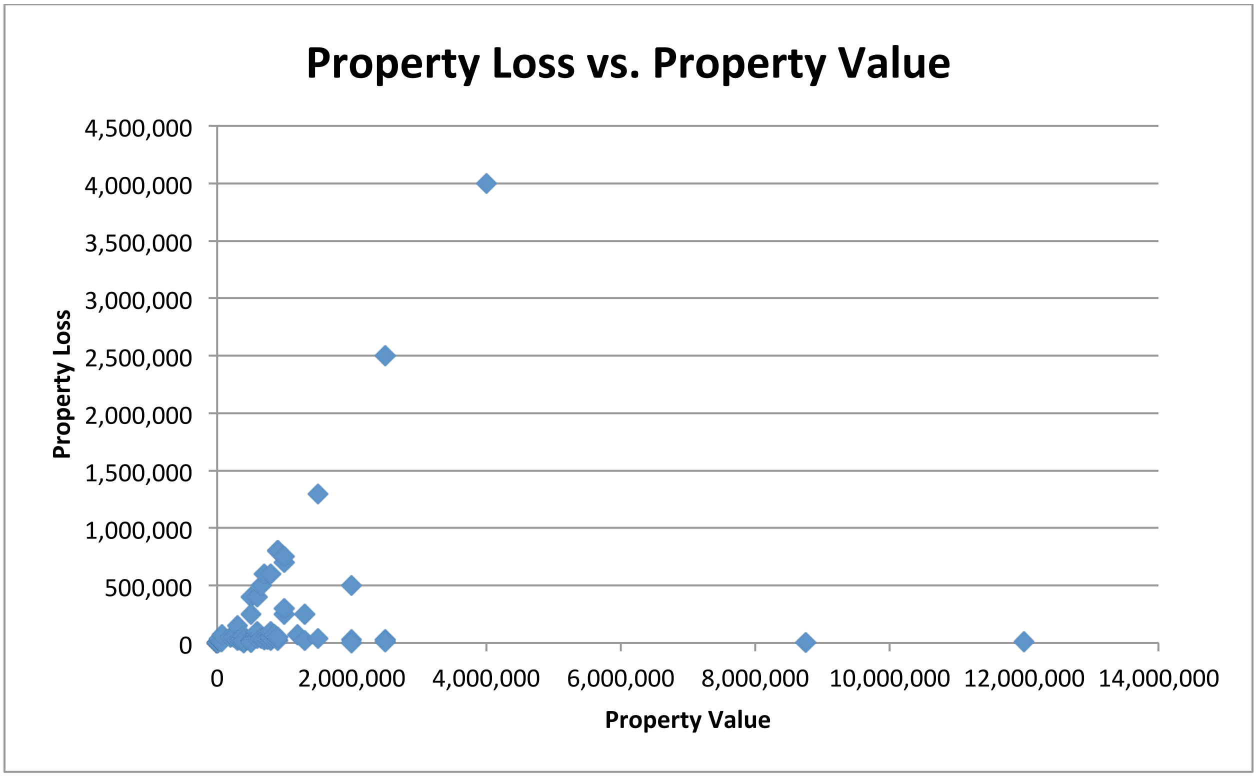


Figure 3.1

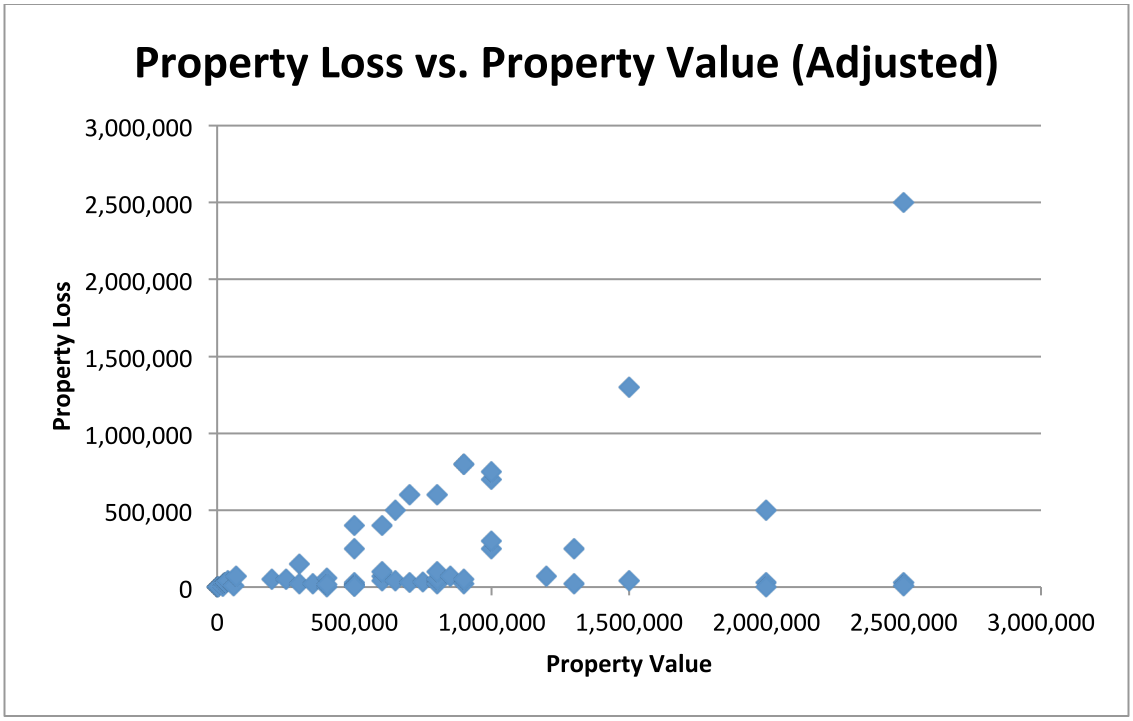


Figure 3.2

It makes sense that we could not simply use a linear line to predict the property loss from the initial property value, because that would imply that most fires suffer the same proportion of loss, but if we graph the proportion of loss against the initial property value (Figures 3.3 and 3.4), we can see that it is clearly not the case.

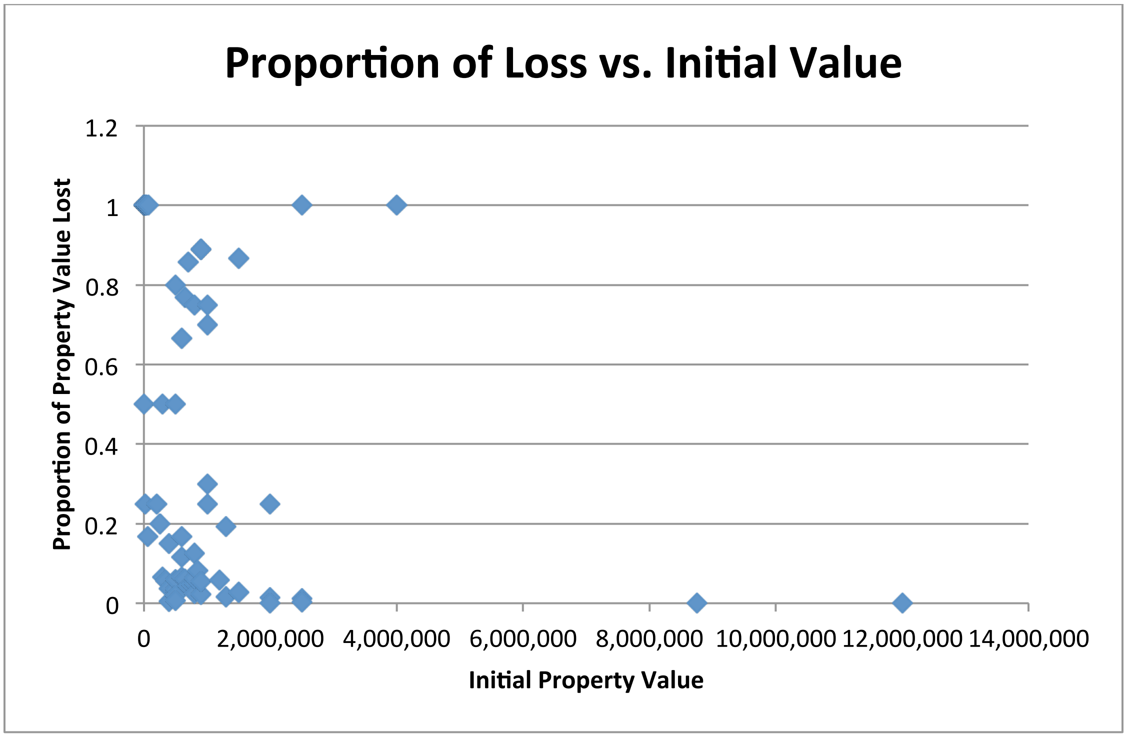


Figure 3.3

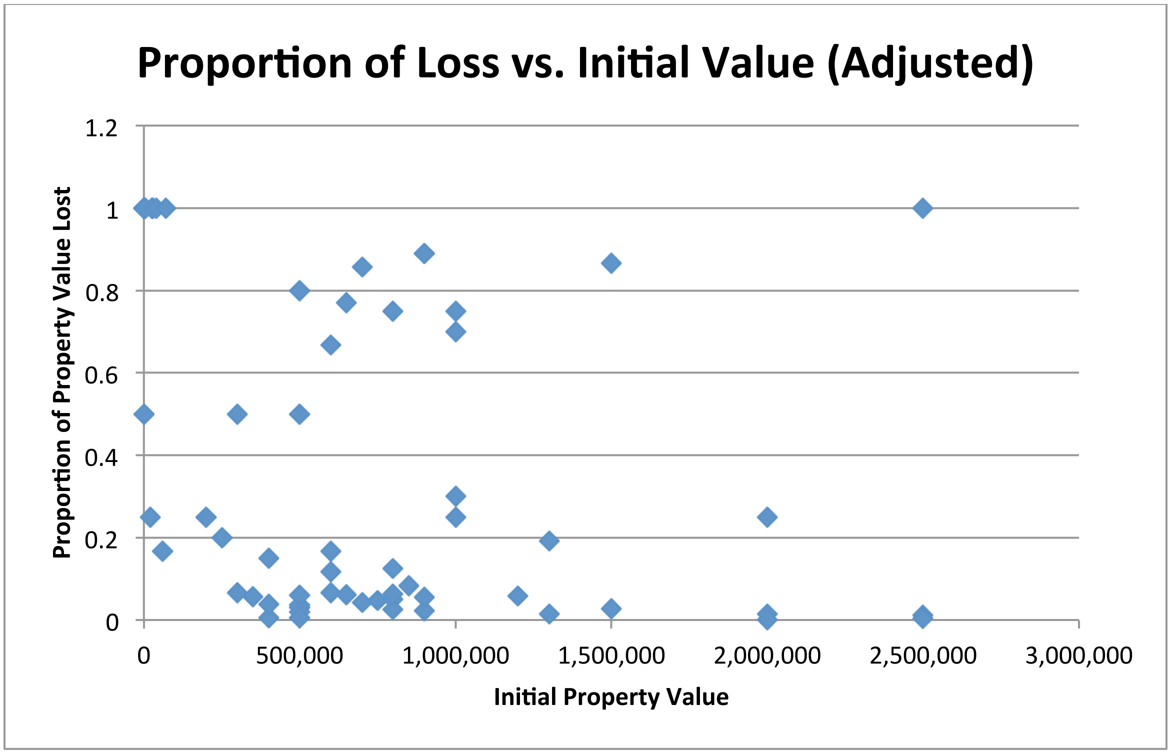


Figure 3.4

It seems like most fires suffer loss proportions close to either 0 or 1, which explains why most data points cluster around the lines y = 0 and y = x in figures 3.1 and 3.2. The loss proportion also seems independent of the initial property value, so we could construct a model to study the loss proportions themselves, which we will do in Part V.

Part IV: The Impact of Extreme Cases

In Part II, we saw that when the four outliers were taken out, the correlation between property loss and content loss increased by 0.05, and in this section, we will further discuss the impact of extreme cases. As we have seen in Figure 1.2 (the same as Figure 4.2 below), the majority of cases suffered losses between $10,000 and $100,000, corresponding to the tallest bar, while the bar representing cases with loss amount greater than $1,000,000 is quite small in comparison. However, the cases in this category (greater than $1,000,000) make up a significant proportion of the sum of total losses. The following figures will demonstrate the magnitude of the impact of these extreme cases.

Category	*Frequency*	*Sum of Losses*
<100	11	545
100's	42	26445
1,000's	63	286,250
10,000's	112	4,322,800
100,000's	44	16,569,000
>1,000,000	10	19,570,000

Figure 4.1

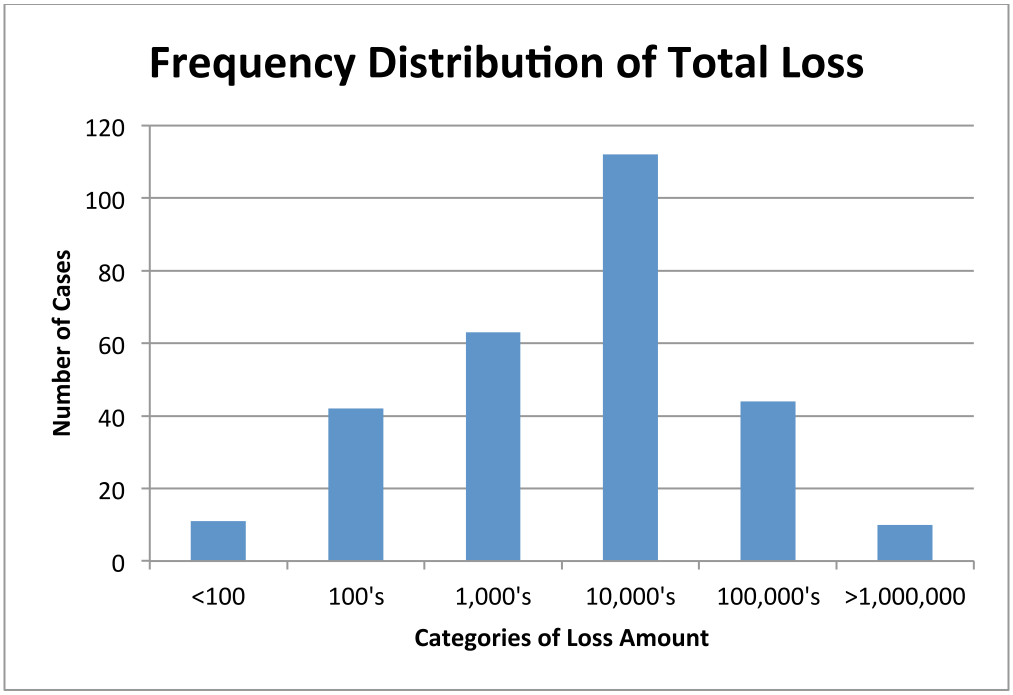


Figure 4.2

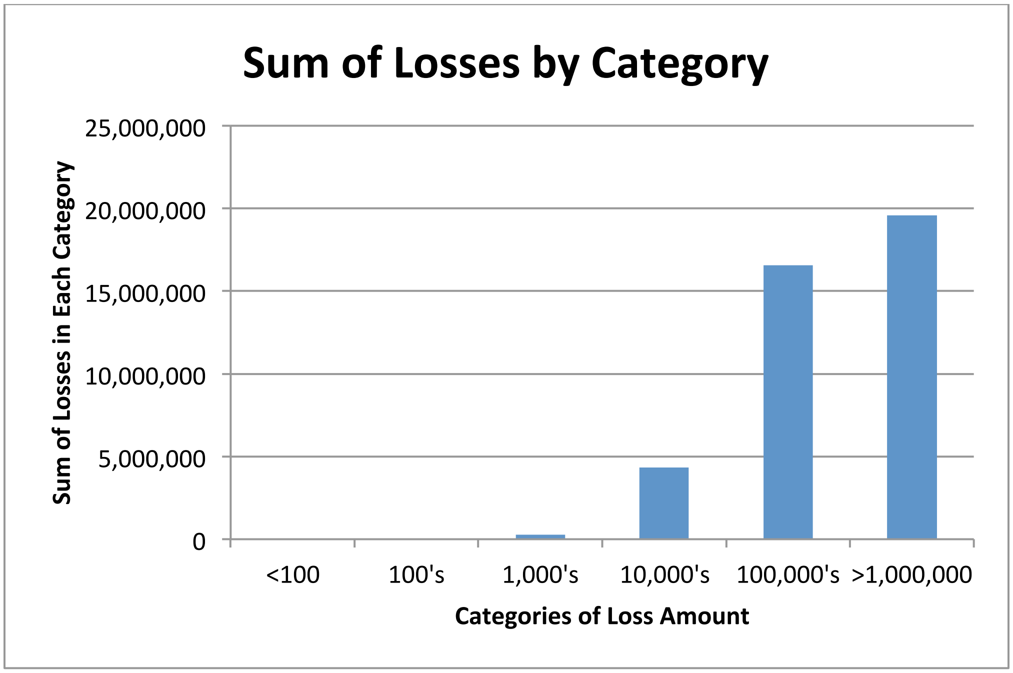


Figure 4.3

Perhaps the proportion of the sum of total losses represented by the largest cases can be better portrayed by the following graph (Figure 4.4). We can see that the worst ten cases, those in the category of more than $1,000,000 of loss, make up for more than 50% of the total losses of the 282cases, while the smallest 182 cases make up less than 5% of it.

Worst N Cases	Percentage of Total
5	32.25
10	50.32
20	66.63
50	87.47
100	95.79

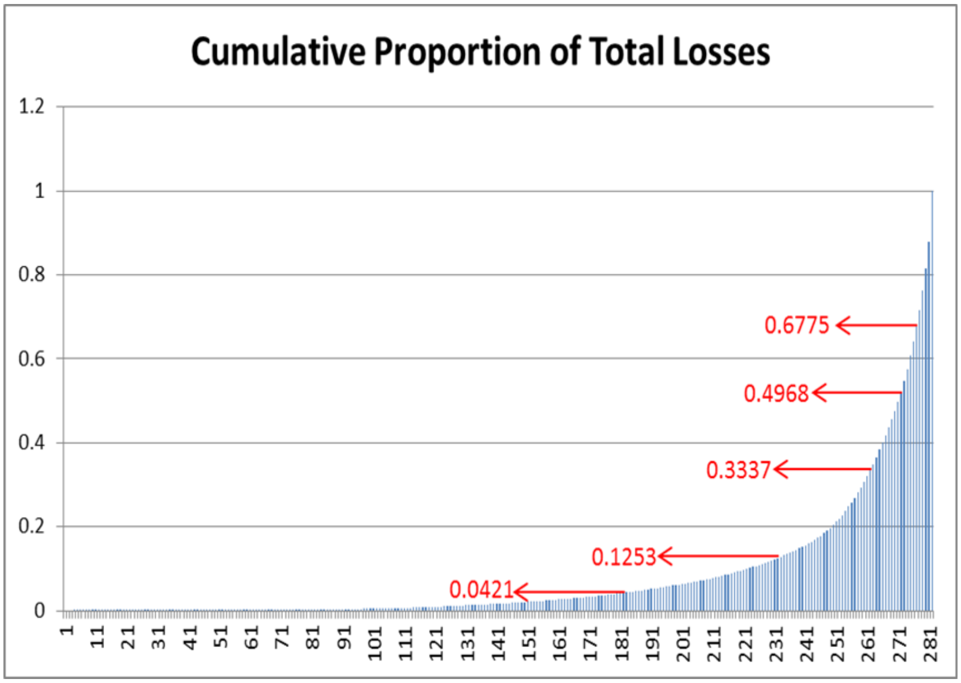


Figure 4.4

Part V: The Probability Density Function

As we have mentioned at the end of Part III, the proportion of losses, which we will define as the random variable X, is independent of the initial property value, therefore independent of such variables as the income level and other indicators of the property owner. Although logically speaking, X would be affected by many factors such as the age of the structure, the distance to the fire department, etc. we will ignore them in our analysis of the fire loss distribution (Figure 5.1).

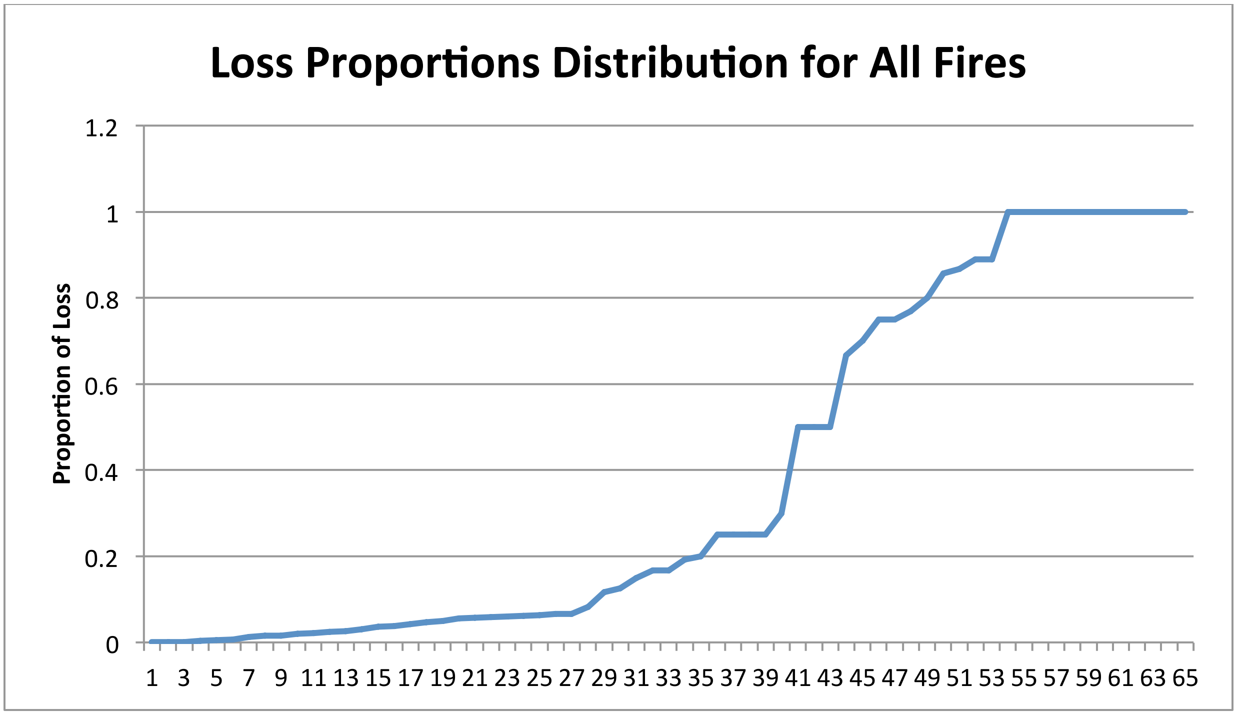


Figure 5.1

The above distribution appears to resemble a logistic function, but instead of approaching a limit, the proportion of loss not only reaches the limit of 1, but a significant number of cases actually had a loss proportion of 1. Therefore, we divided the 65 data points into three categories based on the proportion of loss: small fires (Figure 5.2), medium fires (Figure 5.3), and large fires (Figure 5.4).

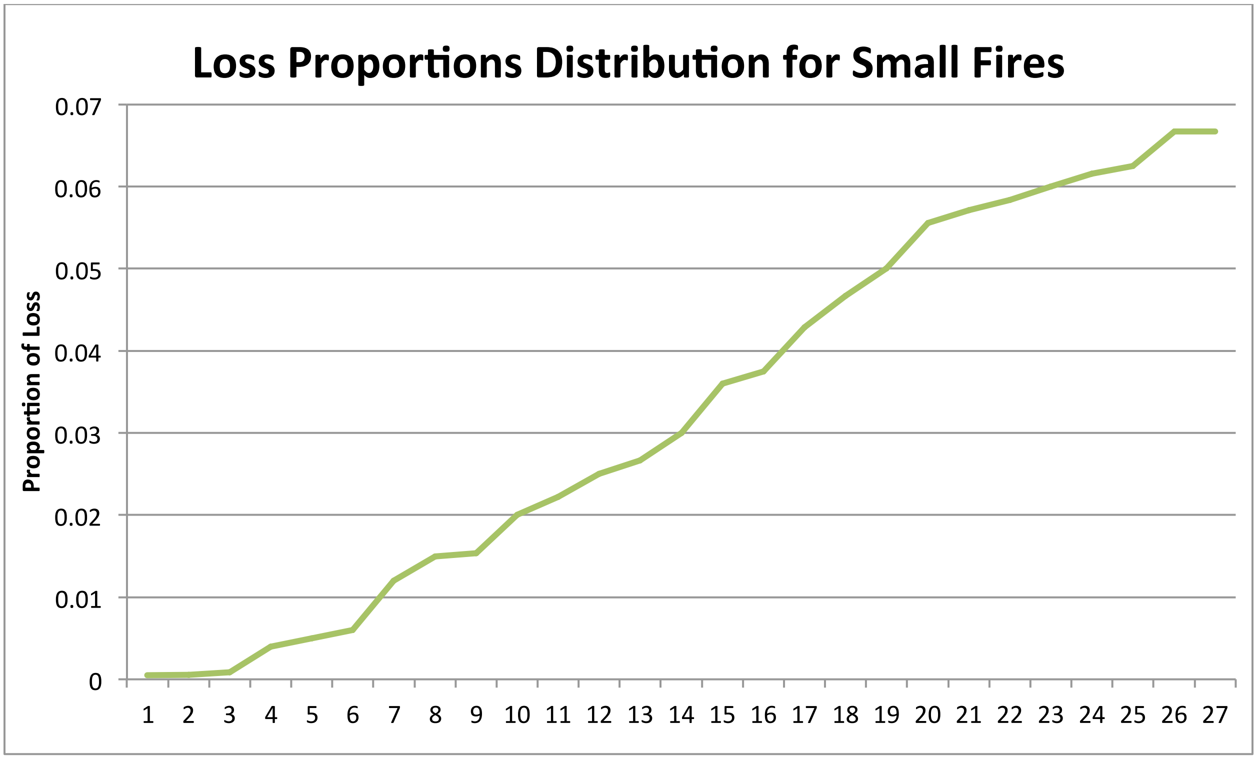


Figure 5.2

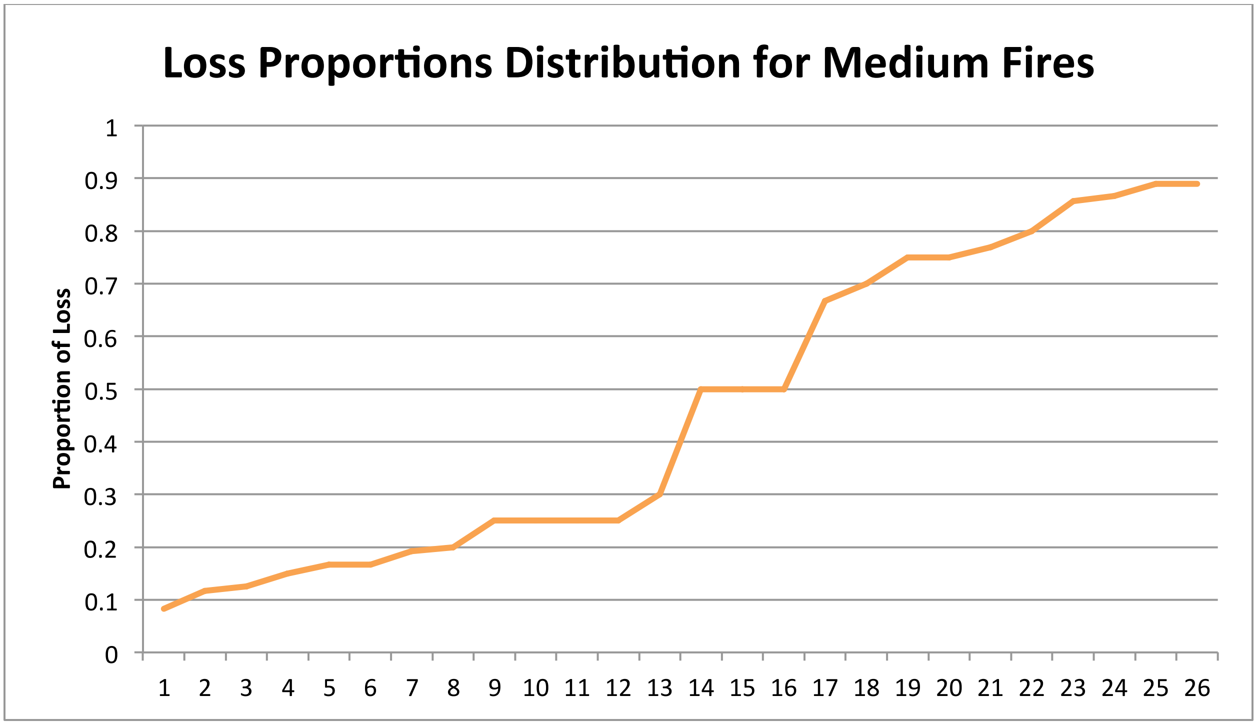


Figure 5.3

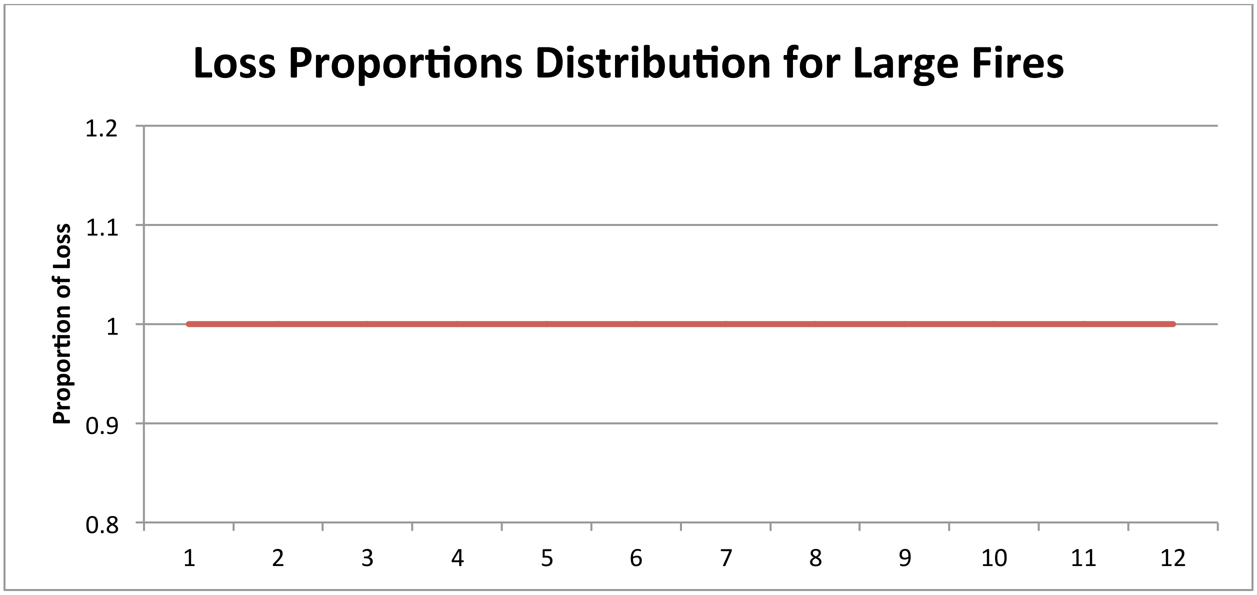


Figure 5.4

For small fires, the relationship appears linear, so the most accurate prediction line would the linear regression line, but this line neither connects with the prediction line for medium fires nor intersects the origin, so we will use the following line to predict the proportion of loss for small fires:

$$G(x) = 6.6554543x$$

This prediction line does not fit the data as well as the regression line (Figure 5.5), but it will make constructing a probability density function more convenient. Notice that the proportion of loss in Figures 5.2 is on the vertical axis, but it will be on the horizontal axis in the following graph, because we are trying to predict the probability that a fire will suffer losses exceeding the proportion x, and this probability is given by F(x) from the original data and predicted as G(x) using our model.

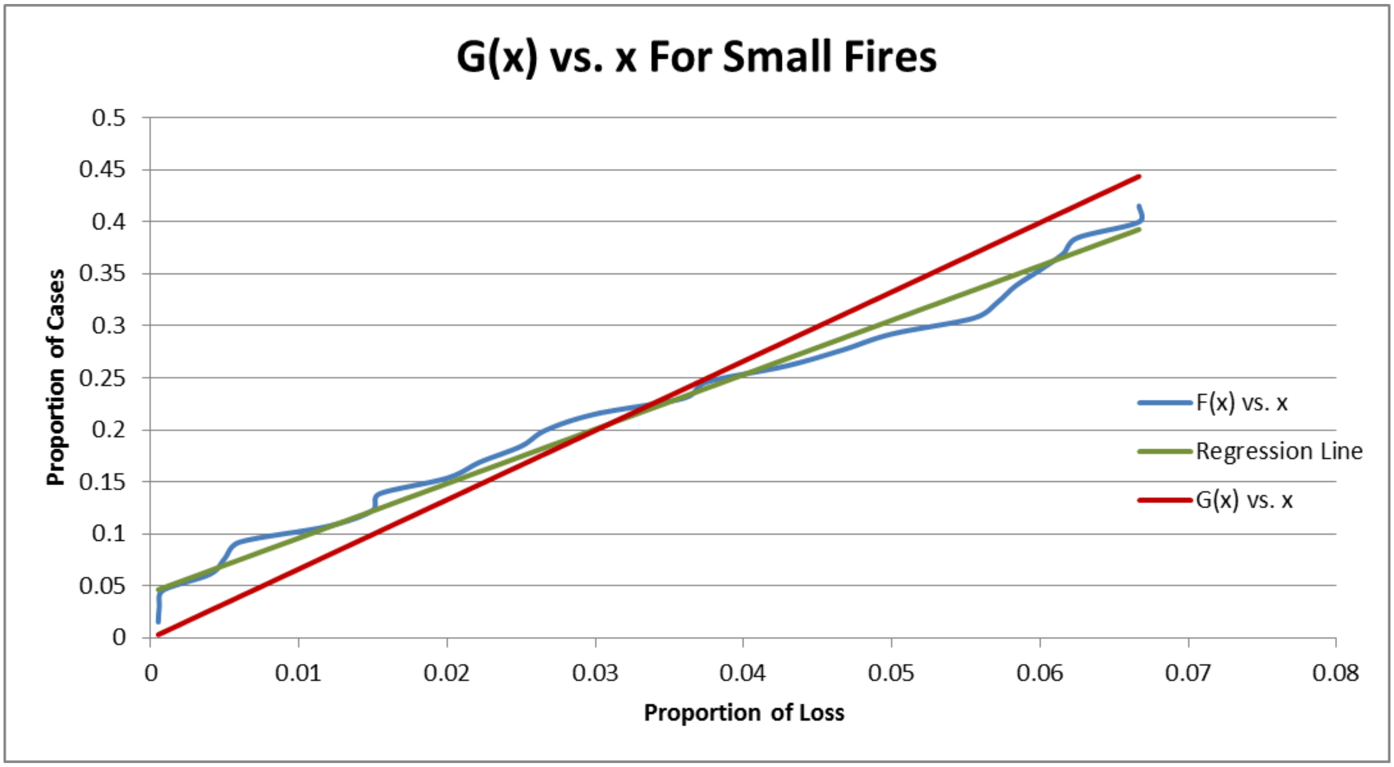


Figure 5.5

Medium fires make up 40% of the 65 cases we are studying here, and they give Figure 5.1 its logistic shape, so we will use the inverse of a logistic function as G(x), which is a logit function (Figure 5.6). The following equation for the prediction line was constructed by trial and error, and since the chance errors for medium fires are much larger compared to small fires, it is sufficient to use the precision of two decimal points.

$$G(x) = 0.18 * \log\left(\frac{x}{1-x}\right) + 0.65$$

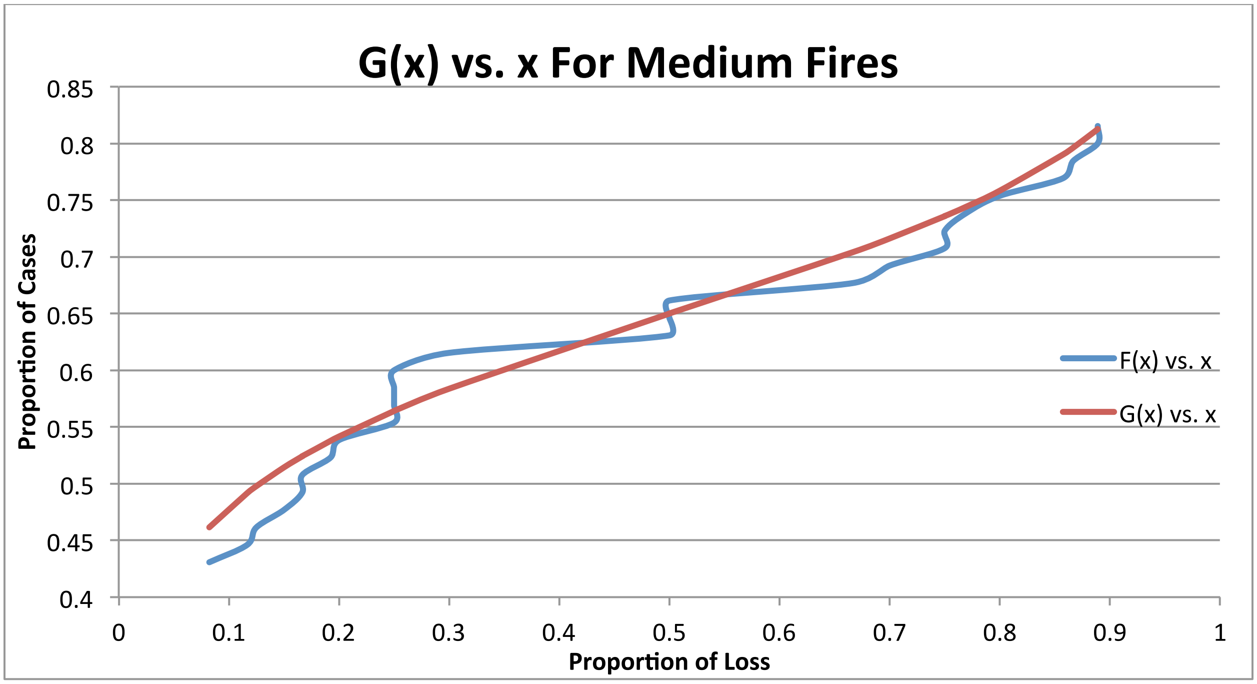
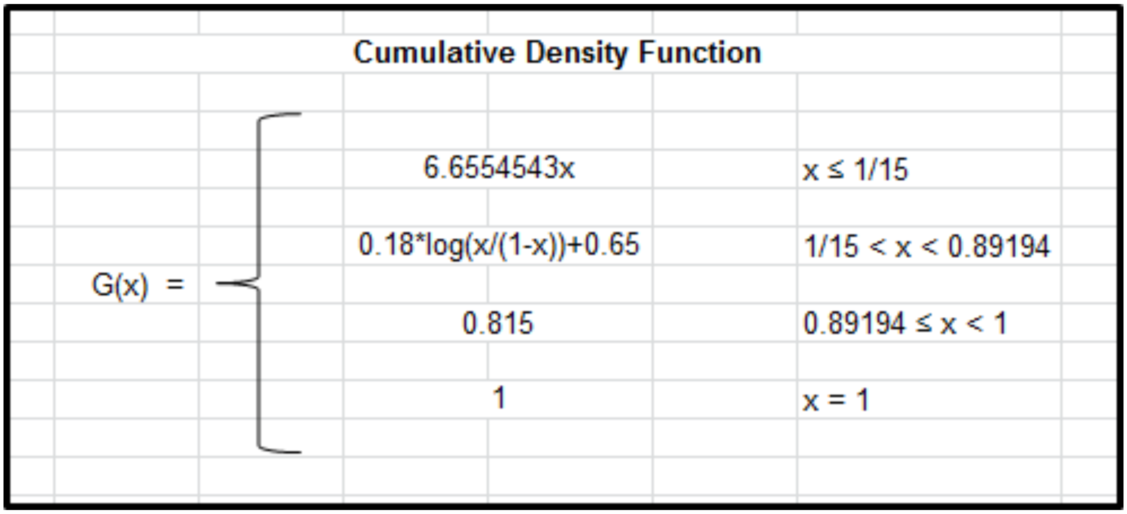


Figure 5.6

A large fire is defined here as a fire with a loss proportion of 1, which has a chance of 0.185 according to the data we have for the city of Berkeley. However, there is a gap between the end of the logit function for medium fires and the point 1, which might be because properties that suffered an almost complete loss were considered to be completely destroyed by the fire, even though a tiny proportion of the property remains. Therefore, when constructing a model for the fire loss proportions, we will assume that no fires will occur in which the proportion of loss is between 0.89194 and 1. The following cumulative density function (Figures 5.7 and 5.8) and probability density function (Figures 5.9 and 5.10) are constructed using the equations we derived above, G(x) for small and medium fires, and the adjustments for fires with a loss proportion of 1.



Cumulative Density Function

$$G(x) = \begin{cases} 6.6554543x & x \le 1/15 \\ 0.18*\log(x/(1-x))+0.65 & 1/15 < x < 0.89194 \\ 0.815 & 0.89194 \le x < 1 \\ 1 & x = 1 \end{cases}$$

Figure 5.7

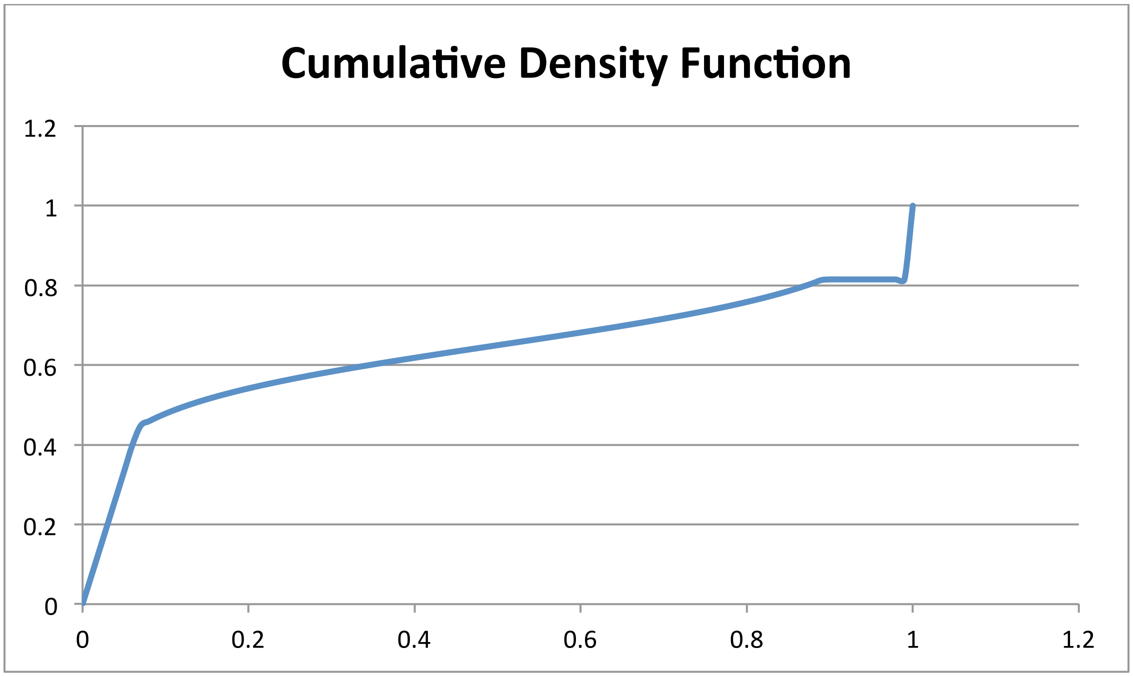
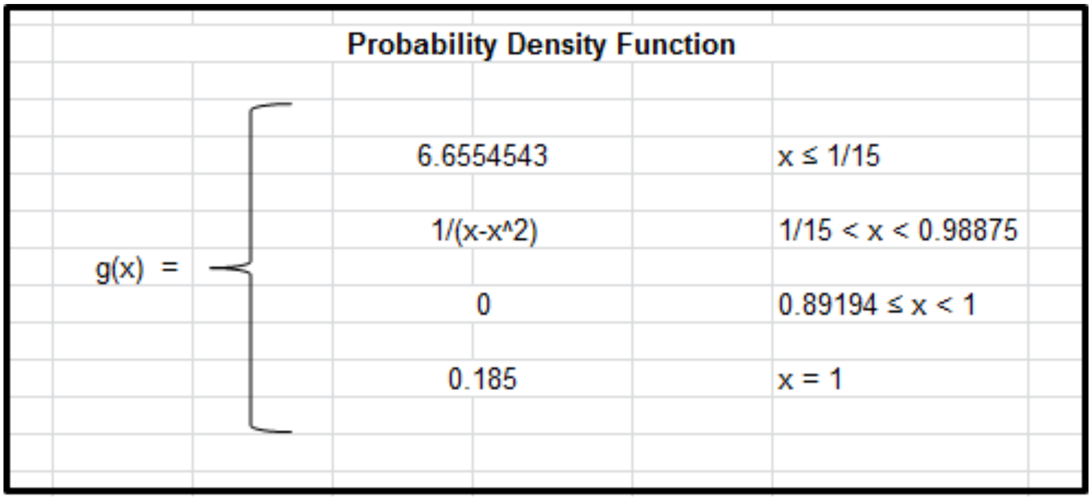


Figure 5.8



Probability Density Function

$$g(x) = \begin{cases} 6.6554543 & x \le 1/15 \\ 1/(x-x^2) & 1/15 < x < 0.98875 \\ 0 & 0.89194 \le x < 1 \\ 0.185 & x = 1 \end{cases}$$

Figure 5.9

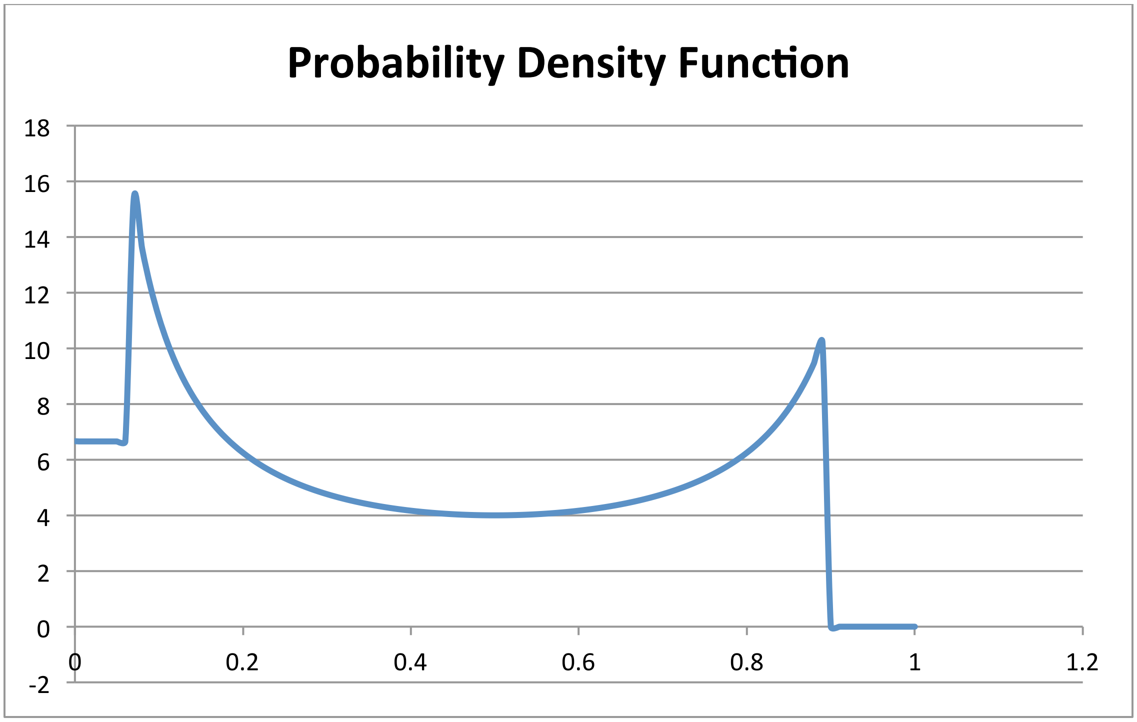


Figure 5.10

The models we have constructed above can be used to analyze the risk of future fires and set proper rates for insurance against such events. However, many factors such as average income and popular awareness of fire hazard may set the city of Berkeley apart from other areas, causing these models to be inaccurate. Nevertheless, these models should reflect a general trend that cities similar to Berkeley may follow.