Colin Quinn

COSC 5610 Data Mining
November 5th, 2020
Professor Guha

# Severe Incidents in Natural Gas Pipelines

Since the 1980's, the natural gas industry has evolved into a major segment of the United States economy. America's dependence on this clean-burning fossil fuel has steadily increased over the last several decades, meaning more gas at higher pressures is being flown through pipeline distribution networks. Severe incidents in natural gas pipelines causes catastrophic supply shortages, significant environmental impact, and potentially loss of life. The pipeline companies responsible for the safe and reliable transmission of natural gas can be held accountable if they do not properly maintain their distribution network. This seems reasonable, however the cost of maintaining a pipeline can be vast and limited resources call for a tool that can identify pipelines prone to severe accidents so companies know which pipe segments need to be repaired immediately.

This document contains an exploratory analysis of the Hazardous Materials Safety Administration (PHMSA) pipeline incident flagged files dataset. Descriptive statistics will be used as measures of central tendency and variability to understand how best the dataset can be used to achieve our goal of reducing the number of severe incidents in natural gas pipelines. An introduction of the dataset is provided in section 1.1. Section 1.2 begins to present the relationships between the variables and severity, and section 1.3 discusses and concludes the final results.

### 1.1 Understanding the Pipeline Incident Flagged Files dataset and Project Objective

The data set used in this project is found on ProPublica but originates from the Pipeline and Hazardous Materials Safety Administration (PHMSA). With the increase in natural gas usage, many governmental regulations and policies have been put in place to help ensure safe and reliable transportation. This data set is built from Title 49 of the Code of Federal Regulations (49 CFR Parts 191, 195), which requires pipeline operators to submit incident reports within 30 days of a pipeline incident or accident. Once a problem occurs in a natural gas network, it is required under penalty of law that both private and public companies fill out a Gas Distribution Incident Report.

The incident report is an 18-page document that asks the pipeline operator to explain the situation in full. Everything from assumed cause of the incident, pipeline installation year, pipe diameter, materials of pipe, and tens of other variables are inspected at the time of the accident. The report is submitted to PHMSA in order to verify if any regulations or policies were violated and for further analysis. Once the incident has been documented, it is added to PHMSA incident reporting database and labeled as SEVERE (1) or NOT SEVERE (0).

This project will use all available information to make an inference on this single descriptive metric the PHMSA uses to define a severe event. A severe incident is likely to cost more money to fix, fatally injure someone, involve an open flame/explosion, or release large abouts of hydrocarbons into the environment. To help a company decide which of their pipe segments is more likely to have a severe incident, a model will be used to makes inferences on such an event in the hopes that the company can preemptively repair or replace the risky pipe.

The objective of this project and intended purpose of the severe event (SE) model is to alert local utility companies or governmental regulators of an impending vulnerability within a pipeline. A vulnerability has the potential to cause an accident, and depending on the nature of the accident, it may be classified as severe or not severe.

A subset of all exogenous variables collected at each event are the inputs to the SE model. The 17 variables used as inputs are listed in Table 1.1.

| **Field Name** | **Data Type** | **Field Name Description** |
|---|---|---|
| FF | CATEGORICAL | Identify if incident was cause by fire first or not |
| GAS_RELEASED | DOUBLE | Estimated volume of gas released in Thousand Cubic Feet (MCF) |
| PIPE_MANUFACTURE_YEAR | DOUBLE | Year of manufacture of Pipe |
| IYEAR | DOUBLE | Year incident occurred, derived from incident date |
| FATALITY_IND | CATEGORICAL | Fatalities (Yes, No) |
| MATERIAL_INVOLVED | CATEGORICAL | Pipe Material involved in Incident |
| INCIDENT_AREA_TYPE | CATEGORICAL | Area of Incident (Aboveground, underground) |
| LOCATION_LATITUDE | CATEGORICAL | Incident Location Latitude |
| LOCATION_LONGITUDE | CATEGORICAL | Incident Location Longitude |
| IGNITE_IND | CATEGORICAL | Commodity ignite (Yes, No) |
| EXPLODE_IND | CATEGORICAL | Commodity explode (Yes, No) |
| EST_COST_PROP_DAMAGE | DOUBLE | Estimated Property Damage - Estimated cost of Operator's property damage & repairs |
| PIPE_DIAMETER | DOUBLE | Nominal diameter of Pipe (in) |

Table 1.1 – List of variables relating to a severe pipeline incident

The dataset was cleaned, removing all observations with missing values. An in-depth analysis of each variable is provided in section 1.2.

## 1.2 Exploratory Analysis of variables

In this section each variable will be inspected and its relationship to the SEVERE variable explained. Conducting an analysis with all variables from Table 1.1, we are able to identify which should be used in a logit-regression model.

Estimated Coefficients:

| | Estimate | SE | tStat | pValue |
|---|---|---|---|---|
| (Intercept) | -351.54 | 4.2443e+07 | -8.2825e-06 | 0.99999 |
| FF_YES | -102.83 | 6.7109e+07 | -1.5323e-06 | 1 |
| GAS_RELEASED | 2.7125e-05 | 8.6346e-05 | 0.31415 | 0.75341 |
| PIPE_MANUFACTURE_YEAR | 0.0027299 | 0.015016 | 0.1818 | 0.85574 |
| FATALITY_IND_YES | 90.066 | 2.5991e+07 | 3.4653e-06 | 1 |
| MATERIAL_INVOLVED_COPPER | -67.867 | 4.2443e+07 | -1.599e-06 | 1 |
| MATERIAL_INVOLVED_DUCTILE IRON | 0 | 0 | NaN | NaN |
| MATERIAL_INVOLVED_OTHER | 33.188 | 7.9404e+07 | 4.1797e-07 | 1 |
| MATERIAL_INVOLVED_PLASTIC | -66.848 | 4.2443e+07 | -1.575e-06 | 1 |
| MATERIAL_INVOLVED_STEEL | -66.777 | 4.2443e+07 | -1.5733e-06 | 1 |
| MATERIAL_INVOLVED_UNKNOWN | 0 | 0 | NaN | NaN |
| IYEAR | 0.20412 | 0.080368 | 2.5398 | 0.011092 |
| INCIDENT_AREA_TYPE_TRANSITION AREA | -100.34 | 6.7109e+07 | -1.4952e-06 | 1 |
| INCIDENT_AREA_TYPE_UNDERGROUND | 0.92621 | 1.2241 | 0.75667 | 0.44925 |
| LOCATION_LATITUDE | -0.036699 | 0.040684 | -0.90205 | 0.36703 |
| LOCATION_LONGITUDE | -0.0042069 | 0.014568 | -0.28877 | 0.77276 |
| IGNITE_IND_YES | 2.2993 | 0.53491 | 4.2984 | 1.7203e-05 |
| EXPLODE_IND_YES | 1.8932 | 0.67188 | 2.8178 | 0.0048353 |
| PIPE_DIAMETER | 0.061178 | 0.064612 | 0.94685 | 0.34371 |
| EST_COST_PROP_DAMAGE | 1.7563e-05 | 6.8229e-06 | 2.5742 | 0.010047 |

157 observations, 139 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 66.2, p-value = 9.45e-08

Figure 1.1 – Initial analysis to determine which features provide a good starting point for a SE model

Figure 1.1 shows only 4 of the 13 variables seem likely to be useful predicting a severe pipeline event. Using a $\alpha = 0.05$, EST_COST_PROP_DAMAGE, EXPLODE_IND, IGNITE_IND, and INCIDENT_AREA_TYPE seem to be significant. You will notice several "1" or "NANs" from the analysis, meaning the dummy variables for MATERIAL_INVOLVED should not be included in the nest model constructed due to faulty data entry (several mislabeled classes). We next take a look at the cross tabulation of the 3 categorical variables in Table 1.2.

| | | SEVERE | |
|---|---|---|---|
| | | NO | YES |
| EXPLODE_IND | NO | 435 | 473 |
| | YES | 53 | 207 |
| IGNITE_IND | NO | 249 | 235 |
| | YES | 239 | 445 |
| INCIDENT_AREA_TYPE | ABOVEGROUND | 237 | 225 |
| | UNDERGROUND | 251 | 455 |

Table 1.2 – Cross Tabulation of 3 significant categorical variables

Table 1.1 shows that if an accident occurs and the gas leaking from the pipe ignites, there is a 53% chance the event will be labeled as severe. This is not an overwhelming proportion, and similar results are seen when comparing the aboveground/underground variable. However, If the accident creates an explosion, there is a 74% chance the event will be labeled significant. This seems logical as an exploration is generally more dangerous than gas ignition. Next, we consider the relationship between severe and the continuous variables.

The initial thought was that there is a relationship between severity and the extremes of an event. The extremes of an event would be high cost of property damage, large amounts of gas being released into the atmosphere, etc…However with the initial analysis of Table 1.1 and Figures 1.1-1.2, we see that these expected relationships are not as clearly defined. In fact, these

seems to be no relationship whatsoever between the two groups of variables, supporting the p-values calculated in Table 1.1.

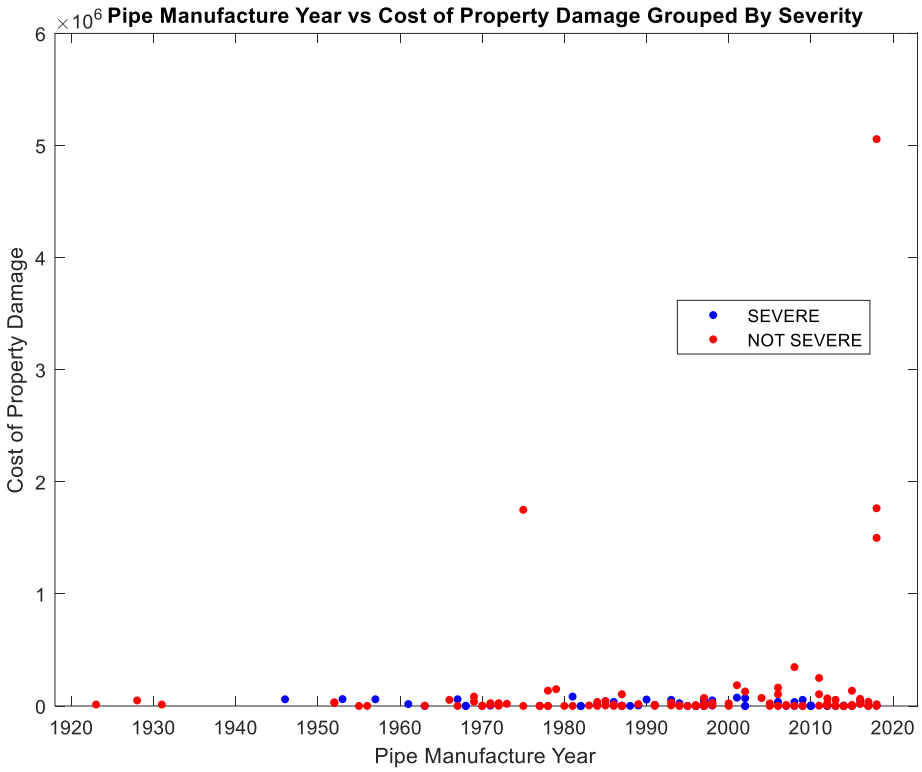


Figure 1.1 -Pipe manufacture year plotted against cost of property damage grouped by severity

Figure 1.1 shows there is no obvious relationship between when the pipe was manufactured and the cost of the property damage. This contradicts the original assumption that older pipes would result in more expensive incidents. Figure 1.2 (below) shows another lack of relationship between pipe diameter and the amount of gas released. Here the original assumption was the larger the pipe, the more gas would be released and severe events occur. However, no distinguishable separation can be made between severe and non-severe events in Figure 1.2.

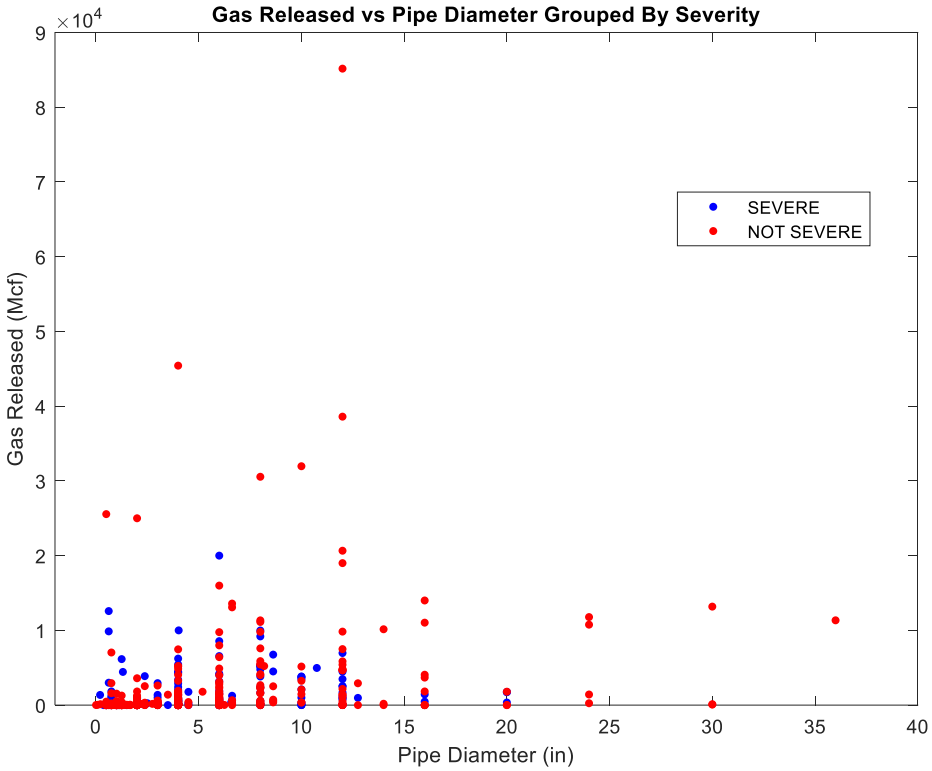


Figure 1.2 -Gas released plotted against pipe diameter grouped by severity

Due to the lack of separation among groups seen in the figures above, we agree with the conclusion from Table 1.1 and that pipe diameter and manufacturer year are not statistically significant in the model. Even when considering the interaction between gas released and pipe diameter, the relationship was not significant. The other continuous variables in the initial model, longitude and latitude, were also included.

The longitude and latitude of the pipe are used in the SE model. One of the common reasons a severe pipeline incident occurs is due to extreme weather events (e.g. hurricane). Areas prone to extreme weather events like in the Gulf of Mexico have historically experienced more severe pipeline incidents than somewhere with little to no weather events. Figure 1.3 shows a heat map of the incident density across the United States.

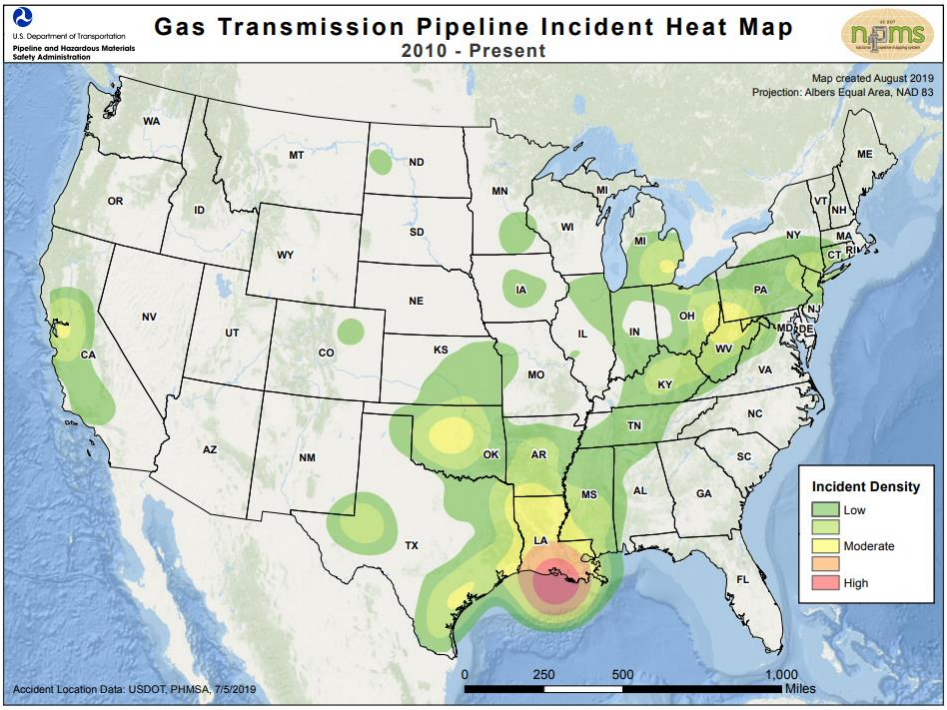


Figure 1.3 – Incident density across the continental United States

Source: [2]

The higher density of incidents in the southern regions did not test significant, however we see the significance for the latitude variable is greater than longitude. This is also inferred from Figure 1.3 as the density of accidents increases from left to right. Using the information gained from this section, a smaller version of the model is presented below in Table 1.3.

Estimated Coefficients:

| | Estimate | SE | tStat | pValue |
|---|---|---|---|---|
| (Intercept) | -160.27 | 40.411 | -3.9661 | 7.307e-05 |
| IYEAR | 0.079466 | 0.02005 | 3.9634 | 7.39e-05 |
| IGNITE_IND_YES | 0.47682 | 0.13619 | 3.5011 | 0.00046337 |
| EXPLODE_IND_YES | 1.7965 | 0.64155 | 2.8002 | 0.0051076 |
| EST_COST_PROP_DAMAGE | -2.205e-09 | 6.9845e-09 | -0.31571 | 0.75223 |
| IGNITE_IND_YES:EXPLODE_IND_YES | -0.67983 | 0.66858 | -1.0168 | 0.30923 |

1163 observations, 1157 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 95.5, p-value = 4.75e-19

Table 1.3 – Reduced model of severe vs non-severe events

These results are further reviewed in the final section, discussion and conclusion.

## 1.3 Discussion and Conclusion

The results from the initial analysis presented above creates several concerns about the dataset. Few of the variables had the expected relationship between a severe and non-severe event. The variables that did test to have a significant effect on the response variable do a poor job explaining the variability within the model ($R^2 = \sim.4$). This exploratory analysis was useful in understanding the dataset, but further work is required to consider this SE model to be effective.

# BIBLIOGRAPHY

[1] "Gas Distribution, Gas Gathering, Gas Transmission, and Liquefied Natural Gas (LNG) Incidents," Pipeline and Hazardous Materials Safety Administration (PHMSA), 2014.

[2] "Gas Transmission Pipeline Incident Heat Map," *U.S Department of Transportation*, 2019. https://www.npms.phmsa.dot.gov/Documents/NPMS_HeatMap_GTIncidents.pdf (accessed Apr. 11, 2020).