COMBINING ENGINEERING AND DATA-DRIVEN APPROACHES TO MODEL THE RISK OF EXCAVATION DAMAGE TO UNDERGROUND NATURAL GAS FACILITIES

by

HAMZA ABUSNINA

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Civil and Environmental Engineering

Written under the direction of

Dr. Jie Gong

And approved by

____

____

____

____

New Brunswick, New Jersey

OCTOBER 2019

ABSTRACT OF THE DISSERTATION

COMBINING ENGINEERING AND DATA-DRIVEN APPROACHES TO MODEL THE RISK OF EXCAVATION DAMAGE TO UNDERGROUND NATURAL GAS FACILITIES

By HAMZA ABUSNINA

Dissertation Director:

Dr. Jie Gong

In the United States there are thousands of gas pipe miles, long grids, and networks of natural gas lines across the states. Recent pipeline leaks and explosions in various regions have driven the industry to re-evaluate on-going efforts aimed at aggressive pursuit of preventive strategies. Considering that safety and environmental risk is a major issue, particularly in cases where underground gas line damages and other explosions are involved, pipeline accidental risk represents both financial and social interests in the gas pipeline industry.

It is possible to, knowingly or unknowingly, damage underground gas services, water services, electrical services, etc. Incidents involving infrastructure damage are far more common than perceived; and these incidents result in hundreds-of-thousands, if not millions, of dollars in repair or replacement. Damages to underground facilities may occur by large construction contractors or by homeowners.

The main objective of this research is two-fold: a) to determine the important risk factors contributing to the underground gas pipe damages; b) to identify inputs required for an effective evaluation and assessment of the risk encountered in exchange of information between different parties involved during the repair of underground gas pipelines. Predictive Model will be developed based on machine learning algorithms (Logistic Regression) to be used in predicting the important risk factors affecting the underground Gas Pipe Damages.

The research will systematically analyze the risk of underground gas pipeline network damage including; process the data collected from agency, organize/classify the data based on certain parameters, process the data, develop integrated risk model and influence diagram. Next, Bayesian Network will be developed based on the derived important factors, and calculated probabilities for each attribute.

Acknowledgements

The years at Rutgers as a Ph.D. student have been an amazing and thrilling part of my life. I have had the privilege and pleasure of being supervised by Dr. Jie Gong. I am deeply grateful for his continuous guidance, support, and encouragement throughout my entire graduate experience at Rutgers. I learned a lot from his professionalism. It is really a great honor for me to have him as my mentor and a friend.

I would like to thank my friends and colleagues starting with Mr. Aravind Reddy Pasham who provided very valuable help for this dissertation. I would also like to thank Mr. Sher Khan, Mr. Kall Dalco for their support throughout my studies. I have felt so blessed to be surrounded by such great friends who made this journey enjoyable.

I would like to extend my appreciation to all my colleagues who provided insightful; discussions, and scientific suggestions to help improve my dissertation.

Mostly, I would like to thank my loving parents and my family members for their immense love, support, and confidence in me.

Table of Contents

Table of Figures

CHAPTER ONE: INTRODUCTION

1.1 Research Motivation and Problem Statement

Currently, in the United States there are thousands of gas pipe miles: long grids and networks of natural gas lines across the states. Recent pipeline leaks and explosions in various regions of the US have driven the industry to re-evaluate on-going efforts aimed at aggressive pursuit of preventive strategies. Considering that safety and environmental risk is a major issue, particularly in cases where underground gas line damages and other explosions are involved, pipeline accidental risk represents both financial and social interest in the gas pipeline industry.

This study compares the failure data from various pipelines to investigate the trend for rates of failure, causes of failure, aging characteristics and relationship between the causes of damage and pipeline parameters. As the construction field continues to expand, it is important to focus on maintaining a high level of safety in order to protect occupational hazard and ensure public safety. Damages relating to excavation practices and procedures directly impact public safety due to the nature of the infrastructure system in place. It is possible to, knowingly or unknowingly, damage underground gas services, water services, electrical services, etc. Incidents involving infrastructure damage are far more common than perceived; and these incidents result in

Figure 1 Aftermath of Harlem Gas Explosion

hundreds-of-thousands, if not millions, of dollars in repair or replacement. Damages to underground facilities may occur by large construction contractors or by homeowners.

The importance of protecting underground utilities is evident, but necessary precautions less known. Contractors and homeowners often disregard or unknowingly excavate with imminent danger in the subsurface. Out of all the underground facilities, natural gas lines pose a major threat to public health and wellbeing. Even when damages are not caused by excavation, gas pipelines may have a lasting impact if not repaired or checked upon. On March 12, 2014 two apartment buildings in Harlem, New York City were devastated by a natural gas explosion (pictured in Figure 1) which resulted in deaths of eight people (Sanchez 2014). While this incident was not directly related to excavation, it reflected on the severity of the impact of natural gas explosions. When excavating into the Earth's surface, it is imperative to know if there is any gas pipeline lies beneath and if so, where does it exactly positioned. Without such knowledge, devastating incidents, similar to the one shown in Figure 1, may happen in the construction field. Therefore, it is critical to have background data that assists in determining causes of explosion. Such documented information may also help creating well-defined laws and

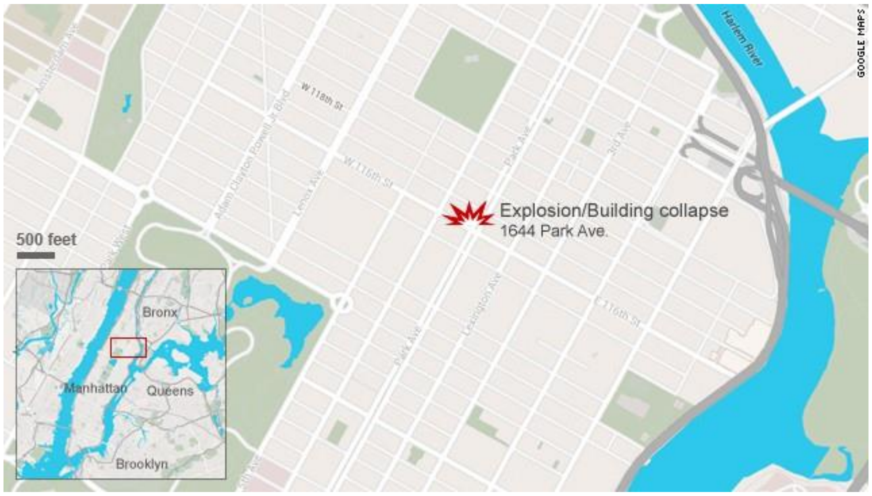


Figure 2 Location of Harlem Gas Explosion

regulations to guide the general public during excavation procedures. For quite some time now, the Public Utilities has been collecting data that accounts for most incidents related to underground utility infrastructure. Tracking all infrastructure damages and persons responsible is a difficult task due to the nature of the construction industry. Records of construction work aid the book keeping process; however, sometimes it is difficult to determine utility damage instantaneously.

Thus, some of the records become improperly tracked and recorded. Nonetheless, the recorded information greatly inform research and data analysis to determine causes of damage and trends. For excavation jobs deeper than 18 inches, it is required to call or fill a form online 3-10 business days prior to job commencement.

In another gas line damage incident, a contractor was hired by the city to replace sidewalk and curbing dug into an unmarked natural gas service line with a backhoe Figure 2. Although the service line did not leak where it was struck, the contact resulted in a break in the line inside the basement of 1816 West 3rd Street, where gas began to accumulate figure 3.

A manager for the contractor said that he did not smell gas and therefore did not believe there was any imminent danger. The manager called an employee of the gas company and left a voice mail. At approximately 1:44 p.m., an explosion destroyed two residences and damaged two others to such extent that those houses had to be demolished. Other nearby

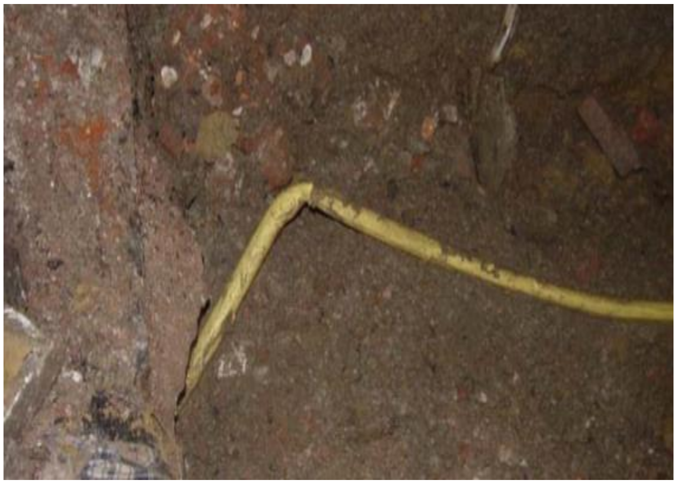


Figure 3 Underground Gas Pipe Explosion NTSB

residences also sustained some damages: the residents on that city block were displaced from their homes for about a week. Three contractor employees sustained serious injuries. Eleven additional people sustained minor injuries (NTSB 2003).

Another Tech Consultants was awarded similar work to be performed at1820 West 3rd Street figure 4. The project manager surveyed the work site and determined that sidewalk and curbing replacement was needed in front of the residences at 1816, 1818, and 1820 West 3rd Street. Then, Tech Consultants showing the work to be done in front of the three addresses shown in Figure 5. On June 23, 2003, Tech Consultants issued a change order to the contractor, which included the 1820 West 3rd Street address location in a list of additional address locations. Tech Consultants did not give the 1816 and 1818 addresses or sketch to the contractor. The underground utility by mistake marked for 820 West 3rd.

Therefore, the damage to the gas line happened. The primary reason the explosion was a miscommunication between the contractor and the consultants. In addition, the wrong marking was part of the problem. Also, the failure of the Tech Consultants to verify that all underground facilities were marked within the proposed dig site before beginning excavation. As responses to the gas leak, the police dispatcher received numerous reports of an explosion on West 3rd Street. The police department responded to the site by evacuating residents, conducting crowd and traffic control. Also, the fire department initially dispatched two engine companies

Based on the later developed best practice, the city stated that the excavators should notify the pipeline operator immediately if their work damages a pipeline and to call 911 or another local emergency (NTSB 2003).

As can be seen from the gas line explosion above and missed mark out, there is a communication gap s between the involved parties in the excavation process. Due to the lack of transfer of the right information to the right party at the right time, the gas line damage happened. Even though best practices proposed was in place after the accident, it did not solve the main

problem of communication gap.

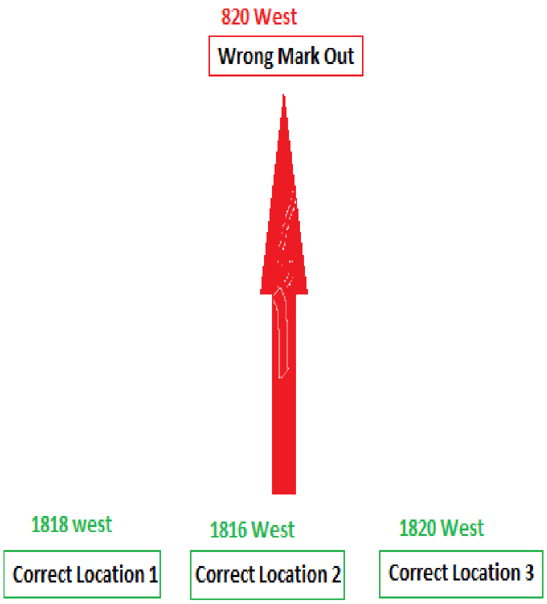


Figure 5 : Locations of Mark out Locations

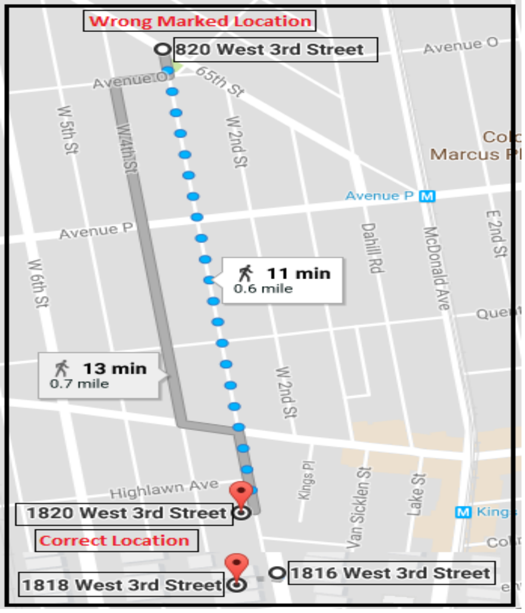


Figure 4: Locations of Mark out Locations & street #

In another case of gas line damage, a backhoe was digging a trench behind a building; then the backhoe operator damaged a ¾-inch steel natural gas service line: as shown in Figure 6. This resulted in two leaks in the natural gas service line, which was operated at 35

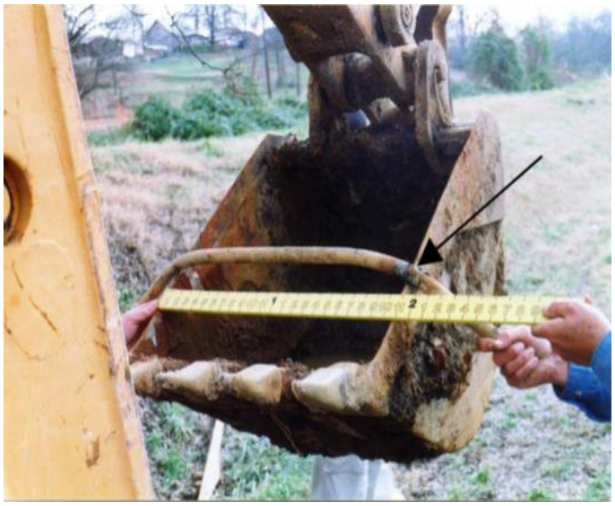


Figure 6: Damaged section of the gas service line NTSB

psig. One leak occurred where the backhoe bucket had contacted and pulled the natural gas service line: shown in Figure 6.

The other one was a physical separation of the gas service line at an underground joint near the meter, which was close to the building. Gas migrated into the building, where it ignited at about 10:02 a.m. An explosion followed, destroying three buildings as shown in Figure 7.

Other buildings within a two-block area of the explosion sustained significant damages. This accident was resulted in three fatalities, five serious injuries, and one minor injury.

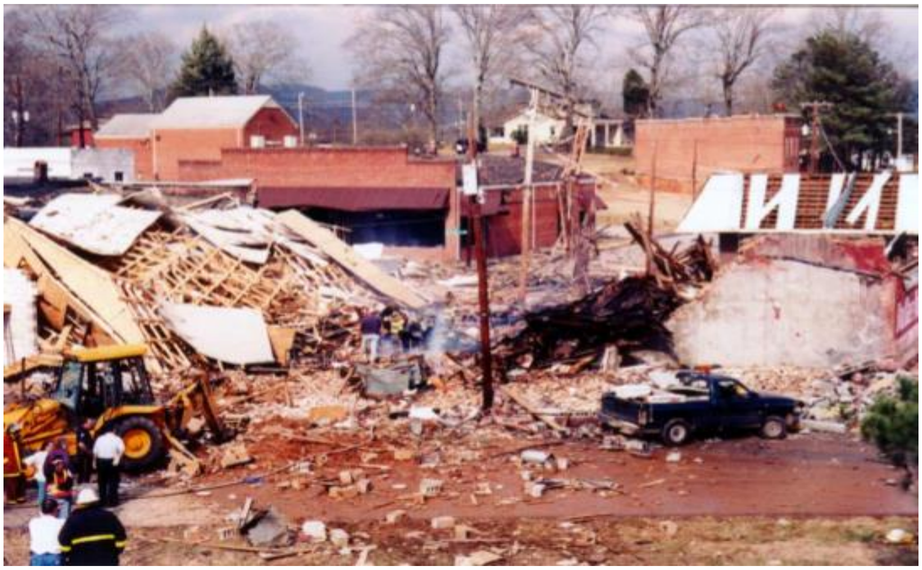


Figure 7: Damage to Buildings (NTSB)

The contractor located and marked the gas and water service lines for the trenching since the accident. The Line Location Center has been predestinated. The contractor told investigators that blue paint was used to mark both service lines because that was the only paint that they had. However, the representative later could not find any blue or other line markings on the ground at the accident scene.

The contractor and the acting supervisor left the excavation site at about 8:15 a.m. to go to the utility shop. They told investigators that, before they left, they had asked the owner to watch the backhoe operator. The backhoe operator arrived at the excavation site sometime after 8:15 a.m. While digging the trench,

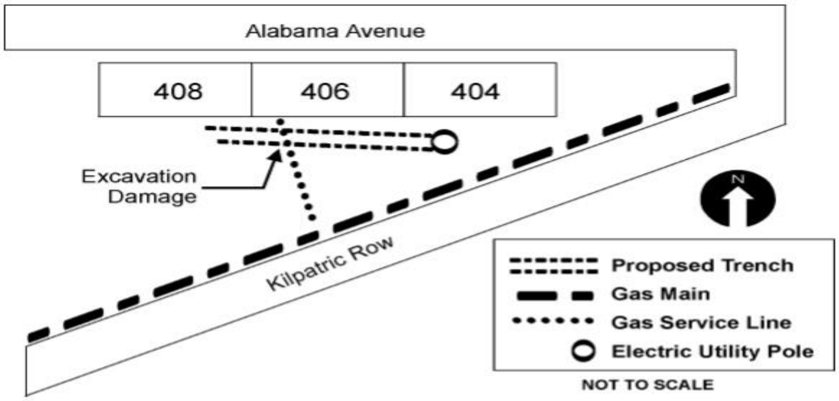


Figure 8: Schematic of Accident Area (NTSB)

The backhoe operator damaged the underground gas and water service lines, resulting in leaks in the water and gas service lines. Figure 8 shows a schematic of the relative position of these elements. The probable cause of the accident was the failure of the contractor to establish and follow safety procedures for excavation activities, resulting in damage to a ¾-inch natural gas service line, and the failure of the Utility Company to provide appropriate emergency response to the resulting natural gas leak. There is missing communication information flow procedures in place to ensure that the contractor is following the right guidelines.

Let us look at another gas line accident case where a 20-inch-diameter steel natural gas transmission pipeline were ruptured and released natural gas near an intersection. The gas ignited and burned; as a result, one resident was killed and another person was injured. About 75 residents required temporary shelter. Six homes were destroyed Figure

9. The contractor called for the 8-inch distribution main to be installed parallel to the 20-inch pipeline with a horizontal center-to-center separation of approximately 5 feet. Before the drilling began, marking out location of the 20-inch pipeline at intervals of 15 to 20 feet was performed. According to the drilling personnel, field measurements indicated that they could maintain an approximate 5- to 7-foot horizontal separation between the new installation and the paint marks: used to indicate the location of the existing 20-inch transmission pipeline. The drilling crew intended to maintain this separation throughout the bore except for one location near the termination of the bore. There, in order to avoid an underground telephone duct, the pipeline separation would need to be reduced to an edge-to-edge horizontal distance of about 1 foot. The drilling area expanded to go out of the mark out boundaries. In addition, there were remaining tools used in the backfill which too had caused some damages.

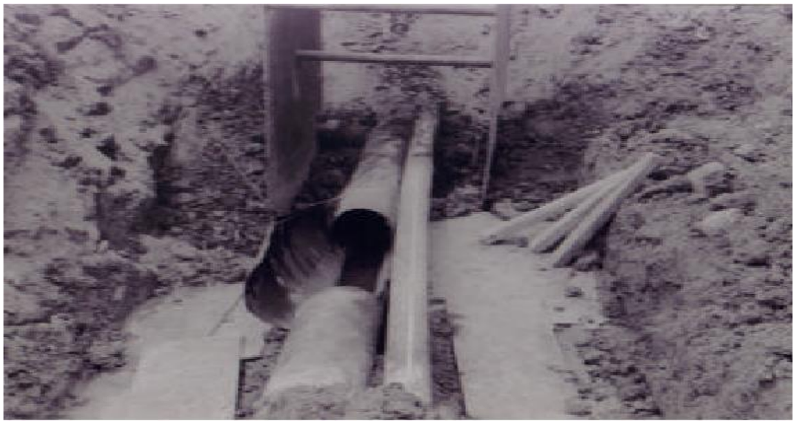


Figure 9: Post accident excavation revealing relative positions (NTSB)

The main cause of this accident was the failure of the utility company & excavator to have adequate controls in place. Control measures needed to ensure that the directional drilling operations carried out in the proximity of existing underground facilities would not cause damage to those facilities. This means digging with the boundaries of the mark out. Later by the agency’s best practices, procedures were put in place to make sure directional drilling is following the national safety procedures. Note the existing gap on the digging processes that misses a step to mandate the excavator to request another mark in case he needs to excavate outside the beaneries. The excavator also need to communicate that information to the utility company to ensure optimum information flow.

While the reasons of underground pipeline damages can be obvious, the underground pipe damage remains hard to predict. It is also unclear that what risks are involved in any pipeline repair or maintenance which required excavation process to fix the damage. Moreover, it was reported by DIRT interactive analysis that in some states damages caused 1,361 service drops among all different customers in different fields, the majority of damages happened to natural gas pipeline which was 1,175 in year of 2015 Figure 10.

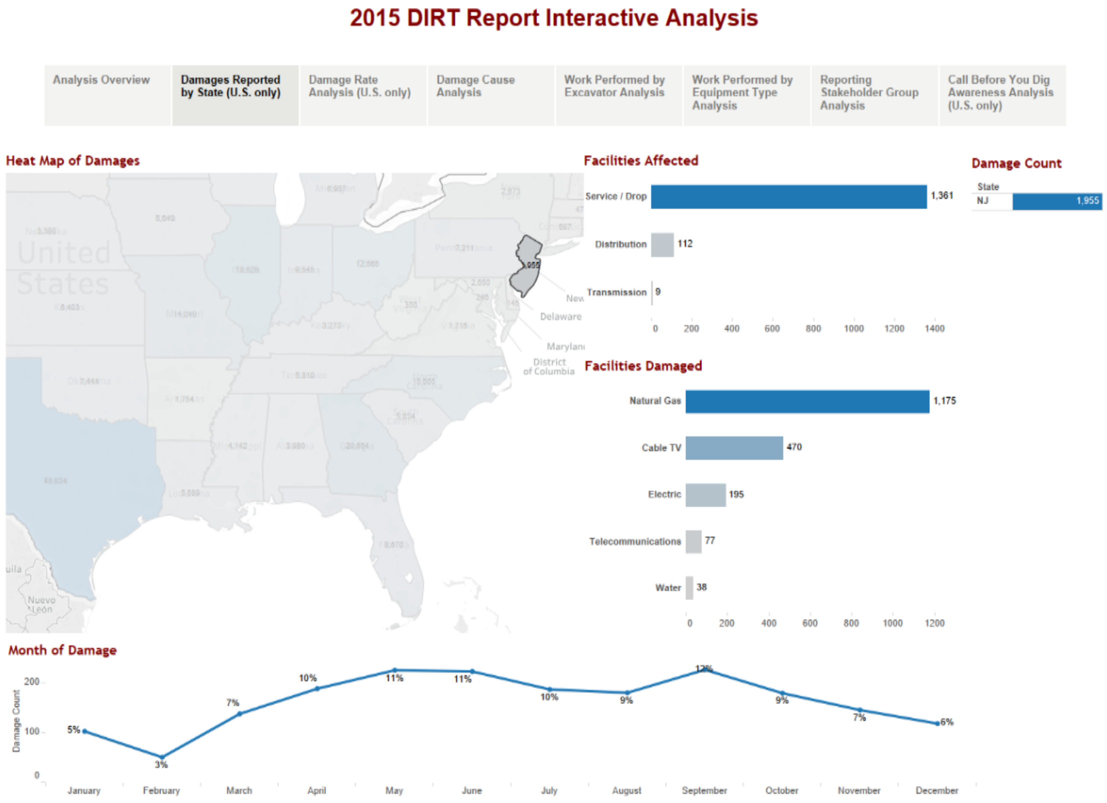


Figure 10:2015 DIRT report by Interactive Analysis by CGA

By clustering and gathering all data available which contributed to the risk of damage were happening as probability. As reported in Figure 10 by Common Ground Alliance CGA, 45% of the damage happened because of unsafe excavation practices, 18% of pipeline damages happened because of insufficient locating practices including mark out process. However, 31% contributed damage because of the excavator did not call the one call canter and raise major concern about the risk involved.

In other words, the causes of pipeline accident fall into many broad categories. Figure 11 below show the number and percentage of significant insufficient excavation practice and insufficient locating practices accidents attributable to different cause categories during 2015.

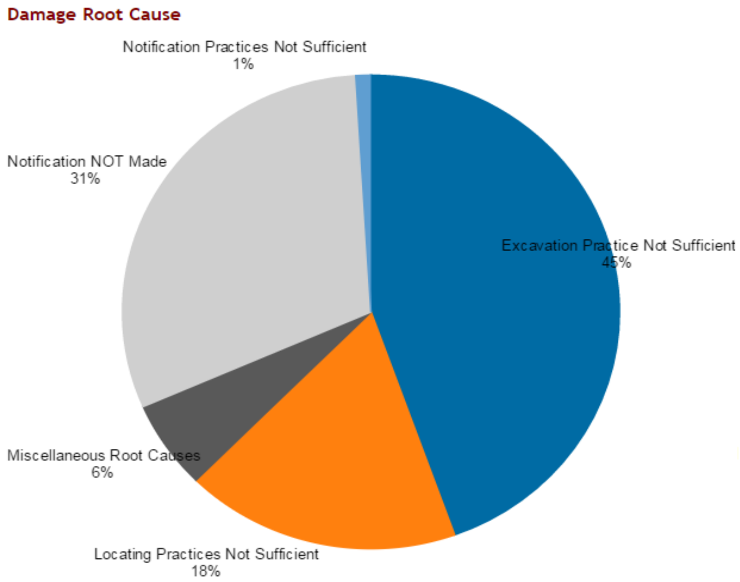


Figure 11 : 2015 DIRT Damage Root Cause in the U.S by CGA

Due to a large number of underground pipeline incidents in utilities sites, significant research has been already conducted to understand the problem. The reviews of underground pipeline damage suggest increased accuracy of the mark out location to prevent the underground pipe being hit by excavator, change excavation practice itself in terms of the operator, excavator, and predicting pipe depth. Other areas, including the notification practices were insufficient within the one Call Center.

Natural gas line infrastructure is a very complex network, integrated with other networks such as water, fuel, and sewer. In current research studies, natural gas pipeline accident is established, based on which the probabilities of evolution stages and consequences of natural gas pipeline network accident can be estimated and analyzed (Wu, Zhou, Xu, & Wu, 2017). Gas pipeline accident in United States, Europe, Canada, and other countries and regions shows that the main reason of accidents are external causes (Yang, Hao, & Xing, 2013).

The decade from 2001–2010 saw a total of 544 major excavation related damages resulting in 37 fatalities, 152 injuries, and close to $200 million in property damage. Lack of accurate position and semantic data of buried utilities coupled with absence of persistent visual guidance are two key problems facing excavator operators (Talmaki & Kamat, 2012). Thus, there are uncertainties for predicting the damages throughout the all excavation practice, starting from initiating the request by the excavator, passing through creating the ticket the One-Call center systems, distributing the information to all utility companies in the particular location, locating/ mark out, sending the information back to the excavator by agency and notify the contractor to start excavation till completion successfully.

1.2 Research Gap

The underground pipeline line damage is a severe wide spread problem. The current studies focus mostly on the excavation and locating process (Talmaki & Kamat, 2012). Furthermore, the existing studies fail to identify the risks involved in the information flow process for ticket damage initiation. The impact of the natural gas line damage can be pervasive and may affect many sensitive sectors that include, but not limited to, hospitals, schools, transportation, and utility companies.

To solve the current problem and eliminate the risk involved, it is important to identify all factors and inputs associated with the causes of the damage. In addition, all the parties needed to be involved in the information flow starting with initiating a request by the excavator, passing through creating the ticket the One-Call center systems, distributing the information to all utility companies in the particular location, locating/ mark out, sending the information back to the excavator by One-call center and notify him to start digging. However, the existing literature did little to = to systematically identify all the elements associated with the damage. This study seeks to identify and further understand the key risk factors that may weaken the different nodes/steps involved in the gas pipe damage process as per Bayesian network. However, Process flow, including excavation and locating, is a very complex process that uses variety of different technologies and a large number of apparatus and equipment where every input is very crucial. (Makowski & Mannan, 2009; Jaw & Hashim, 2013; Dong and Yu, 2005; Jamshidi, Yazdani-Chamzini, Yakhchali, & Khaleghi, 2013).

Risk identification techniques vary according to variables involved in the damage process. While multiple studies focused on how to respond to underground gas pipeline damages but few paid attentions to preventive measures. Such preventive techniques may include more public awareness, more rigorous inspections and prospective analysis of the pipeline dangers. From the Special Report 281 on transmission pipelines and land use (2004), Gas Pipeline companies have also begun utilizing a variety of risk assessment techniques, for instance: scenario-based analysis, fault tree analysis, indexing methods. However, very few gas pipeline companies had adopted a data driven risk assessment approach to proactively mitigate excavation damage risk. Most analyses center around specific factors that could affect the possibility of gas pipeline damage (e.g., wrong Location, off mark out, excavator mistake, digging distance) but not around the consequences of missing/wrong information between different parties. For example, underground operator may not check all utilities exist in the area of the gas line damage, hit during the digging because of the wrong depth given to the excavator.

Even though some of these risk assessments tried to take component interdependencies into account, others focused on specific pipeline system components. As per Muhlbauer (2004), the pipeline risk management and assessment techniques that exist in the current literature involve various methodologies to obtain the probabilities and consequences of processes and events leading to risk. The focus on calculating of a risk probability involved in the gas line damage (e.g., a mathematical product of probability and consequence) are common among these efforts. Even though this calculation provides a quantitative assessment using several components of the gas pipeline damage, it does not consider the all factors involved in the evolution process.

However, it does not pose significant risk (Gas Pipelines and Land Use, 2004) because it does not cover all the parties involved (e.g., initiating ticket by excavator, Underground operator check the utilities in the area, mark out, and safely excavate to fix the damage or replace the gas pipe). More specifically, risk model needed to be developed to solve the problem of the probability the damage occurring because of certain unknown factors. In addition, during emergency requests, the one Call Center takes only two hours to dispatch underground utility person to identify the location and make mark out, which means greater opportunity of risk because of shortened processes.

Another aspect in the process is the risk assessment model. Risk assessment models need to take into consideration all gas line damage data that can be obtained. In addition, risk data analysis must be done with great care to ensure an effective risk assessment model. Moreover, the relationship among all risk factors must be identified; actual data obtained must be validated. Gas pipeline damage is a complex process which depends on a number of factors, including the line pipe characteristics, maintenance policies, underground operator and actions of the excavator. Even though, a great deal of information is known concerning pipeline history and physical processes, there is a lack of information to reliably predict the occurrence of gas line damage under all possible conditions. More specifically, the lack of data regarding the physical conditions of the gas distribution line (e.g., executed depth, location, and mark out) and processes contributed to the damage ticket request information flow. Consequently, substantial uncertainties are associated with the gas line risk frequency predictions. More accurate prediction requires risk model aligned with ticket request processes, policies, and procedures to make precise risk assessment of all risk uncertain parts.

As can be seen above, any small fraction risk occurrence could have severe impact on the gas network, all these gas grids located at various urban cities are exposed to damage due to a number of involved risk factors and parameters. Damages occurring from pipeline accidents bring about a widespread set of consequences. They are typically not restricted to fatalities or human injuries, but also encompass environmental damage caused by fires: and large financial losses due to supply interruptions to customers such as schools, Hospitals, Factories, ports, and other public sectors. It is important to define what is risk incident occurrence rates and probability. Most of these damages occur during construction tasks that does not involve work directly on pipeline systems. Additionally, significant incidents are primarily caused by excavation damage with about 22.5% of all incidents (West 2013). PHMSA defines an incident as a release of gas, liquefied natural gas (LNG), liquefied petroleum gas (LPG), refrigerant gas, or gas from LNG facility that results in either:

1) Death or personal injury that requires in-patient hospitalization

2) Estimated property damage of $50,000 or more, including loss to operator and others, but excluding the cost of gas lost

3) Unintentional estimated gas loss of three million cubic feet or more

Thus, the lack of efficient risk model and defined set of data parameters could lead to misclassifying risk categories involved in the different processes of underground gas pipe request damages. On the other hand, parameters needed to be identified carefully to cover all four parties involved: excavator, One Call center, Underground operator, and Locating to reach better assessment of predicting weak nodes on the information flow diagram of gas pipe request damage.

Gas Pipe Damage Request Information flow process: There are many uncertainties involved during exchanging information. This starts when the excavator requests the service from the agency, then the agency initiates the ticket request and share information with excavator, then the center pass the request to Underground operator which will check which utilities are located within the parameters of the site. After determining all utilities within area, the UG operator contact these utility companies and inform them of the request, then the utilities reply back with certain information about the Gas line location, depth, name. In the next step, the underground operator send person to make the mark outs, then the agency notifies the excavator that he can start excavating within certain period depending on some condition. Finally, the excavation starts excavation. As can be seen, each and every node in these processes has potential of miscommunication of the data which consequently will increase probability of damage occurrence. On the other hand, the lack of coordination throughout the process and prioritizing which of the nodes are more critical makes risk development model to assess the damage very complicated with uncertain parameters involved.

1.3 Research Questions

Regardless of the policies, procedures, and safety measures developed by governing agencies, underground gas lines are still at risk of damage due to excavation, mark out, locating, miscommunicating data between different parties involved in the process. The goal of this research is to develop, process, and organize the available data to efficiently identify risk involved and vulnerable weak nodes on information flow process. The research will systematically analyze the risk of underground gas pipeline network damages: including process the data collected from agency, organize/classify the data based on certain parameters, process the data, develop integrated risk model and influence diagram. More specifically, the goals of this research are the following:

1. Determine the dominant risk factors, and inputs required for an effective evaluation and assessment of the risk encountered in exchange of information between different parties involved during the repair of underground gas pipelines.
2. Design risk predictive model by studying the past underground gas line damages in urban congested cities.
3. Provide research base by using Bayesian Theory to develop risk model to investigate the interactive effects of various factors causing underground gas line damage, and predicting the probability of future damage occurrence.

To realize such research objectives, this research must address the following questions;

1. What are the dominating factors contributing to excavation damages to underground natural gas facilities?

2. How do these dominant factors relate to each other and form up a network of interacting factors triggering different excavation fates?

3. Can these dominant factors serve as leading indicators to predict the risk of future excavation requests in terms of damaging underground natural gas facilities?

1.4 Research Methodology

To address the aforementioned research challenges, the proposed group of processes to develop a solution to the research questions as can be seen in Figure 12. The thin solid arrow in the figure indicates the flow and the sequence of different stages; the thick solid indicate the different stages of data analysis and process through the predictive model. The chart starts by building information flow process which can be used by 811 call. Then the process flow map was integrated with the literature review to develop combined information flow process and to determine research gap. Data assessment conducted throughout the chart is shown below to reassess different process involved in damage of the underground gas pipeline. Followed by e-data organization and selection of the attributes among four parties involved in the ticket request of gas pipe damage, the next step is developing influence diagram cause effect relation within each party itself. Finally, integrated influence diagram is developed for all four parties involved in the ticket request process.

The next step is to propose Bayesian Theory as a possible leading theory for developing risk models. Bayesian network was developed in stages and 3 phases. As shown on the chart, the first phase contains the causes of damage and damage area (Urban City). Followed by, phase 2 of Bayesian network contains two important factors which influence the information flow regarding the underground pipeline damage which is noted by excavator and location. The third phase is a combination of many factor which may lead to a cause or near miss regarding underground gas line damage.

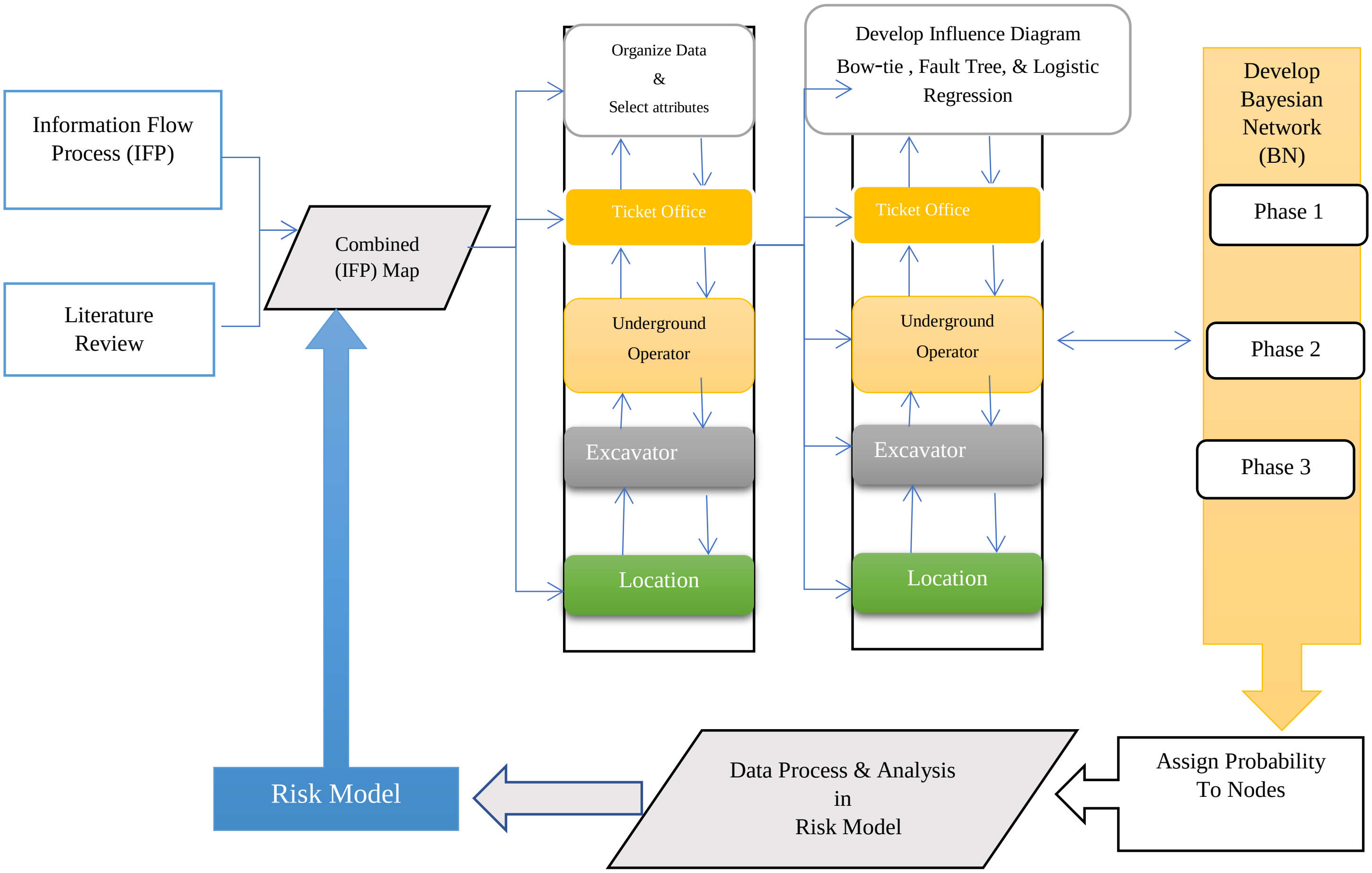

Information Flow Process (IFP)
Literature Review
Combined (IFP) Map
Organize Data & Select attributes
Ticket Office
Underground Operator
Excavator
Location
Develop Influence Diagram Bow-tie , Fault Tree, & Logistic Regression
Ticket Office
Underground Operator
Excavator
Location
Develop Bayesian Network (BN)
Phase 1
Phase 2
Phase 3
Assign Probability To Nodes
Data Process & Analysis in Risk Model
Risk Model

Figure 12 Overview of Components of proposed Framework in Assessing The Risk in Underground Gas Line

Finally, the Bayesian Network will be input into the Agena software after determining the probability. Those data will be processed and analyzed using the Agena Software.

A Comprehensive testing method is proposed to validate the proposed system framework. The data were collected from year 2010-2014. The data from 2010-2013 were processed in Bayesian network through Agena Software; however, the data from

2014 will be used in validating and testing the current proposed model in this research. All collected data represent different urban areas and different municipals to have wide variety and range of collective results.

1.5 Challenges & Research Contribution

1.5.1 Research Challenges

The proposed framework addresses five challenges related to risk assessment for underground gas line damage during and after the damage happened.

Challenge 1: Data preparation**;** the data used was recorded and stored in repositories that are disorganized and therefore render the data as unworkable. Many steps were performed, which are determined by the researcher or analyst, must be taken to convert the raw data into "clean" and workable information. Of all the possible preprocessing steps, a select few are very important for making the database computational friendly. The purpose is to improve risk assessment for underground gas line damage with respect to time, cost, and quality of the results. In this research, a select few important data preprocessing techniques were used. The information flow process and attribute selection criteria which will later be selected to govern the risk model.

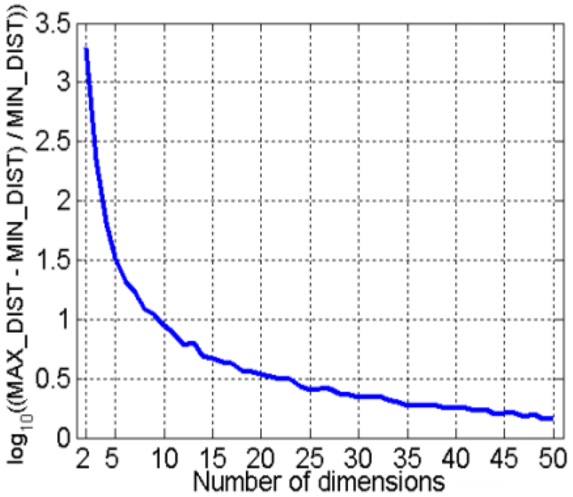


Figure 13:Curse of Dimensionality

Challenge 2: Dimensionality Reduction; **Dimensionality reduction is** one of the most important preprocessing techniques to consider. As technology expands, the chances of running into dimensional problems increase. Many of these big data sets contain hundreds of attributes with varying information which may or may not be useful for meeting the goals of a project. Dimensionality reduction provides the following benefits:

1) Risk assessment precision increased when there are fewer attributes
2) Risk model for one call data can be better visualized
3) Models become less complex and more understandable when the number of attributes decreases.

The goal of dimensionality reduction is to reduce the number of attributes that provide little information. By doing this, it eliminates the possibility of running into the curse of dimensionality. The curse of dimensionality refers to the phenomenon when data analyses become significantly harder as the dimensionality of the data increases (Tan et al. 2006). When applied to clustering, the curse of dimensionality impacts distance calculations between points in data and therefore become less meaningful. Figure 13 shows how the significance of the distance calculation reduces as the number of dimensions, or attributes, increases.

Challenge 3*:* Obtaining the Critical Nodes and Risk Probability**;** data quality affects the probability percentage with direct impact. Precise risk probability in Bayesian network is common when going through raw data files. Data repositories store the raw data that is provided and therefore contain many defects. These defects pertain to noise, outliers, missing values, and duplicate data. Thus, data cross referencing been conducted to verify and validate the real date used in the research which is input to the Bayesian risk model. In addition, the data received were millions in number. Thus, it was necessary to select certain attributes which have direct impact on the underground pipeline damage Figure 14.

Challenge 4: Lack of risk Modeling of Processing, Information Flow Process, Computing Risk Modeling; one of the significant challenges for using one call center data is figuring out the information flow process and integrating that with the risk model. The purpose is to assess the critical elements involved in causing the underground pipeline damage. This process requires a closing loop between ticket request flow mapping and probability of the risk involved in each Bayesian network node. On the other hand, from data processing perspective, it urges a clarified, well-defined goal, which they can convert to a series of feasible computation tasks. Currently, there is a

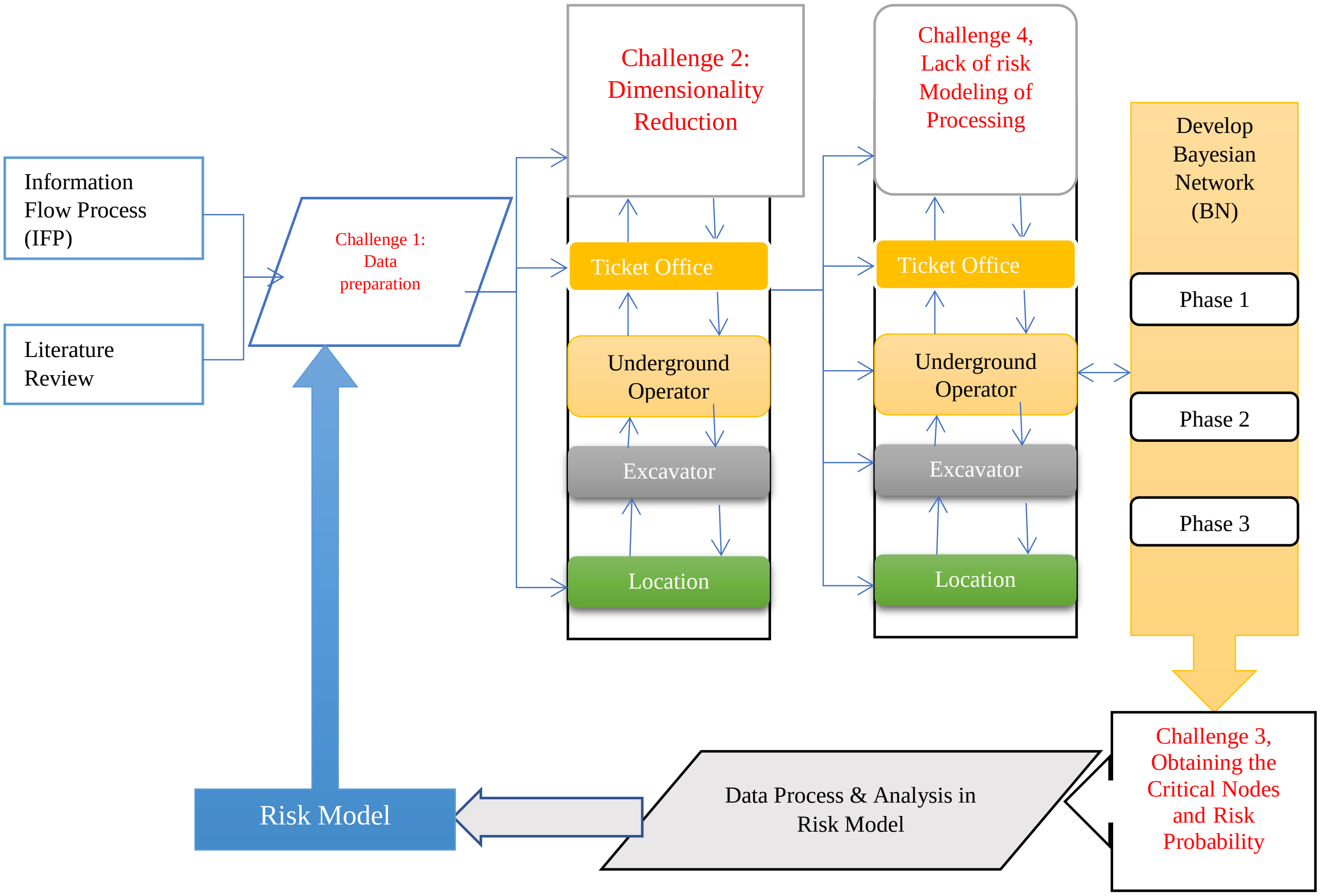

Information Flow Process (IFP)
Literature Review
Challenge 1: Data preparation
Challenge 2: Dimensionality Reduction
Ticket Office
Underground Operator
Excavator
Location
Challenge 4, Lack of risk Modeling of Processing
Ticket Office
Underground Operator
Excavator
Location
Develop Bayesian Network (BN)
Phase 1
Phase 2
Phase 3
Challenge 3, Obtaining the Critical Nodes and Risk Probability
Data Process & Analysis in Risk Model
Risk Model

Figure 14:A Process Map of Faced Challenges During Modeling The Risk in Underground Gas Line

huge shaded area between this between ticket request flow mapping and probability of the risk involved in each Bayesian network: there is lacking of formal risk modeling of information flow processing. In addition, the four parties involved in the damage, one call office, underground operator, excavator, Location Company have wide range to risk probability involved which is difficult to assess unless there is risk flow process to follow.

1.5.2 Research Contribution

This research aimed at using the predictive model tools to describe the interaction between the significant factors required for an effective evaluation and assessment of the risk encountered in natural gas underground damage. This analysis characterized natural gas pipeline failure rates using pipelines obtained data based on the need to minimize pipeline damage rates and factors associated with natural gas damage. A Bayesian risk model was developed to minimize pipeline damage rates through assessing and ranking the risk of various sections of natural gas pipelines. It would explore the interaction among significant factors, and inputs required for an effective evaluation of the risk encountered in exchange of information between different parties involved. In addition, the past data were used to develop a risk model to study future risk associated with the excavation requests and risk factors. Provide a research base by using Logistic Regression to develop risk model to investigate the interactive effects of various factors causing underground gas line damage, and predicting the probability of future damage occurrence. The study assesses and predicts the risk involved in the locating request during the normal process and emergency process in congested cities. It improves the exchange of the data between different parties, which will consequently mitigate the probability of the risk occurrence.

CHAPTER TWO: LITERATURE REVIEW

1.6 Prior studies in underground pipeline damages.

Underground Pipelines play important role in transporting gas, water, and fuel. In addition to cooking and cleaning, the daily commute, air travel and the heating of homes and businesses are all made possible by fuels delivered through pipelines. These routine activities really add up, in terms of energy use. Natural gas and petroleum provides for 24% and 39% of our country's total energy consumption, respectively (Williams, 2015). In addition, underground utilities can be hit and damaged by trucks, excavation causing problem to underground utilities.

Underground pipeline damages can be attributed to two main causes: The lack of reliable data regarding the true location of underground utilities and the lack of communicating the all information. Inaccurate utility location information leads falsely instilled confidence and potentially misleads equipment operators into unintentionally utility strike, proposed ground penetration radar(GPR) to visualize and map underground utility by (Li, Cai, & Kamat, 2015). Although this integrated system showed promise, its accuracy in locating deeply underground buried utilities still concern (Cai, & Kamat, 2014).

On the other hand, many of today's underground utilities are reaching the end of their practical life and need to be replaced or repaired. Thus, precise information of underground utilities is important to utility owners, engineers, and contractors as reference for excavation (Jaw & Hashim, 2013). Jaw & Hashim examines of the accuracy of data used acquisition by scanning technique.

Underground utility damage mainly occurs because of overlapping of the geospatial utility location and the movements of excavation equipment. A proposed computational detail in geometric modeling for geospatial of utility data for 3D visualization and proximately monitoring to support knowledge- based excavation (Talmaki, Kamat, & Cai, 2013). However, there are limitations through the different stages of underground utility excavating cycle. It was estimated that nearly 500,000 utilities damaged in yearly bases in the United States. The decade from 2001–2010 saw a total of 544 major excavation related accidents resulting in 37 fatalities, 152 injuries, and close to $200 million in property damage. The lack of accurate position and semantic data of buried utilities coupled with absence of persistent visual guidance are two key problems facing excavator operators (Talmaki & Kamat, 2012). The third obstacle for safe excavation operations is the lack of real-time spatial awareness of the proximity of the digging implement to the underlying neighborhood utilities (Talmaki & Kamat, 2012).

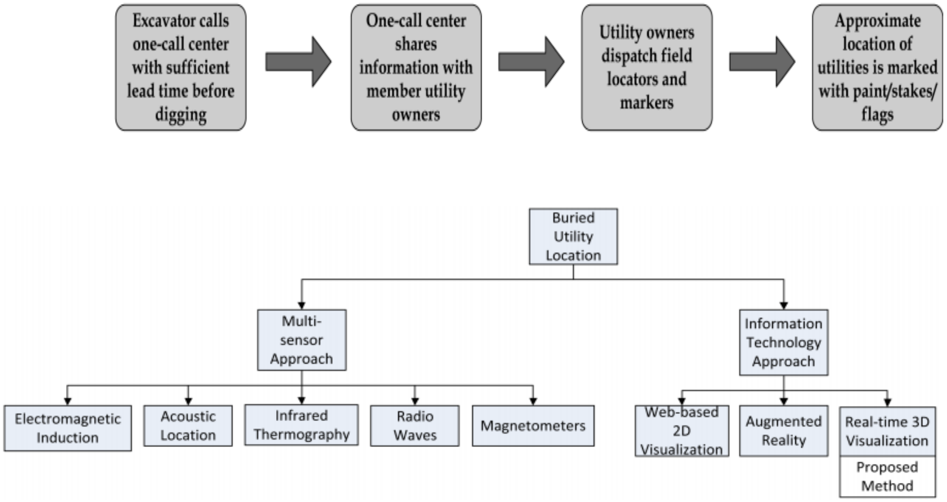

Excavator calls one-call center with sufficient lead time before digging
One-call center shares information with member utility owners
Utility owners dispatch field locators and markers
Approximate location of utilities is marked with paint/stakes/ flags
Buried Utility Location
Multi-sensor Approach
Information Technology Approach
Electromagnetic Induction
Acoustic Location
Infrared Thermography
Radio Waves
Magnetometers
Web-based 2D Visualization
Augmented Reality
Real-time 3D Visualization
Proposed Method

Figure 15: This Workflow is Illustrated & Proposed by (Talmaki & Kamat, 2012)

Even with the introduction of the one-call remarking system (Figure 15), accidents continue to happen. The year 2010 saw over $22 million in damage caused by excavation related accidents (PHMSA 2012a). In addition, Talmaki & Kamat classified buried utility location techniques into two subtypes as depicted in Figure 16. The first group uses a combination of

Figure 16: This workflow is illustrated & Proposed by (Talmaki & Kamat, 2011)

geophysical technologies to accurately determine the location and type of buried utilities. The first category is referred to as the multisensory approach. The second category uses a combination of geospatial databases and tracking technology.

As result, Talmaki & Kamat introduced the concept of HV simulations as a means to emulate real-world operations in a 3D virtual world using tracking and geographic information from the real world. A vision-based pose estimation solution for articulated machine using camera marker network was tested (Feng, Dong, Lundeen, Xiao, & Kamat, 2015): which is basically applying single camera facing each side of the operating machine and marker to find the relationship in united system. Conducting accurate excavation is a challenging task for excavator operators and utility owners who typically use the directions of mark outs person to achieve design grades and levels.

Construction equipment monitoring has been extensively investigated at both macro and micro levels. At the macro level, users are interested in simultaneous localization of several machines in a fleet in real-time for productivity measurement, safety, and fleet management purposes. A variety of technologies including GPS (Navon and Shpatnisky, 2005, Navon et al., 2004), UWB (Teizer et al., 2008), and computer vision (Rezazadeh Azar et al., 2013, Memarzadeh et al., 2013, Rezazadeh Azar and McCabe, 2012a, Rezazadeh Azar and McCabe, 2012b, Gong and Caldas, 2011) have been applied to localize and track construction machines. (Azar, Feng, & Kamat, 2015).

A generic and scalable computer-vision based framework for real-time pose estimation of an excavator’s boom and dipper (stick) using low-cost markers installed on the side of the arms. (Azar, Feng, & Kamat, 2015). This focuses on mark out accuracy based on the pose estimating of an excavator's boom and dipper.

In addition, development of robotic excavator has been a popular topic for the last two decades and in some of the developed prototypes, the control unit of autonomous excavator perceives the pose of the arm elements using various sensing devices (Stentz et al., 1999, Chiang and Huang, 2004, Yamamoto et al., 2009). Moreover, categories of the different stages in the lifecycle of underground utility geospatial data complicates the analyses and precludes its use in downstream engineering applications such as excavation guidance. Five key requirements – Interactivity, Information Richness, accuracy characterization on, and extensibility – were identified as necessary for the consumption of geospatial l utility data in location-sensitive engineering applications (Talmaki, Kamat, & Cai, 2013). As shown in the proposed process chart in Figure 17, the flow chart starts with location, then pass through digging, and the distance between the excavator & the buried underground utility.

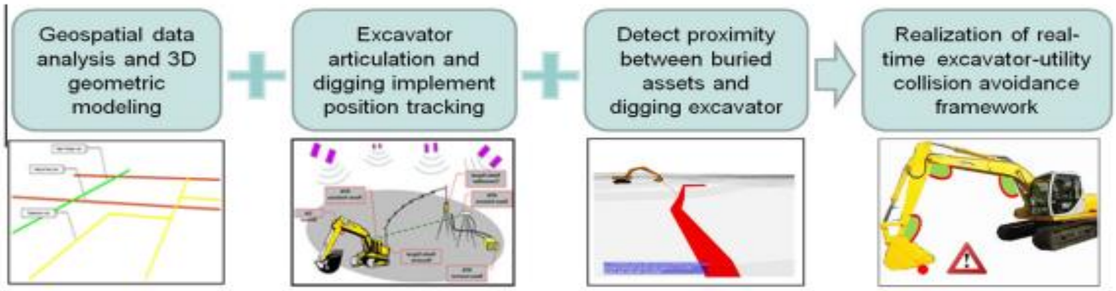

Geospatial data analysis and 3D geometric modeling
Excavator articulation and digging implement position tracking
Detect proximity between buried assets and digging excavator
Realization of real-time excavator-utility collision avoidance framework

Figure 17: Computational Framework for Knowledge-Based Excavation Operations by (Kamat & Cai 2013).

Damage of the underground utilities was influenced by various factors, including the accuracy of the location of the underground pipe in terms of the distance and exact level of the underground pipe. In addition, the operator's errors in terms of following the mark out and locating the arm of the excavator in the exact location. Transmitting the right information to the one call center, underground utilities, and excavator is another factor which if it does fail may cause the excavator to hit or even damage the underground pipeline.

1.7 Categorize One Call Center Process

One Call Centers serve as the clearinghouse for excavation activities that are planned close to pipelines and other underground utilities. One Call Centers help to protect underground telephone service, power lines, water and sewer pipes and energy pipelines. The processes were broken into the following four categories: One Call Center; Underground Facility Operator; Locating Company; and Excavation as shown in Figure18. Then, the tasks broken to more details to include initiation process chart starting from excavator notifying the one call center, the board

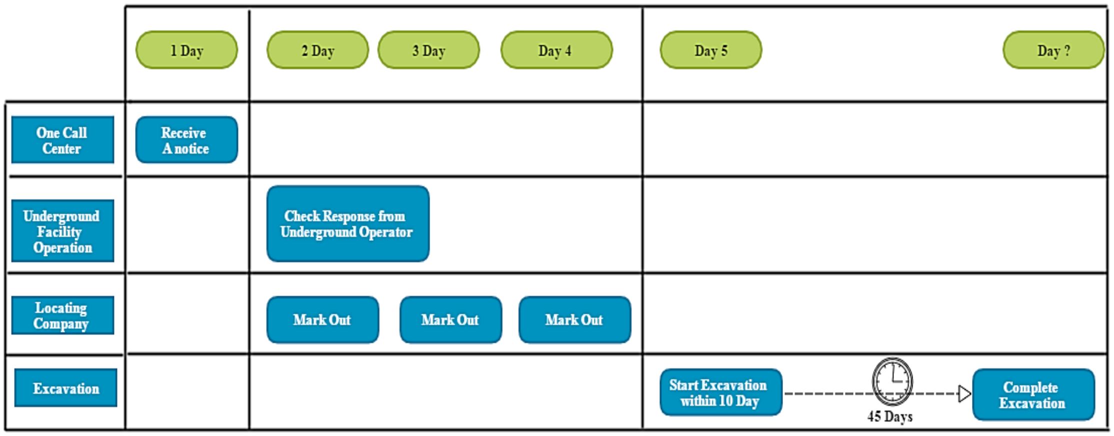


Figure 18 Show High Level One Call Center Processes

designed one –call systems receive the request, Underground facility operator verify the location and existing utilities in the particular site, locating the pipe and mark out, and finally start excavation. In addition, this process is varies between emergency case and regular routine.

Moreover, after the high level one call center was developed, it was necessary to look for subtasks which means breaking the major steps into more details in order to map out the all processes. The Operator receives and records the notice of intent to excavate provided. Then, assign a confirmation number to each notice of intent to engage in an excavation; inform the excavator or responsible contractor of the confirmation number. For each of the notice of intent, the operator maintains a register showing the name, address, and telephone number of the excavator or responsible contractor, the site to which the notice pertains, and the assigned confirmation number. This information is promptly transmitted to the appropriate underground facility operator(s) the information received from an excavator or responsible

contractor regarding intended excavation or demolition. After mark outs are made, the excavator notified and needs to start digging within time frame and boundaries of mark out as can be seen in Figure 19.

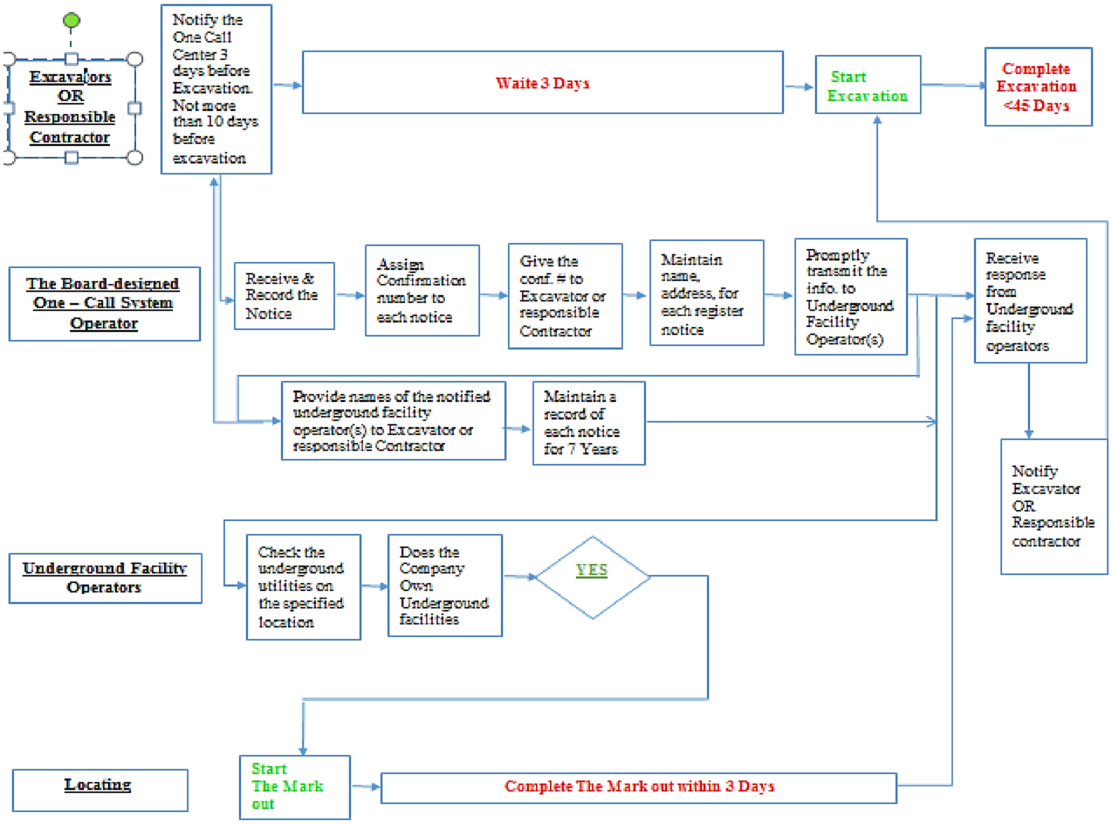

Excavators OR Responsible Contractor
Notify the One Call Center 3 days before Excavation. Not more than 10 days before excavation
Waite 3 Days
Start Excavation
Complete Excavation <45 Days
The Board-designed One – Call System Operator
Receive & Record the Notice
Assign Confirmation number to each notice
Give the conf. # to Excavator or responsible Contractor
Maintain name, address, for each register notice
Promptly transmit the info. to Underground Facility Operator(s)
Receive response from Underground facility operators
Provide names of the notified underground facility operator(s) to Excavator or responsible Contractor
Maintain a record of each notice for 7 Years
Notify Excavator OR Responsible contractor
Underground Facility Operators
Check the underground utilities on the specified location
Does the Company Own Underground facilities
YES
Locating
Start The Mark out
Complete The Mark out within 3 Days

Figure 19: Shows information Flow Chart for Digging Requests

1.8 Data Driven Risk Analysis.

Data-driven analysis (DDA) is an approach to business governance that values decisions that can be backed up with verifiable data been developed through phases or stages of analysis. The success of the data-driven approach depends on the quality of the data collected and the effectiveness of its analysis and interpretation to develop well educated decision. In addition, data-driven analysis methods, such as independent component analysis and clustering, have been effective application in the analysis of functional magnetic resonance imaging data for identifying functionally connected risk assessment analysis. Even though independent component analysis and clustering rely on very different assumptions on the underlying distributions, both give similar results for signals with large variation.

The data-driven analysis was used in many studies to explore the multivariate risk structure of the data: aiming to identify the effective components. These components may reveal structures or patterns in the data, which are difficult to identify apriority: such as unexpected activation and connection, motion related artifacts, and drifts (Biswal, 1995). These data driven analysis methods provide generalizations of connectivity analysis in situations where reference seed regions are unknown or difficult to identify reliably. One important motivation and expectation behind the use of these methods is that in many data sets, data points lie in some manifold of much lower dimensionality than that of the original data space (Christopher, 2006). The four most popular methods are the following: clustering; principal component analysis; independent component analysis; and probabilistic principal component analysis.

Many underground gas line researches used principal component analysis as a statistical technique to linearly transform an original set of variables into a substantially smaller set of uncorrelated variables. It is also known as the Karhunen-Loeve transform (Ringnér, 2008). One of the main goals is to reduce the dimensionality of the original data set. In addition, a group of uncorrelated variables data is assumed to represent the underlying sources for observations, and is more computationally efficient in further analysis than a larger set of correlated variables. Therefore, Principal component analysis method is often used as a pre-processing step for other data-driven analysis methods such as clustering.

We are using a Gaussian latent variable model in developing a more precise probabilistic formulation of PCA. This probabilistic formulation of PCA provides a way to find a low-dimensional risk representation of higher dimensional data with a well-defined probability distribution, and enables comparison to other generative models within a density estimation framework (Tipping and Bishop, 1999). Moreover, there are some advantages of probabilistic principal component. First, this probability model can be used to provide samples from the distribution. Second, it gives an explicit probability model of the data, in the density estimate framework: which allows us to calculate the likelihood of any observation and to compare the result of probabilistic principal component to other exploratory data analysis methods as mentioned above.

Clustering or data segmentation is another method which could be used. , Clustering groups a collection of data points into subsets such that the points in each subset are more closely related to each other than those in other subsets, where each cluster itself is as different as possible from other clusters (Kim, 2008). In many real data cases where multiple clusters are present, a simple probability distribution is insufficient to capture the structure of the data. A linear combination of more basic distributions, known as mixture distribution, gives a better characterization by providing a framework upon which to build a more complex, richer class of density models.

As per (Winter, 2003), a comprehensive methodology that supports the entire process of determining information requirements for data warehouse users, matching information requirements with actual information supply, evaluating and homogenizing resulting information requirements, establishing priorities for unsatisfied information requirements, and formally specifying the results as a basis for subsequent phases of the data. The experts' requirements to information requirements analysis in a data warehousing context call for a demand driven approach. Since the business process oriented approach is not applicable if the data warehouse system has to support decision processes, we focus on a 'conventional' demand driven approach. The proposed methodology should overcome the shortcomings listed, i.e. a multi-stage approach has to be taken, users have to be supported in specifying objective (and not subjective) (Winter Author).

1.9 Risk Identification Method using Bow tie

The bow-tie method is a risk evaluation method that can be used to analyze the risk or danger involved in high level of risk scenarios. Bow-tie diagram does two things. First, a Bowtie gives a visual summary of all plausible accident scenarios that could exist around risk involved in any process. Second, by defining control measures, the bow-tie displays what can be done control those scenarios. Traditional 'bow-tie' approach is not able to characterize model uncertainty that arises due to assumption of independence among different risk events. In other words, Bow-tie does not really have a good handle of analyzing complex risk networks. The traditional 'bow-tie' analysis requires the probability of input events as precise crisp data or defined probability density functions (PDFs) (Markowski et al., 2009). Bow-tie' analysis is an integrated probabilistic technique that analyzes the accident seniors in terms of assessing the probabilities and pathways of

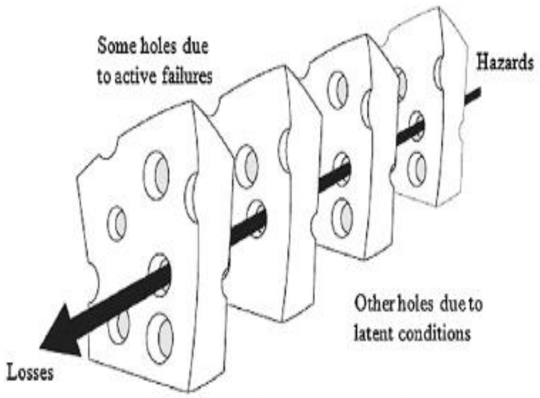

Some holes due to active failures
Hazards
Other holes due to latent conditions
Losses

Figure 20 : Swiss Cheese model adopted from Reason et al.(2001)

occurrences (Duijm, 2009).

Handling risk events in conventional ‘bow-tie’ analysis, basic risk events are limited. Therefore, such probabilities are often hard to come by due to insufficient statistical data and knowledge. Consequently, such rough probabilities may lead to ‘precise’ but unrealistic results. It is used to control and mitigate undesired events by developing a logical relationship between causes and consequences of an undesired event (Dianous and Fievez, 2006). Besides, traditional ‘bow-tie’ analysis uses a default assumption of “independence” among the failure events which has some defects.

Moreover, Bowtie approach by Reason is that early damage barrier model is the classical “Swiss Cheese model” shown in Figure 20 is developed by Reason (Reason, 1990). In this model, each slice is a barrier while the hole rep-resents the weakness or failure of system. If all of holes align, the accident will occur, otherwise, the accident does not occur.

Bow-tie’ is a common platform which couples FTA and ETA by considering a common top-event named as critical event (Cockshott, 2005; Cauchois, 2006; Fiévez, 2006; Duijm, 2009). Specifically, ‘bow-tie’ model is a constructive risk management tool, providing a graphical representation of the relationship between risks, initiating events, controls and consequences. It is widely used by engineers, management, process operators and maintenance personnel involved in risk management. It is more of a risk ranking method which is commonly used to evaluate the risks of simple likelihood-consequence pairs and are straightforward in application to define the structure of the model.

The bow-tie model has entered the field of occupational safety through the European Workgroup for development of the Occupational Risk Model (WORM): which started with the aim of decreasing by 10–15% the occupational accident rate in the Netherlands (Hale et al., 2005). Moreover, the bow-tie method seems particularly useful to represent the influence of safety systems (and barriers) on the progression of risk scenarios. Safety systems, either technical or organizational elements, can be placed in the two main branches of the diagram.

The building of the bow-tie diagram is a complex task: it not only requires reliable data on the frequency of all events, but the failure probabilities of the barriers need to be known as well. This type of assessment also calls for the involvement of highly specialized people from different expertise areas. For all these reasons, it is unlikely that individual enterprises will be able to apply the model in this way. Despite this, the developed diagram contains an attractive basis to support the risk analysis. Therefore, it becomes apparent that the bow-tie approach represents a step forward in the current state of the art concerning the management of risks, including those associated with Fault Tree. This is the context in which the authors equated the use of the bow-tie diagram in combination with a matrix approach, based on accident statistics of the activity under analysis (Jacinto, 2010).

1.10 Risk Identification Method using Fault tree

The FT is a graphic expression to show how an event can occur in different ways and systematically identify the probable sequence of events. FTA is a systematic method for analyzing the cause of risks by adopting a deductive method. In this approach,a specific risk that is only qualitatively recognized from a relevant primary system is placed as the top event in the tree for deductive reasoning (Hyun, 2015). The occurrence of the top event can be quantitatively estimated based on the probability of each risk factor occurring. FTA also permits the theoretical relation between the risk categories (top events), the risks (gates or sub-gates) and the risk factors (events) to be clarified on the basis of AND and OR logic. The method can explain, for instance, how equipment defects and human mistakes can be combined to cause a risk in the relevant primary system (Hyun, 2015).

The fault tree is a logic diagram based on the principle of many damages or accidents: which traces all branches of events which could contribute to the risk damage (Shahriar, 2012). In order to facilitate that, it uses sets of symbols, labels and identifiers. Fault tree analysis (FTA) and event tree analysis (ETA) are two graphical techniques used to perform risk analysis: where FTA represents causes (likelihood) and ETA represents consequences of a failure event. ‘Bow-tie’ is an approach that integrates a fault tree (on the left side) and an event tree (on the right side) to represent causes, threat (hazards) and consequences in a common platform (Shahriar, 2012).

Chang used FTA to set up a model to reduce the probabilities of data misuse or system crash (Chang 2007). Doytchev and Szwillus (2009) used FTA to analyze accidents and incidents in order to prevent the propagation of a chain reaction due to a single failure. Ortmeier and Schellhorn used FTA for safety concerns associated with automation and transportation systems (Ortmeier and Schellhorn 2007). Lindhe et al. (2009) used FTA on a drinking water supply system to help decision-makers minimise sub-optimization of risk-reduction options.

Park and Lee (2009) performed FTA to investigate causes for faults in hand-washing processes, as part of a hospital hygiene management program. In addition, FTA has been used in industries, such as energy resource estimation of failure probability in oil and gas transmission pipelines (Yuhua and Datao 2005). The printed circuit board industry has used FTA to calculate the fault interval of system components and to determine the most critical component in the production process (Shua, Cheng, and Chang 2006). The service industry has used FTA to analyze a large-scale and complex service process and to reflect the customer participation perspective (Geum et al. 2009).

To introduce Fault Tree in simple terms, we consider a simple fault tree comprised of two independent basic events, B1 and B2, linked to the top event A through an OR-gate (Flage, 2013) Figure 21. As can be seen above, these studies clearly show that FTA is widely used, especially in high-risk, complex or multi-element systems, or when there are numerous potential contributors to a mishap (Cheng, 2013). Since the reasons for risk identification are multiple and complex, i.e. high risk, this study used FTA to investigate the root causes of underground gas pipe damage, identify the key factors and risk model. It is very important to evaluate safety and reliability of complex and large scaled underground gas pipe network (Yuhua, 2005). Fault tree is commonly used to predict reliability of the complex network in many fields, such as UG operator, pipe locating, third part damage, and gas pipelines. In conventional fault tree analysis, the failure probabilities of

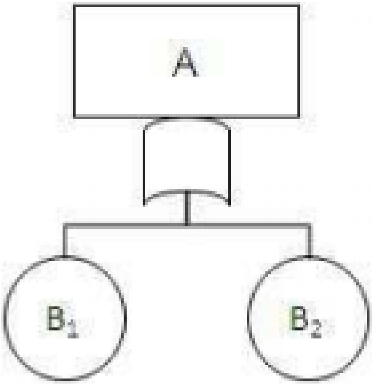


Figure 21Simple fault tree

components were considered as exact values (Chen, Wang, & Meng, 1995; Liao, Yao, & Zhang, 2001).

1.11 Machine learning Method

The straight rule probably would not be immaculate. However, it will give more exactness over the long run, since it will more successfully sum up a restricted arrangement of information to the population at large. It was noticed that for more unpredictable guidelines, the algorithms for machine learning must utilize greater informational collections to combat the generalization of errors. Since the algorithm of machine-learning work to streamline basic leadership, utilizing code and informational indexes that can be held up to the public scrutiny, the decision-maker think machine learning is not biased. In any case, separation can emerge in a few non-evident ways. To start with, the data which encode the biases. For instance, the algorithm that utilizes the preparing information for prediction whether somebody can commit a crime and should know whether the general population represent the set of data that really carried out violations or commit crimes. However, that data is not accessible—rather, an eyewitness can know just whether the general population were arrested for that crime and police can also arrest certain gatherings of individuals that may well create the biases. Second, the machine learning algorithm made utilizing deficient measures of training data that can cause a loop of feedback that makes the unfair results: regardless of whether the person did not intend to encode bias or he is guilty.

It was clarified that a moneylender can see whether a credit was paid back just if that it was conceded in any case. If that training data erroneously demonstrate that a gathering with a specific element is less inclined to pay back a credit. In light of the fact that the bank did not gather enough information, at that point the moneylender may keep on denying those individuals loans to maximize an earnings. The moneylender could never realize that the gathering is really credit-commendable, just because that the lender could never have the capacity to watch the rejected gathering's loan reimbursement behavior as well. An artificial intelligence turns into a consideration center in later a long time halfway because of the achievement of deep learning applications.

Deep learning algorithms and advances in GPUs alongside with the substantial datasets enable the large learning algorithms to address the real-world issues or problems in numerous areas or territories: from picture order to social insurance expectation, and from auto game playing to reserve engineering. Numerous logical and building fields are energetically embracing the deep learning. These energetic appropriations of new machine learning algorithms have started the advancement of various deep learning structures. For example, Tensor Flow, Torch. These structures empower quick advancement of deep learning applications. A structure gives normal building squares for layers of a neural system. By utilizing these systems, engineers can center around show outline and application particular rationale without stressing the coding subtle elements of the GPU enhancements, input parsing or matrix duplication. Petroleum gas and unrefined petroleum generation is normally conveyed through long-separate transmission metallic pipelines. Due to the idea of condition and extraordinary temperature, metallic pipelines are subjected to erosion. It has been accounted for that about 30% of pipeline collision are because of outside consumption. These pipeline deformities can result in immense money related misfortunes, harm to the earth, and death toll. Consequently, pipeline administrators are required to use powerful and productive shrewd instruments to recognize and find pipeline deserts. Efficient intelligent instruments use Magnetic Transition Leakage (MFL) signals and ultrasonic waves and utilize them to identify and localize defect types (e.g., corrosion, cracks, dents, etc.). MFL accounts around the focal point of a loss of metal defect and do have an unmistakable example of the behavior. The sensor passing specifically over the imperfection focus has most elevated sufficiency of the radial components and axial of the MFL flag. The sufficiency of these parts gets bring down for sensors facilitate far from the center of the defect. Utilizing the MFL estimations of the area sensors, the sort and size of the defect can easily be resolved.

In the literature, a few methods have been proposed to detect also, confining pipeline defects. utilizing the MFL signals, Artificial Neural Networks (ANNs) are utilized to order patterns of signals of different kinds of defects. These surrenders were fabricated and intentionally embedded. The ANN could recognize non-defect signals and defect signals with incredible exactness (94.2%) experienced pipeline administrators use Magnetic Motion Leakage (MFL) sensors to test oil and gas pipelines to localize and estimating distinctive deformity composes. The large number of sensors is normally used to cover the directed pipelines. The sensors are similarly, appropriated around the perimeter of the pipeline; and each three millimeters the sensors measure MFL signals.

Therefore, the gathered raw information is big to the point that it makes the pipeline examining process troublesome, error-prone and exhausted. Machine learning methodologies is a key. For example, neural systems that have made it conceivable to successfully deal with the unpredictability relating to huge information and take in their inborn properties. We think, in this work, on the appropriateness of artificial neural systems in imperfection profundity estimation and present a definite investigation of different system designs. Discriminant highlights, which characterize different defect depth patterns, are first acquired from the raw data. The Neural systems are at that point prepared utilizing these types of features. The Levenberg-Marquardt back- propagation machine learning algorithm is received in the preparation process: in which the weight and inclination parameters of the systems are tuned to enhance their performances. Contrasted and the execution of pipeline assessment methods revealed by specialist Organizations. For example, GE and ROSEN, the outcomes got utilizing the technique we proposed (Stephen Hawking,2018).

Applying artificial learning and machine learning to your basic decision making can enable your business to remain focused. In any case, a considerable measure can turn out badly in your route. Without the best possible governing rules, machine learning efforts can be out of control: presenting your organization to risks or dangers. Machine learning is regularly used to make predictions or forecasts. The examples related to this anticipating movie, search results, anticipating customer purchasing behavior, or product selections, predicting new types of hacking techniques. One reason expectation might be erroneous that may have something to do with "over fitting." Over fitting happens when a machine learning algorithm adjusts excessively to the noise in the data: as opposed to revealing the fundamental signal. There is a lack of machine learning ability out there. In the interim, machine learning is being democratized as the capacities discover their way into more applications and simple to-utilize stages that can cover the hidden many-sided quality of machine learning. The most likely best-known impediment of Neural Networks is their “black box" nature, implying that you do not know how and why your NN thought of a specific yield.

For instance, when you put in a picture of an animal into a neural system and it predicts it to be a car or anything else, it is difficult to comprehend what made it came up with this forecast. When you have highlights that are human interpretable, it is considerably simpler to comprehend the reason for its mistakes. In Comparison, the algorithms like Decision trees are exceptionally interpretable. This is vital in light of the fact that in a few areas, interpretability is very imperative. This is the reason a considerable measure of banks doesn’t utilize Neural Network to predict whether a man is financially weak since they have to disclose to their clients why they do not get a loan. Something else, the individual may feel wrongly debilitated by the Bank, since why he does not get a loan: which could lead him to change his bank. A similar thing is valid for destinations. If that they would choose to erase an user account in light of a Machine Learning algorithm, they would need to disclose to their users why they have done it (JONATHAN VANIAN, 2018). Therefore machine learning process lacks general intelligence and multiple domain knowledge integration. The intelligence of human civilization accelerates due to connectivity between people. Neural networks fed inaccurate or incomplete data will produce the wrong results. The outcomes can be embarrassing as well (I. Ioannou,2012).

Why we should employ this algorithm in our study model? With the measure of data that is accessible in the hiring procedure, machine learning can uncover significantly more effective techniques for recognizing solid competitors. Machine learning is the investigation of computational learning hypothesis in artificial intelligence and pattern recognition. By making a "model", which is basically a prepared set of data produced using the sample inputs, an organization can make accurate decisions or predictions as yields as opposed to following a static system. This model can be utilized to settle on more brilliant decisions in recruitment procedure (Azahara,2016).

The precision medication is a quickly developing territory of present day therapeutic science and open source machine-learning codes that guarantee to be a basic part for the successful improvement of institutionalized and mechanized investigation of patient information. One critical objective of exactness growth medication is the precise forecast of ideal medication treatments from the genomic profiles of individual patient tumors. We present here an open source programming stage of machine learning algorithm that utilizes a profoundly versatile support vector machine (SVM) algorithm that joined with a standard recursive feature elimination (RFE) and it is the way to deal with various drug reactions from the gene articulations profiles. The drugs particular models were constructed utilizing quality articulation and the various drug reaction information from the National Cancer Institute board of 60 human

growth cell lines. The models are exceptionally exact in anticipating the drug responsiveness of an assortment of tumor cell lines including those containing the ongoing NCI-DREAM challenge. The developed model shows that prescient exactness is advanced when the machine learning dataset uses all test set articulation esteems from a diversity of cancer cell that composes without pre-sifting for genes that are for the most part and thought to be "drivers" of cancer progression/onset. Utilization of the model to public ally accessible ovarian cancer (OC)

patient gene expression datasets generated predictions consistent with observed responses previously reported in the literature. By influencing this machine learning algorithm, this would encourage its testing in different types of cancer;the context that must be leading to driven changes and refinements in resulting applications (Bradley A Fritz,2017).

The deep learning presents a great tool for breaking down medicinal pictures. The Retinal infection recognition by utilizing PC determination from fundus picture has risen as another strategy. By using the algorithm of the deep learning, the neural system utilizing for a computerized identification or automated detections of numerous diseases(retinal).

The Dataset was worked by extending information to 10 classifications: which includes the nine Retinal disease and normal retina. The ideal results were obtained by utilizing an irregular backwoods transfer learning in view of VGG-19 design. The characterization results depended enormously on the quantity of classifications. As the quantity of classes expanded, the execution of deep learning models was decreased. Besides, a few group classifiers upgraded the multi-straight out characterization execution (Joon Yul Choi,2017). Just because of the small size of datasets, the deep learning procedures in this investigation were insufficient to be connected in facilities where various patients experiencing different kinds of retinal disorders visit for finding and treatment.

For example, functions or regularization methods, representation transformations may at first be communicated in numerical documentation, they should be transcribed into a PC program for true use. For this reason, there exist various open sources as well as business machine learning programming libraries and systems. Among these are scikitlearn. These libraries were reached out by Tensor Flow, a novel machine learning programming. According to the underlying production, the Tensor Flow means to be an interface for communicating machine learning algorithm in vast scale on heterogeneous frameworks of distributed (Joon Yul Choi,2017).

We will be building predictive model to predict the future risks. Because of the advance innovation related with Big Data, information accessibility, and processing control, most banks or loaning organizations are restoring their plans of action. To predict the future expectations, observing, display dependability and compelling credit handling are vital to basic decision making. In this work, parallel classifiers were constructed in view of machine and deep learning models on genuine data to predict the probability of the loan. The best 10 essential highlights from these models are chosen and after that utilized in the displaying procedure to test the security of paired classifiers by contrasting their execution on the separate data. It was noticed that the tree-based models are stable enough than the models in view of multilayer neural systems. This opens a few inquiries in respect to the escalated utilization of the deep learning frameworks.

To assemble a model which predict the future risk, more data (historical) were required that enables us to catch data about various events that is leading towards the risk. Moreover, general "static" highlights of the framework can likewise give important data. For example, average usage, operating conditions and mechanical properties. If more information isn't in every case better. The accomplishment of the predictive models relies upon three primary parts: having the correct data; evaluating the predictions properly; and framing the problem appropriately. The life length of machines is as a orders of years: which implies that the data must be gathered for a broadened timeframe with a specific end goal to watch the system all through its debasement procedure (Peter Martey Addo, 2008).

1.12 Random Forest Method

The random forest which was presented by (Dasgupta et al. 2014) is a reliable, non-parametric technique of regression procedure that when connected to the binary results and empowers the calculation of estimation of effect size predictor. Utilizing the simulation, the random forest is found to appraise fundamental impacts for categorical predictors and binary. Besides it produces an interaction impacts with minimal bias for binary predictors. These assessments are nearly as productive as those from a logistic regression effectively which display when the information creating model must be logistic. The method of intuitive interaction detection is appeared to be a moderately fast screening procedure to distinguish any potential communication impacts. However, we have to careful when utilizing the random forest to gauge the impact of a continuous predictor: which produces the estimation with insignificant predisposition when the impact estimate is small and linear. The random forest strategies are connected to an extensive Nova Scotia dataset to distinguish and to measure the risk components for fetal development variations from the norm (Chunrong Mi, 2017).

The clinical datasets are generally constrained in estimate, along these lines limiting utilizations of Machine Learning (ML) strategies for prescient demonstrating in clinical research and organ transplantation. The capabilities of was investigated that is Random Forest models of classification, with regards to little dataset of 80 tests, for result expectation in kidney transplantation and the result is quite risky. The RF and DF models distinguished the key hazard factors related with intense dismissal: the levels of the contributor particular IgG antibodies and the levels of IgG4 subclass and the quantity of human leucocyte antigen jumbles between the recipient and the donor. Besides, the DT show decided risky levels of benefactor particular IgG subclass antibodies. It also exhibits the capability of finding new properties in the information when tools of traditional statistical can't catch them. The DT and RF classifiers created in this work anticipated early transplant dismissal with exactness of 85%, consequently, offering a precise choice help device for the doctors that is entrusted with foreseeing results of kidney transplantation ahead of time of the clinical intercession (A.V.Lebedev,2014).

In clinical and biomedical building space ML offers prescient models. For example, artificial Neural Networks (ANNs), Random Forests (RFs), Support Vector Machines (SVMs), Decision Trees (DTs) which can delineate non-straight heterogeneous information and the designs of patterns, when physiological connections between display factors couldn't be resolved because of multifaceted nature, pathologies, or absence of understanding of biological. Moreover, the random forest models are once in a while seen with regards to little information, where inadequate number of preparing tests that can bargain the learning achievement (A.V.Lebedev,2014).

The expanding utilization of electrical vitality has yielded more necessities of electric utilities including transmission lines and electric arches which require a continuous hazard checking to avoid gigantic gas pipes damages. As of late, Airborne Laser Scanning (ALS) has turned out to be one of essential information procurement instrument for mapping because of its capacity of direct 3D estimations. For power line risk management, a quick and exact arrangement of electrical cable articles is a critical undertaking. As a base classifier, Random Forests (RF) was utilized. RF is a composite descriptor comprising of various choice trees populated through learning with bootstrapping tests. The Two unique arrangements of highlights are researched that are separated in a point space and an element (i.e., line and polygon) area. Minimum Description Length (MDL) and RANSAC are connected to make lines and a polygon in each volumetric pixel (voxel) for the line and polygon highlights. Two RFs are prepared from the two gatherings of highlights uncorrelated by Standard Component Analysis (PCA). Results from these two RFs are joined for conclusive characterization. To explore different avenues regarding two genuine datasets exhibits that the proposed methods of classification strategy indicates 10% changes in classification exactness that is contrasted with a solitary classifier.

The particular area of inconsistencies on bearer pipe inside housings is 2% of the packaging length, or around 3 ft overall from either end of the packaging. Past 3 ft the peak anomalies are moderately consistently appropriated. The particular area contains 25% of pinnacle abnormalities, or 10 times the probability of anomalies somewhere else in the packaging (A.V.Lebedev,2014).

An Administrators for the most part consider cased pipe sections to be protected in light of the fact that time-autonomous dangers, including outside force damage and third-party excavation are generally disposed of. The potentially upgraded outer consumption of the bearer caused by the packaging, moreover, bargains this safe contention. There is no evidence of sound that exists to propose that cased sections are more secure than uncased ones as far as holes per mile or booked or prompt reactions per mile. A 1984 overview rather uncovered that, of 14 nations, five announced the major damage on the gas pipe when casing was utilized. However,, none of the studies detailed the damage(corrosion) on the carrier pipe at intersections when casing is not even used. The carrier pipe can turn out to be more extreme when the external corrosion takes place on uncased fragments in presence of electrolyte. Such electrolyte exposure may happen due to the following issues: the condensation from the funnels open to air, or within the sight of a short if ground water obtains entrance into the packaging transporter annulus when casing end the seals are either missing or don't legitimately seal (A.V.Lebedev,2014).

Significantly expanded random forest damage has as of late been seen in the Republic of Croatia. The goal is to investigate the potential outcomes of reliable or simple detection, reviewing (mapping) and checking the safety forest condition by methods for shading infrared (CIR) symbolism and geostatistical strategies.

Following methods and material will be used to predict the future damage: The Four trees (crowns) nearest to the point of the raster (100 × 100 m) was set up in the computerized photo for the zone; and deciphered in CIR pictures. The random forest damage was ascertained for the entire territory under perception. The identification of spatial distribution of these damage indicators and assessment and was performed utilizing raster point information: from which an irregular (966 focuses) and an orderly (445 focuses) test were made. The outcomes on the random forest damage that is procured by deciphering CIR pictures. A model of hypothetical parameters were utilized to add the both the damage pointers with customary kriging. The Nonstop maps of the damage degree circulation was then developed. The consequences of insertion were tried with the cross-approval technique. The Damage marker maps are the after effect of the accompanying: method, data variability, and sampling intensity. The Tree damage for the most part does not have consistent but instead arbitrary spatial circulation. This is the reason the essential point in distinguishing the random forest damage is to consolidate the entire territory of enthusiasm into testing. The Testing force ought to be adjusted to the required precision and to the time and subsidies available to us. This exploration depends on the utilization of CIR elevated photos and geostatistical apparatuses in spatial investigation of the forest damage. The Persistent maps of damage pointers procured with kriging give a superior understanding into the spatial dissemination of harm than do topical maps acquired by translating CIR flying symbolism based on a precise example (the raster strategy). The Combination of understanding consequences of CIR flying pictures and statistical approach guarantees a more exact circulation of the damage forest indicators. Therefore,, the likelihood of better spatial investigation of the event, patterns and improvement of harm in the examination territory.

Why we should employ this algorithm in our study model? The random forest is a group machine learning calculation, which is best characterized as a "mix of tree indicators to such an extent that each tree relies upon the estimations of an arbitrary vector examined freely and with a similar appropriation for all trees in the forest. In numerous applications this calculation produces a standout amongst other exactness’s to date and has critical favorable circumstances over different procedures; for example,, the capacity to deal with exceedingly robustness to noise, tuning simplicity non-linear biological data (contrasted with other gathering learning algorithms and it is open door for effective parallel preparing (De Bruyn et al., 2013; Menze et al., 2009). These variables additionally make RF a perfect possibility for taking care of high-dimensional issues, where the quantity of highlights is regularly excess.

In spite of the fact that Random forest would itself be able to be considered as a compelling component choice algorithm and a few methodologies to set the reduction for feature inside and outside the setting of Random forest that have been proposed to additionally enhance its execution (Tuv et al., 2009). In the present examination, we utilize recursive element disposal (Kuhn, 2012a) to enhance the models. The past work uncovered that thickness of parceled cortical, together with the subcortical volumetric estimations (utilized as a contribution to a multivariate model) brought about the best execution, contrasted with different modalities (Westman et al., 2013). Here, it has been pointed not exclusively to evaluate the exactness's of the classifiers prepared with various morphometric modalities. Yet additionally, to investigate the effect of the system on computation/memory/time costs of model training, feature selection and dimensionality.

At long last, the past examinations have effectively utilized example acknowledgment systems to group MRI pictures from various associates just inside the consolidated sets (Westman et al., 2011; Lebedev et al., 2014). The present investigation was arranged as one of the first to survey classifiers' between-companion power in two autonomous substantial scale datasets (Luckyson Khaidem,2016). We guessed that with the utilization of more disease particular chart books for this situation, when the estimations are separated from the predefined districts that is known to be influenced by Alzheimer's illness or disease. It conceivable to accomplish AD-recognition exactness equal to that of the models prepared with high-dimensional contribution without percolations with shorter computational time. Also, we estimated that it is conceivable to accomplish great between-associate speculation of the models if the MRI conventions are fit (Akbar K Waljee,2014).

Building predictive model to predict the future risks. A particular quickly creating field of neuroimaging with solid potential to be utilized every day. In this specific circumstance, evaluation of models' power to commotion and imaging convention contrasts together with post-handling and tuning techniques are key assignments be tended to keeping in mind the end goal to move towards the clinical applications. In this examination, the viability of Random Forest model was researched by utilizing diverse basic MRI measures: with and without neuroanatomical requirements in the recognition and expectation of AD as far as exactness and between-associate heartiness. The Huge information is changing each industry. The Pharmaceutical is no exemption. With quickly developing volume and assorted variety of information in medicinal services and biomedical research, the conventional measurable strategies regularly are deficient. By investigating different businesses where present day machine learning strategies assume focal parts in managing enormous information, numerous health and biomedical scientists have begun applying machine figuring out how to remove significant bits of knowledge from consistently developing biomedical databases: specifically with prescient models. The adaptability and ability of machine learning models likewise empower us to use novel, however, to a great significant sources of data: for example, electronic health record information and wearable gadget data.

In spite of its prominence, it is hard to discover an all-around settled upon definition for machine learning. Many machine learning strategies can be traced back to as early as 30 years prior. The surveys by Jordan give open reviews to machine learning. This paper centers around machine learning prescient techniques and models. These include the random forest, the vector machines, and different techniques recorded. They all offer a vital distinction from the customary measurable techniques, for example, analysis of variance or logistic regression and the capacity to make exact forecasts in unseen information. To enhance the forecast exactness, regularly the strategies do not endeavor to deliver the interpretable models. This additionally enables them to deal with the large variable in hugest data issues.

Most issues with applying machine learning techniques like random forest in biomedical research start from few basic concerns: including over fitting and data leakage which can be maintained strategic distance from by receiving an arrangement of best practice models. Perceiving the critical requirement for such a standard, we made a base rundown of detailing things and an arrangement of rules for ideal utilization of prescient models in biomedical research. The previous studies exhibited that these sample acknowledgment strategies are delicate to MR-convention contrasts (Westman et al., 2011; Lebedev et al., 2013) and that a harmonization step is thusly required. Another significant issue relates to the examination of high-dimensional imaging data input versus estimations. This is removed by neuroanatomical parcellation chart books. With the territories isolated by practical and histological maps of the human cortex for effortlessness, we will utilize the algorithm of random forest. The random forest has some conspicuous focal points, i.e.,lower calculation, memory cost and handling time. In any case, it is conceivable that it could be one-sided by these points of interest. Standardized high-dimensional estimations without parcellation, interestingly, are fair, yet in the meantime are harder to deal with utilizing multivariate and machine learning algorithms because of calculation and memory costs. In addition, circumstances where the quantity of estimations is considerably bigger than the quantity of perceptions ($p \gg n$) are regularly connected with the purported "revile of dimensionality" (Bellman, 1961).

The random forest (RF) is an outfit machine learning calculation, which is best characterized as a mix of tree indicators to such an extent where each tree relies upon the estimations of an arbitrary vector examined freely.

1.13 K-nearest neighbors Method

The k-nearest algorithm is a powerful and adaptable classifier that is frequently utilized as a benchmark for more perplexing classifiers: for example, Support Vector Machines (SVM) and Artificial Neural Networks (ANN).KNN can beat all the greater classifiers in terms of effectiveness and is utilized in an assortment of uses: for example economic data compression, genetics and forecasting.

The k-nearest is likewise an instance-based learning and non-parametric algorithm. Being non-parametric, it makes no explicit presumptions about the practical type: keeping away from the damage of miss modeling of the hidden dissemination of the data. For instance, assume the information is exceedingly non-Gaussian yet the picked learning model accept a Gaussian shape. All things considered, k-nearest algorithm would make greatly poor forecasts. Another example is instance-based learning which implies that this algorithm does not expressly take in a model. Rather, it remembers the preparation cases which are thusly utilized as "learning" for the forecast stage. Solidly, this implies just when an inquiry to the database is made that is when the request is made that it anticipates a name given an information and the algorithm utilize the preparation examples to release an answer.

In the present investigation, the k-Nearest Neighbor arrangement technique have been examined for financial determining. Due to the impacts of organizations' monetary trouble on partners, money related misery expectation models have been a standout amongst the most appealing regions in money related research. As of late, after the worldwide money related emergency, the number of bankrupt organizations has increased. Since organizations' monetary trouble is the principal phase of insolvency, utilizing money related proportions for foreseeing monetary pain have pulled in an excessive amount of consideration of the scholastics and in addition monetary and budgetary organizations. Despite the fact that lately considers to predict the companies with financial distress in some countries have been expanded, most endeavors have abused the statistical methods of traditional and only a couple of studies have utilized nonparametric strategies.

The prediction of the credit risk by KNN is characterized as the hazard that borrowers will neglect to pay its credit commitments. Lately, a substantial number of banks have created modern frameworks and models to help investors in aggregation, measuring and overseeing the risk. The outcomes of these models likewise assume progressively essential parts in banks' risk administration and execution estimation forms. In this examination, the goal was to handle the topic of default forecast of here and now credits for a business bank. 924 credit records were utilized from national firms that is allowed by businesses bank from 2003 to 2006. The K-Nearest Neighbor algorithm outcomes show that the best data set is identifying with collection and income and the great order rate is arranged by 88.63 % (for k=3). A bend ROC is plotted to survey the execution of the model. The outcome demonstrates that the AUC (Area Under Curve) basis is in request of 87.4% (for the primary model), 95% (third model) and 95.6% generally advantageous show with income data (Jaime Vitola,2017).

KNN normally handles the conceivable non-linearity of Information takes care of the blame identification issue of gas pipe damage. In conventional blame identification strategies, the recognition procedure of each new test includes all examples in the whole preparing test set. In this way, these strategies can be an intensive computation in observing procedures with a huge number of variables and preparing tests and might be unthinkable for monitoring of real-time. To address this issue, a novel grouping rule for KNN is exhibited. Landmark spectral clustering with low computational multifaceted nature, is utilized io separate the whole preparing test set into a few clusters as well. Further, the productivity of the blame identification techniques can be upgraded by diminishing the quantity of preparing tests associated with the discovery procedure of each test. The execution of the proposed clustering rule of KNN is completely checked in numerical reproductions with both non-linear and linear models and a genuine gas sensor exhibit test framework with various types of issues.

The outcomes of recreations and investigations show that the clustering control of KNN can enormously improve both the precision and productivity of the detection of fault methods and give an amazing answer for reliability and continuous observing of gas pipe damages. The primary utilization of gas sensor clusters was proposed by (Persuade et al 2011) to segregate between the odors(simple). Afterwards, the endless endeavors have been made to address the location of gases with sensor exhibits in numerous fields. The outcomes to date which shows that semiconductor metal oxide gas sensor exhibits can give a particular and special reaction answer for various individual substance gases or gas damage problems. These days, gas sensor clusters are utilized in more building applications. For example, synthetic designing, aviation design and ecological building. To keep up the centralizations of risky gases inside the points of confinement determined by Controls and the real-time monitoring and the ability of reliability is essential for gas sensor clusters or gas damage problems. Moreover, the gas sensors are inclined to failed in light of the idea of sensors of chemicals detecting the material corruption because of irreversible substance responses including poisoning, external effects or sensing the layer ageing that is to electrical interference or power supply instability. Also, the gas sensor cluster dependably faces a more serious risk due to the incorporated repetitive of the k-nearest algorithm. If the faults occur, the broken sensor ought to be distinguished in an opportune way. Therefore, numerous accomplishments which takes places on the detection of faults have been made to manage the failure of the gas damage problems (Florian Nigsch,2006).

Researchers connected the k-closest neighbor (kNN) demonstrating procedure to the forecast of the melting points. An informational index of 4119 assorted natural particles and an extra arrangement of 277 medications were utilized to look at execution in changed districts of substance space. We examined the impact of the quantity of k-nearest algorithms utilizing distinctive kinds of atomic descriptors. To register the prediction which is based on the dissolving temperatures of the k-nearest neighbors. Four distinct techniques were utilized that is an inverse distance weighting, exponential weighting, arithmetic and geometric average, of which the exponential weighting plan yielded the best outcomes. The surveyed model showed means of a Monte Carlo that is 25-fold cross-approval with roughly 30% of the aggregate information as a test set and it will have optimized it utilizing the algorithm of genetic. The KNN expectations for drugs in light of medications which isolate preparing and test sets each taken from the set of data and these data were observed to be impressively better. It is demonstrated that the k-nearest algorithm intrinsically presents an efficient mistake in the prediction of the melting point. A great part of the rest of the mistake can be credited to the absence of data about associations in the fluid state, which are not very much caught by atomic descriptors.

Moreover, the KNN investigate premiums focused on zones of future predictions of stock costs developments which make it demanding and challenging. The Analysts, business networks, and intrigued clients who expect that future event relies upon present and past information. They are quick to recognize the stock value expectation of developments in securities exchanges (Kim, 2003). The money related information is considered as perplexing information to gauge or to predict. The Foreseeing market costs are viewed as problematical, and as clarified in the productive market speculations that was advanced by (Fama, 1990). The EMH is considered as conquering any hindrance between money related data furthermore: the monetary market. it likewise confirms that the vacillations in costs are just an outcome of the recently accessible data; and that all accessible data reflected in showcase costs. The EMH affirm that stocks are by any stretch of the imagination times in harmony and are troublesome for designers to conjecture. Besides, it has been certified that stock costs don't seek after an irregular walk and stock forecast needs more confirmation (Gallagher and Taylor, 2002). Also, different examinations were performed to decide stock value forecasts (Khalid Alkhatib,2013).

Why we should employ this algorithm in our study model? The military and civil structures are helpless and defenseless against harm due to the natural and operational conditions. Along these lines, the usage of the k-nearest algorithm provides a robust solution and decrease operational and upkeep costs. In this sense, the utilization of sensors for all time connected to the structures has shown an extraordinary adaptability and advantage since the examination framework can be mechanized.

This work introduces the depiction of a structural health monitoring framework in light of the utilization of a piezoelectric dynamic framework. The SHM framework incorporates the following: the utilization of a piezoelectric sensor system to energize the structure and gather the estimated dynamic reaction, in a few incitation stages; advanced signal processing technique to characterize the component vectors; lastly; the k-nearest neighbor algorithm as a machine learning way to deal with arrange various types of damages. A depiction of the trial setup, the test approval and a dialog of the outcomes from two distinct structures are included and broke down. The administration life of structures is influenced by a few structures as well for example, the nature of the materials and operational conditions, components, environmental effects and the nature of the working. Hence, it is fundamental to review the structure during its

structural life. The maintenance task and revision may rely upon the sort of structure. Moreover, in a system of automated framework, some regular components are of intrigue, classification, damage detection and localization being probably the most essential. The harm ID unwavering quality is related with the utilization of a dependable sensor networks since the fault occur in the sensors can prompt false encouraging points in the identification process of damage. The damage or sensor fault is ordinarily in light of bad connections, deboning, piezoelectric Fractures: created at the simple snapshot of the establishment of the observing framework or on the other hand during its lifetime. To distinguish these sorts of failures or defects, a few methodologies have been created. Among them, an algorithm of data driven to distinguish the at different temperatures and crystal cuts and the impacts of breaks in the convenience of the signs for basic detection of damage and crystal removals as well (Jingli Yang,2016).

It is conceivable to recognize the risk early and to educate both the researchers and the instructors. While a few colleges have begun to utilize principles based evaluating the risk prediction models have not been adjusted to receive the rewards of guidelines based evaluating in courses that use this reviewing. Many researchers contrasted the predictive methods to recognize the risk at students in a course that utilized models-based evaluating. Just in- semester an execution information that were accessible to the course teachers were utilized in the techniques of predictions. While recognizing the risk, it is essential to limit false negative that is first blunder while not expanding false positive that is the second mistake fundamentally. To build the generalizability of the models and exactness of the predictions, we utilized an element determination strategy to diminish the quantity of factors utilized in each model. The Naive Bayes Classifier show and an Ensemble demonstrate utilizing a grouping of models that is the Naive Bayes Classifier, Support Vector Machine and K-Nearest Neighbors had the best outcomes among the seven modeling methods(tested) (FarshidMarbouti,2016).

1.14 Support Vector Machine Method

Most of the studies researches the profitability of an exchanging methodology in light of preparing a model to distinguish stocks with high or then again low anticipated returns. A tail set is characterized to be a stocks groups whose instability balanced value change is in the most noteworthy or least quintile, for instance the lowest or highest 5%.

Each stock is spoken to by an arrangement of specialized and major highlights figured utilizing CRSP and Composted information. A classifier is tested on future information and prepared on historical tail sets and tried. The classifier is being a (SVM) that is nonlinear help vector machine because of its straightforwardness and viability. The SVM is prepared once per month: keeping in mind the end goal to conform to changing economic situations. The Portfolios are shaped by positioning

stocks utilizing the classifier yield. The highest stocks are utilized for long positions and the least positioned ones for short deals. The Global Industry Classification Standard is utilized to fabricate a model for every segment with the end goal that an aggregate olong-short portfolio for Consumer Staples, Financials, Energy, Health Care, Materials, Industrials, Buyer Discretionary and Information Technology are framed. The information extends from 1981 to 2010. Without estimating exchanging costs, yet utilizing multi day holding periods to limit these, the system prompts yearly overabundance returns of 15% with volatilities and under 8% utilizing the main 25% of the distribution of the stock for preparing long positions and the last 25% for the short ones (Ramon Huerta,2013).

Previous research studied whether there are includes in data of accounting and in authentic value data that will predict the stock cost of changes of organizations. To address this inquiry, the predictive model was prepared on sets of stocks that experience noteworthy value changes. For example, a 5% quintile determination implies that we take those stocks whose positive (negative) and volatility balanced value returns are in the best (base) 5% among all stocks. These 10% of all stocks are utilized towards the nonlinear support vector machine (SVM) to learn relationships between the stock features and the class it has a place with (top or on the other hand base). The huge relationship between future changes of the stock cost and principal and specialized information is a key issue that we explore from which the quantile threshold best captures the significant correlations.

The SVM depends on the auxiliary risk minimization guideline and has another relapse approach with great speculation capacity. It has been effectively connected to issues of finance issues, which are accounted. When displaying of the financial data arrangement utilizing the SVM,the noise occurs in the information and could prompt under-fitting issues or over-fitting. Oil and gas pipeline condition checking is a conceivably extremely difficult process because of shifting temperature conditions, and the harshness of the streaming item and unusual territory. Gas pipe damage can possibly cost a great many dollars' worth of misfortune and also the genuine natural damage that is caused by the spilling item. The proposed procedures will be executed on a lab scale trial fix, then goes to analyze a monitoring system or framework utilizing a variety of sensors deliberately situated on the surface of the pipeline. The Sensors utilized are the piezoelectric ultrasonic sensors. The signal of raw sensor will be first handled utilizing the (DWT) that is Discrete Wavelet Transform and afterward arranged utilizing the intense learning machine called (SVM) that is Support Vector Machines. The Starter tests demonstrate that the sensors can distinguish the presence of the initiated artificially divider. This is done in a steel pipe by arranging the recurrence changes of the spreading signals and the attenuation. The SVM algorithm could recognize the signals in the presence of wall thinning that demonstrates abnormal behavior. At present, a built up type of pipeline examination utilizes the pigs in a procedure called "pigging". Smart pigs travel inside the pipeline recording basic data like structural defects, corrosion levels and cracks utilizing its various sensors. The Pigs can give pinpoint data on the area of abandons utilizing methods like lux leakage of magnetic what's more the ultrasonic recognition. Moreover, utilizing the brilliant pigs in pipeline investigation has a couple of defects (R. RAJKUMAR,2017).

The proposed strategy goes for giving a constant checking framework utilizing a variety of various sensors deliberately situated on the outer surface of the pipeline. The Sensors that are utilized will chiefly be piezoelectric acoustic sensors. The signal of raw sensor will be first prepared utilizing the Discrete Wavelet Transform (DWT) and after that ordered utilizing the great learning calculation called the Support Vector Machines (SVM).

The model is utilized here as an element extraction device to single out any remarkable features in the sensor information. Developed model has some helpful properties: it packs signals and thusly; it has the inclination to dispense with high recurrence noise. The model is utilized here to eliminate the noise in the sensor signals and furthermore to pack a lot of continuous sensor information for the faster processing. The packed information or the model coefficients are then utilized as data sources to the SVM classifier, which will combine the unique sensor information together and afterward perform arrangement.

The SVM has been broadly utilized of late for various applications due to is amazing speculation capacity with little preparing tests. The SVM will be prepared with simulated defect conditions and typical utilizing a trial pipeline fix in the research facility. The electronic, circuit, and systems. The quality of the SVM classifier will then be judged on its precision in deciding the presences of defect in the gas pipeline damage.

The support vector machine is consistently being connected to a developing arrangement of fields, including the sociologies, business, and drug. A few fields introduce issues that are not effortlessly tended to utilizing standard machine learning approaches and, specifically, there is

developing enthusiasm for differential forecast. In this sort of assignment, we are occupied withdelivering a classifier that particularly describes a subgroup of enthusiasm by expanding the distinction in prescient execution for some result between subgroups in a populace. We talk about adjusting most extreme edge classifiers for differential expectation. We initially present various methodologies that don't influence the key properties of most extreme edge classifiers, yet which additionally don't attempt to endeavor to streamline a standard proportion of differential in prediction. We next propose a model that specifically improves a standard measure in this field, the inspire measure. We assess our models on genuine information from two therapeutic applications and show amazing outcomes. These applications lessen to well-known tasks for example, regression or classification. Moreover, there are vital risk that is the issue for the state of art. One such tasks are propelled by the studies and produces differential prediction where one submits two distinct subgroups from some populace to the stimuli. The objective is then to pick up understanding on the distinctive responses by creating, or basically distinguishing, a classifier that exhibits essentially better prescient execution on one subgroup (regularly called the control subgroup over another target sub-group as well.The failure of heart is a dynamic disorder that denotes the end-phase of heart illnesses, and it has a high death rate and huge cost trouble. Specifically, non-adherence of medicine in HF patients may result in genuine outcomes, for example, doctor's facility death and readmission. This examination means to distinguish indicators of prescription adherence in HF patients. In this work, we connected a Support Vector Machine (SVM), a machine-learning strategy valuable for information grouping (Finn Kuusisto,2015).

The Differential forecast has wide and vital applications over a scope of areas. As particular motivation applications, we will think about two restorative assignments. The first task in which we need to explicitly distinguish more established patients with the breast cancer who are great possibility for "attentive pausing" instead of treatment. The other is an undertaking in which we need to explicitly distinguish patients who are most powerless to unfavorable impacts of COX-2 inhibitors: in this manner not recommend such medications for these patients.

The unfavorable medication occasion undertaking alone is of major overall centrality: the noteworthiness of the breast cancer that cannot be exaggerated. Finding a model that is prescient of an antagonistic occasion for individuals on a medication as opposed to not could help in isolating the key causal relationship of the medication to the occasion, and utilizing the support vector machine that is figuring out how to reveal causal connections from observational information is a major subject in momentum look into. Also, finding a model that can distinguish patients with the breast cancer disease that may not be sufficiently debilitating in their lifetime to require treatment could incredibly diminish overtreatment and expenses in human services all in all.

Why we should employ this algorithm in our study model? We exhibit a possibly helpful elective approach in view of the (SVM) that is support vector machine strategies to arrange people with and without normal illnesses. We delineate the strategy to identify people with diabetes and pre-diabetes in a cross-sectional agent test of the U.S. We utilized information from the Nutrition Examination Survey and National Health to create and approve SVM models for two grouping plans. The SVM models were utilized to choose sets of factors that would yield the best grouping of people into these diabetes classes (Finn Kuusisto,2015). In the Classification Scheme I, the arrangement of diabetes-related factors with the best grouping execution included family history, age, race and ethnicity, weight, tallness, midsection circuit, weight list (BMI), and hypertension. For Classification Scheme II, two extra factors - sex and physical action - were incorporated. The discriminative capacities of the SVM models for Classification Schemes I and IIwere 83.5% and 73.2%, respectively. The online instrument Diabetes Classifier was created to exhibit an easy to use application that considers individual or gathering evaluation with a configurable, client characterized limit (Finn Kuusisto,2015).

The SVM algorithm has shown elite in taking care of arrangement issues in numerous biomedical fields, particularly in bioinformatics. As opposed to strategic relapse, which relies upon a pre-decided model to anticipate the event or not of a parallel occasion by fitting information to the curve of logistic, the SVM segregates between two classes by creating a hyper plane that ideally isolates classes after the information have been changed scientifically into a high-dimensional space. Since the SVM approach is information driven and demonstrate free, it might have essential discriminative power for grouping, particularly in situations where test sizes are little and an expansive number of factors are included (high-dimensionality space).

This procedure has as of late been utilized to create robotized grouping of sicknesses and to enhance techniques for identifying illness in the clinical setting.

To test the potential intensity of SVM as an approach for grouping people characterized by illness status, we picked diabetes for instance. In the U.S., diabetes influences an expected 23.6 million individuals, of whom around 33% are unconscious that they have the illness. Another 57 million individuals have pre-diabetes, with raised blood glucose levels that expansion their danger of creating diabetes, coronary illness, and stroke. Ongoing investigations demonstrate that diabetes can be anticipated by way of life changes or pharmacotherapy among people with pre-diabetes. Early screening and finding is in this manner integral to viable anticipation techniques. Various hazard scores and expectation conditions have been produced to distinguish individuals at high danger of creating diabetes or with pre-diabetes in view of regular hazard factors, for example, weight list (BMI) and family history of diabetes (Wei Yu,2015).

For instance, a published risk calculator for the logistic regression utilizes logistic regression to recognize individuals with pre-diabetes and undiscovered diabetes by utilizing mixes of basic hazard factors. Our goal was to create a SVM-based way to deal with recognize individuals with either undiscovered diabetes or pre-diabetes from individuals without both of these conditions. The factors or variables used to produce the SVM models were restricted to basic clinical estimations that don't require research facility tests. The Expectations from this approach were contrasted and the forecasts from logistic regression models containing a similar arrangement of the variable. A last objective was to show the pertinence of the SVM approach by making an exhibition of web-based tool of classification (Wei Yu,2015).

The motivation behind prescient stock value frameworks is to give irregular comes back to money related market administrators and fill in as a reason for risk administration devices. In spite of the fact that the (EMH) that is Efficient Market Hypothesis states that it isn't conceivable to envision showcase developments reliably, the utilization of computationally concentrated frameworks that utilize the support vector machine is progressively basic in the improvement of stock exchanging components. A few investigations, utilizing day by day stock costs, have introduced prescient framework applications prepared on settled periods without thinking about new model updates as well. In this specific circumstance, this examination utilizes a machine learning strategy called Support Vector Machine to anticipate stock costs for extensive and little capitalizations and in three distinct markets, utilizing costs with both day by day and uto-the-minute frequencies. The Expectation blunders are estimated, and the model is contrasted with the arbitrary walk demonstrates proposed by the EMH. The outcomes recommend that the SVM has prescient power, particularly when utilizing a procedure of refreshing the model intermittently. There are likewise demonstrative outcomes of the expanded forecasts accuracy during the bring down instability periods (Dino Isa,2009).

The value forecast instruments are principal to the arrangement of the systems of investment and the improvement of hazard administration models. The Efficient Market Hypothesis (EMH), moreover, states that it isn't conceivable to reliably acquire the adjusted risk returns over the productivity of the market all in all. Computational advances have prompted a few machine learning algorithms used to anticipate the advertise developments reliably and hence gauge future resource values for example, organization stock costs. The Models in view of the Support Vector Machine (SVM) are among the most generally utilized procedures (Dino Isa,2009).

1.15 Logistic Regression Method

A continuous issue in evaluating logistic regression models is a very kind of failure of the probability expansion algorithm to join. Much of the time, this failure is a result of data designs known as quasi-complete or complete. For these examples, the greatest probability evaluates just don't exist. Anybody with much down to the practical experience encounter utilizing logistic regression will have apparently experienced issues with the convergence as well. Such issues are generally both exasperating and puzzling. Most analysts have no idea as to why certain models and certain informational collections prompt combination challenges. Also, for the individuals who do comprehend the reasons for the issue, usually indistinct whether and how the issue can be settled. A typical issue to maximize the function is the nearness of nearby maxima. Luckily, such issues can't happen with the logistic regressions just because that the log-probability is all around concave as well, implying the function can have at most one greatest. Sadly, there are numerous circumstances in which the likelihood function has no most extreme maximum, in which case we say that maximum likelihood does not exist.So, the general guideline is obvious: Whenever there is a zero in any cell of a 2 × 2 table, the most extreme likelihood gauge of the slope of logistic coefficient does not exist. This rule additionally, stretches out to multiple variables of multiple. For any dichotomous autonomous variable in a logistic regression, if there is a zero in the 2 × 2 table shaped by that dependent variable and variable, the regression coefficient of the ML evaluate does not exist.

This is by a wide margin the most well-known reason for in logistic regression of the convergence failure. Clearly, it will probably happen when the size of sample is little. Indeed,even in the sample of large, it will often happen when there are outrageous parts on the recurrence dissemination of either the reliant or free factors. Consider, for instance, a logistic regression foreseeing whether a man has some allergy or disease whose general commonness is under 1 of every 1000. Assume further, that the variable of explanatory incorporates an arrangement of 7 dummy factors speaking to various age classes. Regardless of whether the example contained 20,000 cases, we could sensibly expect that for no less than one of the classes, nobody would have that disease.

This research also focuses on predicting the effect on the likelihood of failure of adding hydrogen to the gaseous petrol dissemination network. Hydrogen has been shown to change the behavior of split like imperfections which may influence the security of pipeline or make it costly to work. A device has been created to evaluate the failure of the gas pipeline because of the presence of split lie surrenders including the operational parts of the pipeline. for example, investigation and repair systems. With different parameters (i.e., split sizes, material properties and the internal pressures), a reliability quality investigation in light of failure appraisal outline is performed through direct Monte Carlo reproduction. Examination and repair strategies are incorporated into the simulation to empower reasonable pipeline situations for maintenance. In the information arrangement process, the exactness of the probabilistic meaning of the vulnerabilities is critical as the results are exceptionally delicate to specific factors: for example, the break profundity, length and split development rate. The failureprobabilities of each imperfection and the entire pipeline framework can be acquired during the simulation. Diverse examination and repair criteria are accessible in the Monte Carlo simulation process whereby an ideal maintenance procedure can be acquired by looking at the combinations of assessment and repair strategies (Hitinder S. Gurm,2014).

This simulation gives not just date on the probability of failure, but also the anticipated number of repairs required over the gas pipeline life along these lines giving information appropriate to financial models of the gas pipeline administration. This instrument can be likewise used to fulfill certain objective dependability necessity. An illustration is displayed contrasting a petroleum gas pipeline and a pipeline containing hydrogen. The failure of an auxiliary part, for example, a pipeline may happen when a break spread in the not stable way to cause hole or blast of the pipeline. Break mechanics joined with a probabilistic approach has been used in numerous fields of investigation including critical auxiliary segments such as nuclear piping and pressure vessels. In view of probabilistic crack mechanics, the methods of statically are connected keeping in mind the end goal to survey the reliability and quality of pipeline containing break like defects. In other words, to give a solitary number which speaks to the likelihood that a pipeline could come up short. The point of the naturally venture is to research the likelihood of utilizing the current gaseous petrol pipelines to convey hydrogen or blended common hydrogen gas. As zero resistance to hydrogen spillage is generally acknowledged in the pipeline business the idea of 'damage' of a pipeline incorporates the two gas spillage and pipeline breakage (Hitinder S. Gurm,2014).

Factual examinations of the pipe-related data(incident) for gas transmission pipelines damages somewhere in the range of 2002 and 2013 gathered by the Pipeline and Hazardous Material Safety Administration (PHMSA) of the United States

Department of Transportation (Spot) are directed. It is discovered that the aggregate length of the inland gas transmission pipelines in the US is around 480,000 km starting at 2013(PHMSA website). Outside material failure, interference and external corrosion and an inner corrosion are the main foundations for the pipe-line incidents: accounts for75% of the incidents somewhere in the range of 2002 and 2013. In view of the pipeline mileage and a data(incident), the normal rate of burst incidents over the 12-year time frame somewhere in the range of 2002 and 2013 is figured to be 3.1 × 10-5 for every km for each year (PHMSA website).

A logistic model is created to assess the POI that is the probability of ignition given a crack of an inland gas transmission pipeline utilizing the greatest probability technique in view of an aggregate of 188 burst incident that takes place somewhere in the range of 2002 and 2014: gathered from the PHMSA pipeline incident database. The result of the pipeline pressure of internal at the time of outside diameter squared and time of incident is seen to be unequivocally associated to POI while the area class of the pipeline is not. The 95% certainty interim is assessed, and for functional designing utilize, the 95% upper certainty bound is classified in a look-into table. The proposed demonstrate is additionally approved utilizing an autonomous dataset announced in the writing.

The logistic regression displaying is utilized to figure out what factors are related with nonzero item misfortune cost, nonzero property harm cost and nonzero cleanup and recuperation costs. The elements inspected incorporate the framework part associated with the accidents, area attributes which occurs in a high outcome region or areas and whether there was liquid spillliquid ignition and an explosion. For these accidents, related with nonzero values for these outcome measures (weighted) minimum squares regression is utilized to comprehend the variables related to them, and additionally how the distinctive starting reasons for the accidents are related with the outcome measures (Sandro Sperandei, 2012).

The regression model is then used to develop illustrative situations for perilous fluid gas pipeline damages. These situations propose that the size of outcome estimates. For example, property damage, product lost, and cleanup and recuperation costs are profoundly subject to other accidents qualities. The regressions model used to develop these situations constitute an expository instrument that industry chiefs can use to evaluate the conceivable outcomes of accidents in these pipeline frameworks by cause (and different qualities). Moreover,to allocate the resources for maintenance and to decrease the chance factors in these frameworks (Sandro Sperandei,2012).

Why we should employ this algorithm in our study model? The logistic regression is utilized to get chances proportion within the sight of in excess of one informative variable. The strategy is similar to multiple linear regression; with the exemption that the reaction variable is binomial. The outcome is the effect of every factor on the odds proportion of the watched occasion of intrigue. The biggest advantage to use this algorithm in our study model is to abstain from jumbling impacts by examining the relationship of all factors together. In this article, we clarify the logistic regression method which utilizing cases to make it as straightforward as could be expected under the circumstances. After meaning of the system, the fundamental elucidation of the outcomes is featured and afterward some unique issues are examined.

The logistic regression works fundamentally as the same as response of binomial, however with a binomial reaction variable. The best preferred standpoint when contrasted with Mantel-Haenszel OR is the way that you can utilize consistent illustrative factors and it is less demanding to deal with in excess of two variables of explanatory. This last trademark is basic when we are occupied with the effect of different illustrative factors on the reaction variable. If, that we look up at numerous variables independently, we disregard the covariance among variables and are subjected to an effect of confounding, as was exhibited in the case above when the impact of treatment on death likelihood was incompletely covered up by the impact of age (Zhang,2014).

The logistic regression is a paired classifier that can be utilized as a parallel classifier and consequently: the model can utilize the standard measurements for classifiers. The measurements you utilize are the standard ones, logistic regression being the most total (however the less instinctive regarding meaning). An accuracy for example, is the level of focuses that have been effectively grouped. UG gas pipe model appears not terrible by any means. An irregular classifier that is your model predicting random guesses . The precision relies upon the dataset. In a simple dataset numerous fundamental models would have the capacity to get comparable (or best) values than our UG Gas pipe model can contrast your model with others to see whether your model gets remarkable outcomes or it is only that the order undertaking was simple.

To finish the assessment, you can likewise plot a confusion matrix to see whether your model tends to complete a great job in ordering one class however not as great in distinguishing individuals from alternate class. In a confusion framework, you need the vast majority of the components to fall in the inclining (genuine positive and genuine negatives).

The other two squares in a confusion matrix of a double classifier that are the false negative and false positives. Plus, you ought to do this in a test set, a subset of your information that your model has never observed that will be, that you didn't use to tune the coefficients of your regression as well (Chio Lam,2012).Many predictive modeling techniques including neural systems (NNs), grouping, thel ogistic regression model, and affiliation rules, exist to help make an interpretation of this information into understanding and esteem. They do that by learning designs covered up in substantial volumes of verifiable information. When learning is finished, the outcome is a prescient model. After a model is approved, it is esteemed ready to sum up the information it learned and apply that to another circumstance. Given that prescient demonstrating strategies can gain from the past to foresee the future, they are being connected to a bunch of issues, for example, recommender frameworks, extortion and mishandle recognition, and the prevention of accidents and diseases. The accessibility of enormous data and cost-proficient handling power is growing the materialness of prescient information driven procedures in various businesses. In doing that, the clever mathematics is helping an ever-increasing number of organizations understand the genuine potential covered up in their information (Carlos E,2008).

1.16 Gap in Existing and Prior Research

Previous studies discussed the damages to the underground pipeline in terms of location including locating the pipe by (GPR, GIS), locating by penetration Radar, and data visualization, and also in terms of excavation including 3 D excavation, using Camera, 3 D Operator (Figure 22). However, none of the existing studies has considered studying the information process flow chart used by contractors to request damage tickets to identify weak nodes, potential risk vulnerability and to assess other risk involved in the digging process.

There are many causes of gas pipe damages were identified by prior studies, and yet, there are many ways that damage can be avoided on gas pipe networks. Furthermore, there has been risk models developed to study the risk involved in the gas pipe damage such as causes of miss locating the gas pipe, and the causes of excavator damage the gas the pipe. However, none of the prior researches has considered studying a risk model based on some derived risk attributes from underground facility operator, the board designed one call system, and locating party. This research will develop a risk model to predict future risks involved on the information flow process, which will help utilities agencies prioritize the risk involved in the information flow processes, and focus on risk involved with each process.

The literature review map (Figure 22) shows ticket request process of gas pipe

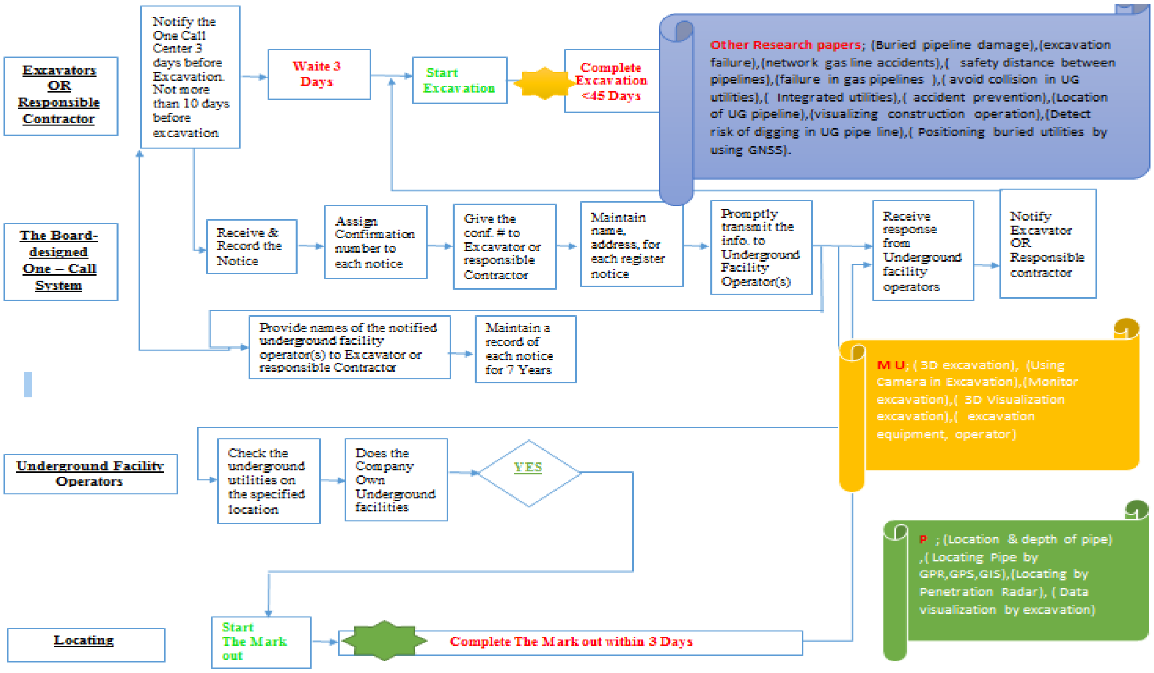

Excavators OR Responsible Contractor
Notify the One Call Center 3 days before Excavation. Not more than 10 days before excavation
Waite 3 Days
Start Excavation
Complete Excavation <45 Days
Other Research papers; (Buried pipeline damage),(excavation failure),(network gas line accidents),(safety distance between pipelines),(failure in gas pipelines),(avoid collision in UG utilities),(Integrated utilities),(accident prevention),(Location of UG pipeline),(visualizing construction operation),(Detect risk of digging in UG pipe line),(Positioning buried utilities by using GNSS).
The Board-designed One – Call System
Receive & Record the Notice
Assign Confirmation number to each notice
Give the conf. # to Excavator or responsible Contractor
Maintain name, address, for each register notice
Promptly transmit the info. to Underground Facility Operator(s)
Receive response from Underground facility operators
Notify Excavator OR Responsible contractor
Provide names of the notified underground facility operator(s) to Excavator or responsible Contractor
Maintain a record of each notice for 7 Years
M U; (3D excavation), (Using Camera in Excavation),(Monitor excavation),(3D Visualization excavation),(excavation equipment, operator)
Underground Facility Operators
Check the underground utilities on the specified location
Does the Company Own Underground facilities
YES
P ; (Location & depth of pipe) ,(Locating Pipe by GPR,GPS,GIS),(Locating by Penetration Radar), (Data visualization by excavation)
Locating
Start The Mark out
Complete The Mark out within 3 Days

Figure 22 Shows the Literature Review Map & the Gap in Existing Research.

damage passing through excavator initiating the request, then it goes to receive and record the notice, assign confirmation number, and follows through all processes.

In addition, the map detail out the processes of the Board designed One Call System, underground operator, and locating party. As shown on Figure 22 most the conducted research addressed mainly the excavation, and mark out risk factors only of gas pipe damage. However, risk assessment and potential impact of other risk factors involved in the information flow process chart has not been covered yet. In this research, we will address the risk factors related to the Board designed One-call System, underground facility operator, and locating party and address their potential impact on damaging the underground gas pipe (Figure 22).

Previous research focused on risk assessment of natural gas pipeline network problems from qualitative and quantitative perspectives based on the conventional risk analysis methods (Fault Tree, Event Tree) and in addition to other methodologies (Petri Network and Bayesian Networks) (Cagno et al., 2000; Dong and Yu, 2005; Markowski and Mannan, 2009; Han and Weng, 2010, 2011; Baksh et al., 2015; Guo et al., 2016; Li et al., 2016; Kabir et al., 2016). However, all these researches have limitations. First, they did not use the ticket request process information flow in diagnosing and analyzing the underground pipeline damage. Secondly, most of these reaches are based on probability which does not include the uncertainties when analyzing the risk which leave clear gap in the results.

There are many potential risks involved in the digging and locating process for the underground ground pipeline (Cooke & Jager, 1998). The same goes for exchanging data between parties. Yet, there are many other ways which can cause the pipe damage.

However, there has been no significant research on studying all the information flow through request initiation process chart starting from excavator notifying the one call center, the board designed one –call systems receive the request, Underground facility operator verify the location and existing utilities in the particular site, locating the pipe and mark out, and finally start excavation (Figure 23).

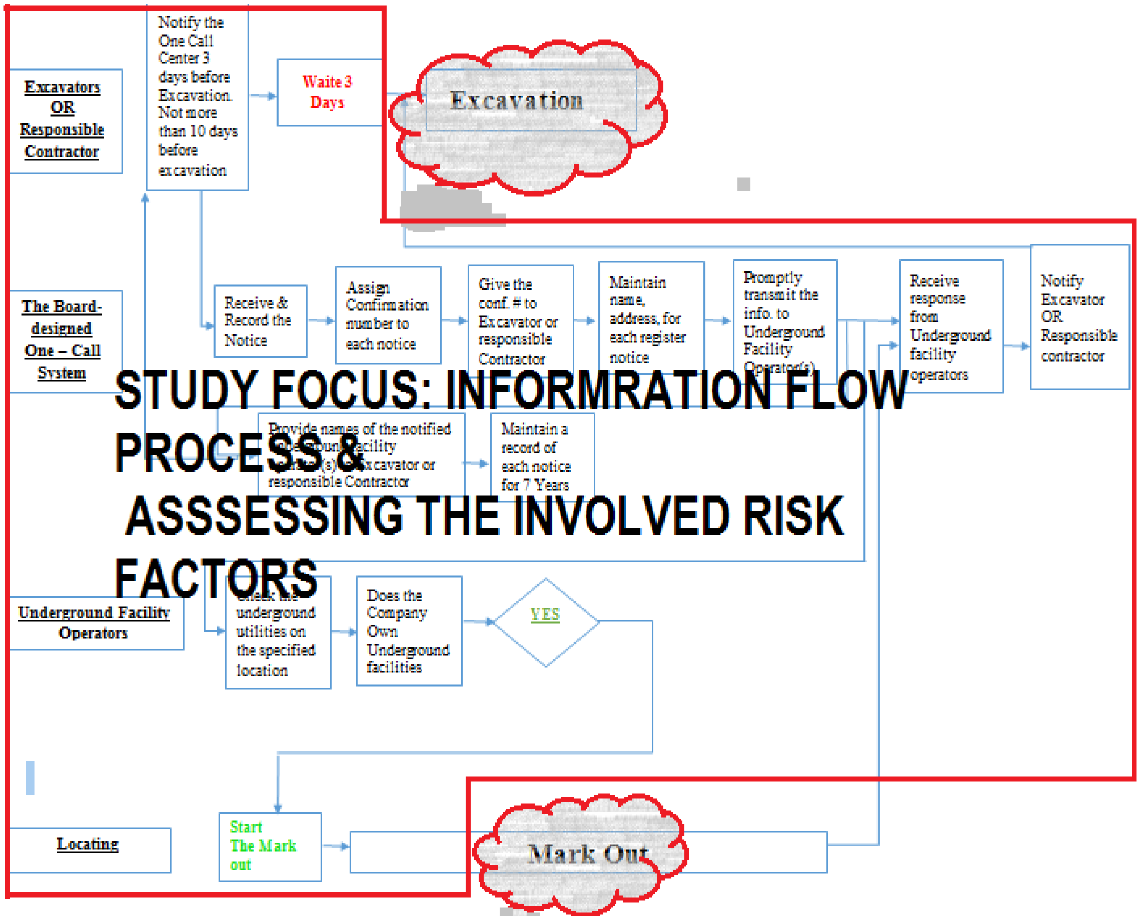

Excavators
OR
Responsible
Contractor
Notify the One Call Center 3 days before Excavation. Not more than 10 days before excavation
Waite 3 Days
Excavation
The Board-designed One – Call System
Receive & Record the Notice
Assign Confirmation number to each notice
Give the conf. # to Excavator or responsible Contractor
Maintain name, address, for each register notice
Promptly transmit the info. to Underground Facility Operator(s)
Receive response from Underground facility operators
Notify Excavator OR Responsible contractor
STUDY FOCUS: INFORMRATION FLOW PROCESS &
ASSSESSING THE INVOLVED RISK FACTORS
Provide names of the notified underground facility operator(s) to Excavator or responsible Contractor
Maintain a record of each notice for 7 Years
Underground Facility Operators
Check the underground utilities on the specified location
Does the Company Own Underground facilities
YES
Locating
Start The Mark out
Mark Out

Figure 23Study Focus Map

The proposed research will fill the gap on all prior researches in underground gas pipe damage prevention. More specifically, gas pipe damages and the involved risk throughout the digging process contributed to 1350 pipe damages on 2009 by contractor, and 1513 gas pipe damages on year 2013. In addition, number value of each categorical breakdown on a per year basis. 2009 to 2010 showed a slight increase in contractor type. 1,656 records in 2009 increased to 1,727 records in 2010, which equates to a 4.29% increase as per Figure 18. Furthermore, both Homeowner and Utility groups decreased by approximately 27%, while the contractor group, which is the largest, decreased the least at 5.5%. The gradual decrease is interrupted in 2013 where records skyrocketed from 1,555 in the previous year to 1,840. Between these years, an 18.33% increase occurred where all contractor types increased other than the Utility group. Thus, a developed risk model is needed to predict the future potential risk factors causing all the damage mentioned above Figure 24.

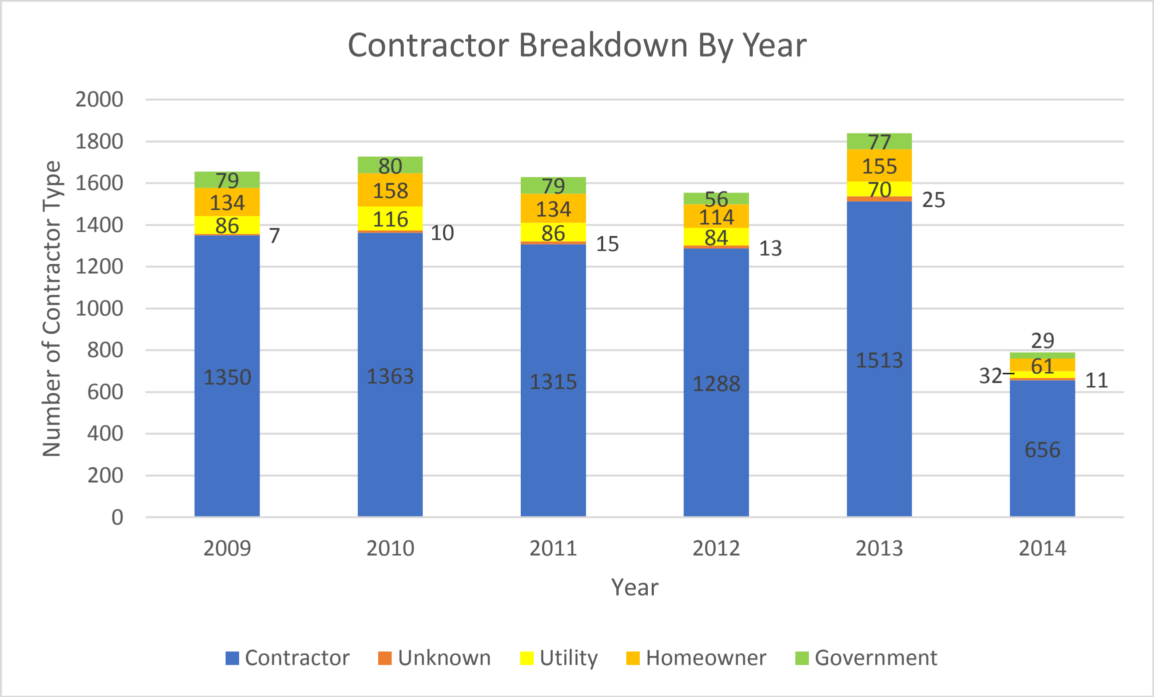


Figure 24 Yearly contractor breakdown

CHAPTER THREE: PROPOSED METHODOLOGY

1.17 Methodology Overview

This chapter provides an outline of research methods used in the study. This chapter discusses in detail the research methodology that has been adopted in this study. This methodology involves ten steps. It starts with defining the system and collecting the data set which includes organizing the collected data in categorized attributes and compile them in tables. In addition, this step includes exploration of data, and data set processing. Second, identifying the risk process by using Bow-tie method. Third, Mapping out underground gas pipe damage network by using Fault tree method, then involving machine learning algorithms such as Logistic Regression, Support vector machine, Random Forest, and KNN. Finally, processing UG gas pipe damage network with Bayesian network and defining the nodes probability. The tasks involved in the methodology are delineated in Figure 25.

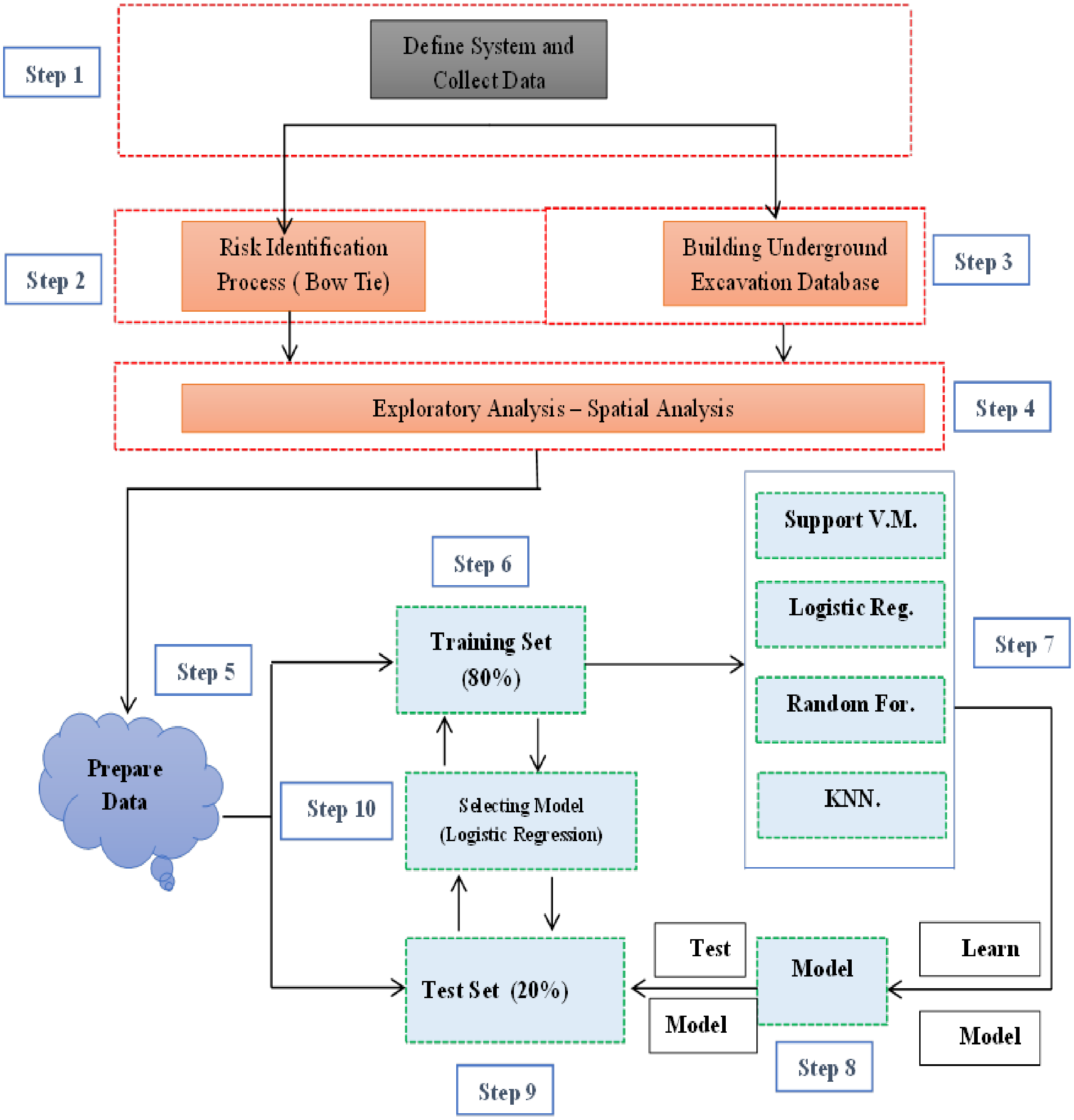

Step 1
Define System and Collect Data
Step 2
Risk Identification Process (Bow Tie)
Building Underground Excavation Database
Step 3
Exploratory Analysis – Spatial Analysis
Step 4
Support V.M.
Logistic Reg.
Random For.
KNN.
Step 7
Step 6
Training Set (80%)
Step 5
Prepare Data
Step 10
Selecting Model (Logistic Regression)
Test Set (20%)
Test
Model
Model
Learn
Model
Step 8
Step 9

Figure 25 Overview of Components of Methodology in Predicting Damages in UG Gas pipes

1.18 Collecting the data & Defining the System

<u>STEP 1: Defining the system and collecting data</u>: the data used in this research was both for damaged, and undamaged UG gas pipes. Most of the undamaged data contains attributes such as ticket number, date of incident, and address of the incident. In the data structure, damaged UG gas pipe has attributes such as excavator type, and type of request. The data,organized into years 2010-2014, were categorized by ticket number and address to be used later.

STEP 2: Risk evolution process of underground gas pipe damage modeling with Bow-tie: This is a widely used graphical process for damage modeling. Bow-tie process can present a complete accidental scenario starting from the causes and ends with the consequence. The Bow-tie method was selected for UG gas pipe risk assessment because it can identify where resources should be focused for risk reduction, i.e. prevention or mitigation. Bow-tie includes two parts, the left of bow-tie is a FT that describes the latent causes for an initial event, the right of bow-tie is an ET which describes the sequential failure of damage preventive barrier and presents the evolution process from initial event to final latent consequence. The FT and ET is linked through a pivot node that is the top event of FT and the induced event of ET.

1.19 Building Excavation Database, and Perform Exploratory Analysis

STEP 3: Building Underground Excavation Database: Excel and Python was used in this research to combine all the attributes of the UG gas pipe damages in columns, and rows. This step was done for many reasons. First, duplicate data can be flagged out and easily removed to get better results. Second, understating the data trend: By organizing the data in one database, the trend of the data attributes can be easily specified by performing preliminary analysis which can be used in the future steps of the research. Third, cleaning the data: the data were received as text files which was not useful. Thus by transferring the data into excel spreadsheet, it was possible to clean the data, and prepare it to be useful for the research.

<u>STEP 4: Exploratory and Spatial Analysis:</u> We will use Exploratory and Spatial analysis to extract the latent risk factors. More specifically, clustering analysis will be used. We will group the data to groups containing similar attributes such as depth, pipe size, year, zip code, number of outgoing calls…etc. The goal is to observe the characteristics of each cluster and to focus on particular set of clusters for further analysis Figure 26. Rapid miner will be used to further study different clusters patterns of the damaged data.

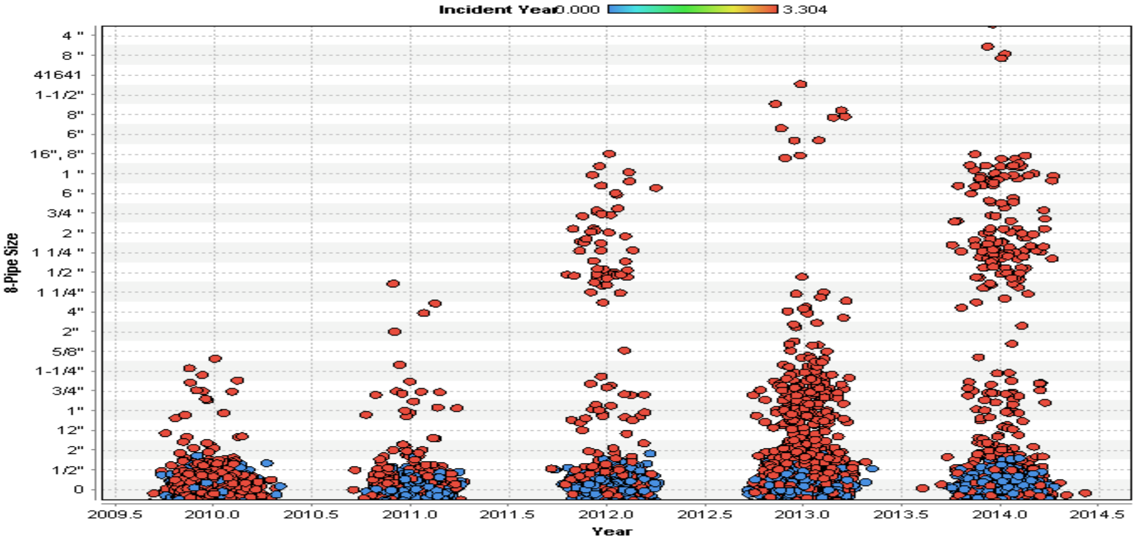


Figure 26: Cause of Damage between Pipe Size & Year (Phase 2)

In addition, Hot Spot analysis will be performed to identify the damage hotspots. We will be using Arc GIS to plot the data and to generate hot spots which will enable us to focus on the most affected areas. Then we will analyze the attributes of the damaged data at certain areas at the Hot spots to identify the risk factors Figure 27.

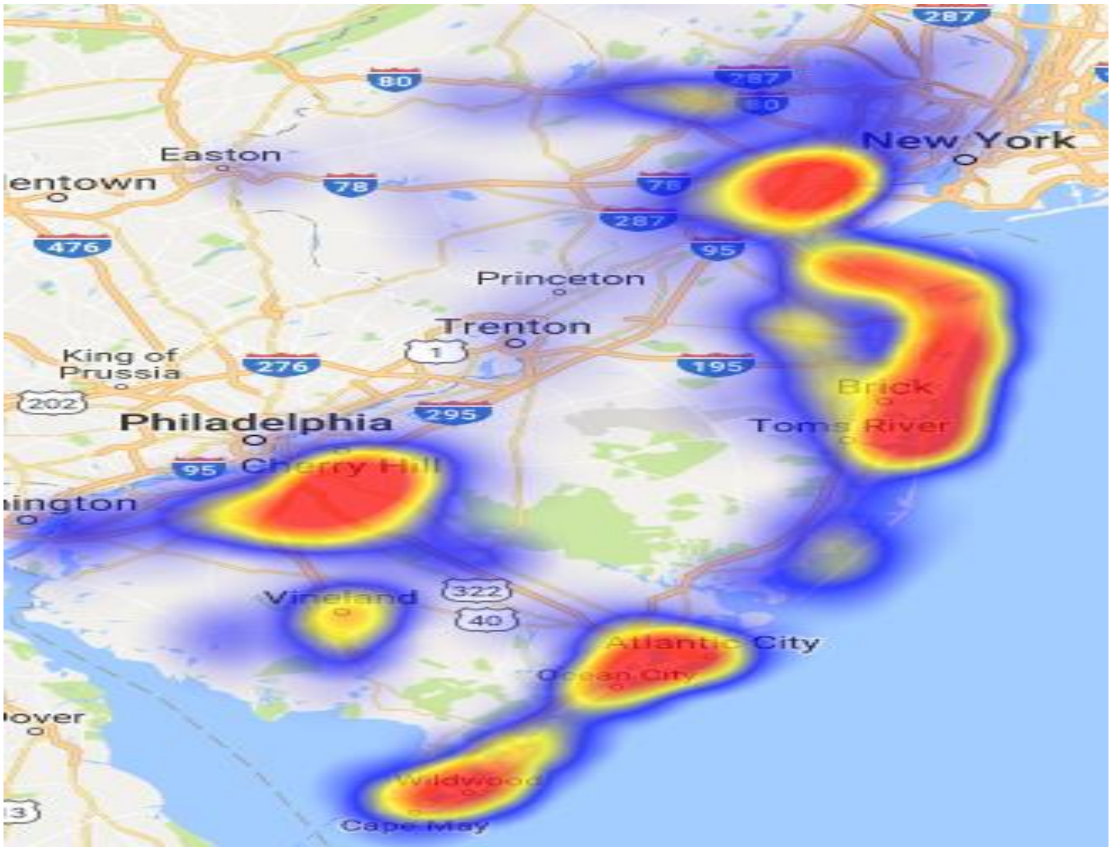


Figure 27: Hot Spot map of damaged data

<u>STEP 5: Prepare the Data of The UG Gas pipe:</u> Several steps were followed to better prepare the data to be used in the Predictive model and risk factors identification. Since a large portion of the data set contained missing values, it was important to focus on correcting these instances. In addition, during data testing it was evident that some gas distributors kept extra records. Often there were variables that were recorded by a single company Therefore, the damaged data was missing some attributes, the undamaged data was missing most of the attributes. , The first data preprocessing step eliminated attributes that were deemed unnecessary for further computations. Eliminating attributes cleaned the data set up and assisted with the removal of a large number of missing values.

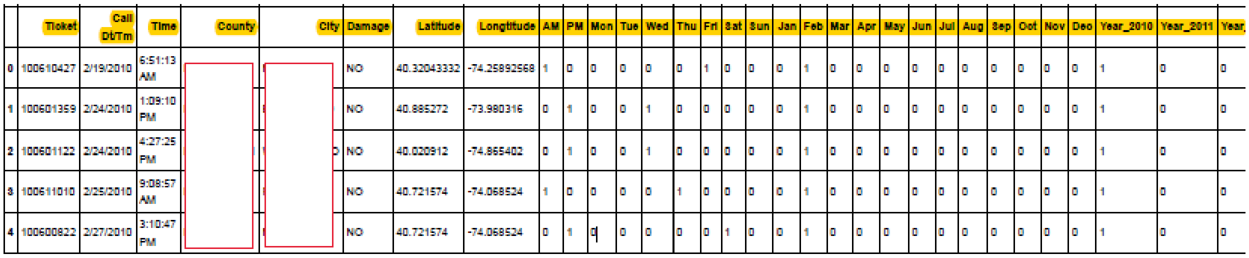
Furthermore, eliminating attributes actively reduced the dimensionality of the data set and allowed for increased model performance with a reduced computation time.

Preparing the data will be explained in more details in the chapter of preparing the data, and perform exploratory analysis. That will cover all the steps of data preprocessing including, incomplete data, Geo-coding the data, missing data, organizing the data, converting the text data into columns, rows in excel, data integration, cross validation of the UG gas pipe data. The steps also include preparing the data maps, plotting the data, preparing the data attributes for damaged data, preparing the data attributes for undamaged data, and deriving new attributes from the data.

1.20 Building The Predictive Model

STEP 6: Training The Data Set (80 %) : The data were split into 80 % training, and 20 % testing. The training data set was chosen to be large enough to yield meaningful results, and is representative of the data set as a whole including all selected data feature or attributes. As part of this step, data preprocessing was conducted to make sure we have quality data and meaningful attributes to yield meaningful results. In addition, Python Anaconda was selected to be used in the analysis of the data. Python is a popular free source tool with many add-on libraries.

Jupyter Note book was used as a tool as well. First, the operating system was imported to be used in the analysis. Python contains many Libraries, Library name called Panda is needed to perform the analysis. In order to use Pandas in Jupyter Notebook (Jupyter come with Anaconda by default). Importing a library means loading it into the memory and then it’s there ready for use. In order to import Pandas, we ran import code. Then the data was saved in CSV file, given a name " Log" in coding. We ran a code to import the data. Next, the data was displayed in Python to view all columns to make sure all the required attributes of the data was imported, and no missing columns see figure 28.



	Ticket	Call Dt/Tm	Time	County	City	Damage	Latitude	Longtitude	AM	PM	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Year_2010	Year_2011	Year_
0	100610427	2/19/2010	6:51:13 AM	[redacted]	[redacted]	NO	40.32043332	-74.25892568	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0
1	100601359	2/24/2010	1:09:10 PM	[redacted]	[redacted]	NO	40.885272	-73.980316	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0
2	100601122	2/24/2010	4:27:25 PM	[redacted]	[redacted]	NO	40.020912	-74.865402	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0
3	100611010	2/25/2010	9:08:57 AM	[redacted]	[redacted]	NO	40.721574	-74.068524	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0
4	100600822	2/27/2010	3:10:47 PM	[redacted]	[redacted]	NO	40.721574	-74.068524	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0

Figure 28Show displayed data in Jupyter Note

All the input data was defined in Anaconda by using codes. Next we checked how many null values there were in the inserted data. Python treats null values as NAN or NAV(not available number or Value), consider it a text. and the model cannot run a algorithms like Logistic Regression containing NAN. Data Shape code was run to view the shape of the data, and a gain making sure we have all attributes we planned to have. Next, Damage, Undamaged was defined by running code, (Damage = "Yes = 1"), (Undamaged = "NO = 0"). Then doubled checked some of the attributes, for instance, there were 851 cities, and 21 counties which same as what we have in CSV file. In addition, Code was run to drop season, Latitude, Longitude because they are objects. More Libraries we imported into Python, Seaborn Library, matplotlib, and Sklearn. Standard deviation was performed to make sure all data was normally distributed.

STEP 7: Selecting training Algorithm Model : Four algorithms were selected to be used in the training model, Logistic Regression Method , k-nearest neighbors Method, Support vector machine Method, and Random Forest Method. Logistic Regression analysis was selected because it indicates the strength of impact of multiple independent variables (data attributes) on a dependent variable (Damage/ Not). In addition, Logistic Regression indicates the significant relationships between dependent variable (Target (YES/NO)) and (Predictors (data Attributes)) independent variable. k-nearest neighbors algorithm was selected because K nearest neighbors stores all available cases (Predictors) and classifies new cases based on a similarity measure (e.g., distance functions). In addition, KNN works as pattern recognition through specific distance for each neighborhood, then it classifies the data based ion the distances between the generated neighborhoods. Random Forest Method was selected because it builds multiple decision trees for the data attributes Time, Week days, Months, and other attributes and merges them together to get a more accurate and stable prediction. More specifically, it creates a forest of data attributes, and makes random selection. In addition Random Forest used because it will be used for both regression and classification tasks. Random forest also makes it easy to view the relative importance it assigns to the input Attribute. Support Vector Machine was used because it classify data: based on either a priori information or statistical data mined from raw data set.S. V Machine provides pattern recognition and makes it an extremely powerful tool for UG gas pipe damage data separation.

STEP 8: Run the training Model : The process of training a predictive model involves providing PM algorithm (that is, the learning algorithm) with training data. The term PM model refers to the model artifact that is created by the training process. The provided training data to the algorithm contains the correct answers: known as a target or target attribute. The learning algorithm finds patterns in the training data that map the input data attributes to the target (the answer that we want to predict): it outputs an PM model that captures these patterns. We will use the model to get predictions on new data for which the model do not know the target. In our study we provided the algorithm with 40 attributes called predictors, and the Target which is YES or No, damage or undamaged. The data was divided to X, Y , where is X is equal to all attributes, and Y is equal to Damage / Target Variables. The data split was 80 % of the data into training, and 20% into testing. Moreover, 70% of the data is in X-train, Y-train. Then, the classifier was chosen to be either Logistic regression, KNN, Random Forest, Or Support Vector Machine. Then the equation, CLF.fit(x-train, Y-train) finds the patterns in all the attributes with respect to Y- train, and save the patterns. Results, the algorithm, found the pattern in the training data. Next, the model uses the pattern found from the training data which maps all the attributes to target value which is damage (YES/NO). Using the pattern found above combined with another function called "PREDICT". Equation Data-Predict = CCF. Predict (X – test), Data – Predict = = (Y – Test) Actual. Finally, we score the model or test the model by comparing Data-Predict Vs Y- Test.

1.21 Testing, and Validation

STEP 9: Testing, and validation (20 %): 20 % of the data was used in testing the predictive model. The metrics used to evaluate the model were Confusion Matrix, Recall, and Precision. These three Metrics were conducted to test and validate all of the outcomes of the algorithms including Logistic Regression, Random Forest, KNN, and Bayesian. Results, Logistic Regression Model gave the best result according to Confusion Matrix, Recall, precision. Sample was used for illustration purposes.

A confusion matrix is a table that describes the performance of a classification models we employed in our study (Bayesian, Logistic Regression, KNN, and random Forest) on a set of test UG gas pipe data for which the true values. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing. Therefore, these the definitions regarding confusion matrix, true positives (TP): These are cases in which we predicted yes (they have the UG gas pipe damage), and they do have the damage. true negatives (TN): means the model predicted no, and they don't have the gas pipe damage. false positives (FP): means the model predicted yes, but they don't actually have the damage. False negatives (FN): means the model predicted no, but they actually do have the damage. Accuracy: We used accuracy measure in our study so it tell us how often is the classifier correct? Equation = (TP+TN)/total. Precession : We used accuracy measure in our study so it tell us when it predicts yes, how often is it correct? Equation =TP/predicted yes.

STEP 10: Selecting The Model: The Logistic Regression was selected as model to predict the future gas pipe damages and the dominant risk factors contributing to the damage of the UG gas pipe damages. The confusion matrix was produced for all algorithms, including, Logistic Regression, S.V Machine, KNN, and Random Forest. The Logistic Regression demonstrated a precision level equal to 0.89 and accuracy equal to 0.98: which is the highest as compared to the other algorithms.

CHAPTER FOUR: PREPARING THE DATA, And PERFORM EXPLORATORY ANALYSIS

1.22 Description of Data Sets

The data used in this research includes excavation requests. Most of them did not result in damage to the underground facilities. Small portion of the excavation requests caused damage to the underground facilities. In addition, the data were received in the form of raw data: shown in Figure 29 and Figure 30. The analysis will be conducted on a period from years 2010 to mid-2013 and the data of year 2014 will be used to test and validate the results. The data set includes both damaged and undamaged UG gas pipes. All the received data was updated progressively as records were made available. In total, there were 2 Million records provided and used within the analysis procedures.

The raw data was processed and organized in a way where rows represented individual incidents and whose columns represented provided attributes. For damaged date, 20 attributes were labeled for damaged data of UG gas pipes as follows:

1) Incident ID
2) Incident Date
3) Underground Facility Operator
4) Damage Address
5) Damage City
6) Damage Zip
7) Damage County
8) Excavator Name
9) Excavator Type
10) Excavator Address
11) Excavator City
12) Excavator State
13) Excavator Zip
14) Locate Ticket Number
15) Locate Ticket Type

16) Locate Performed

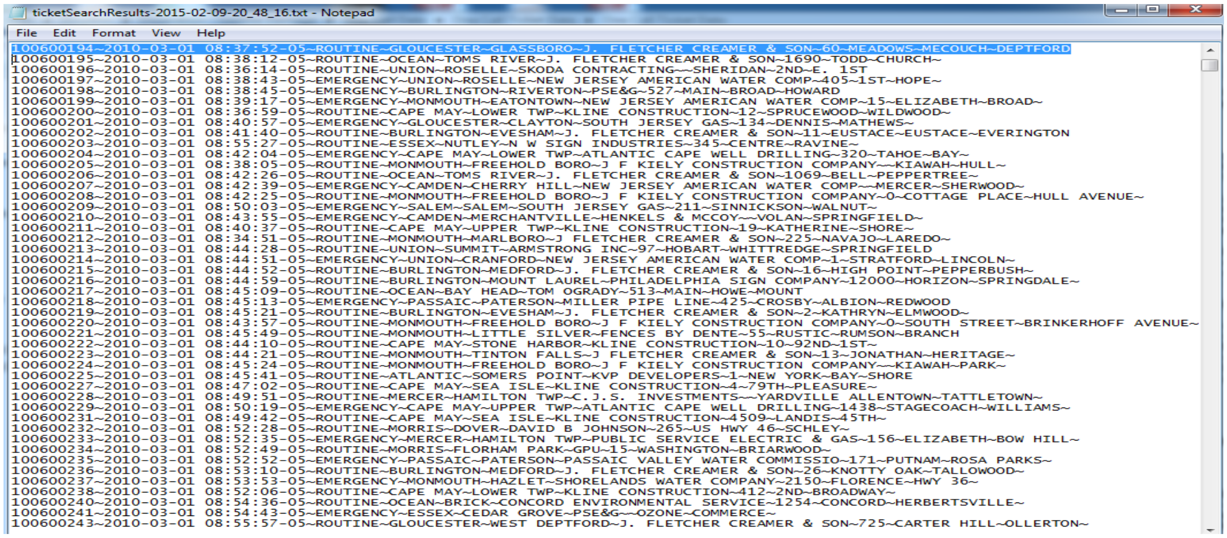
17) Damage Description

18) Pipe Material

19) Pipe Size

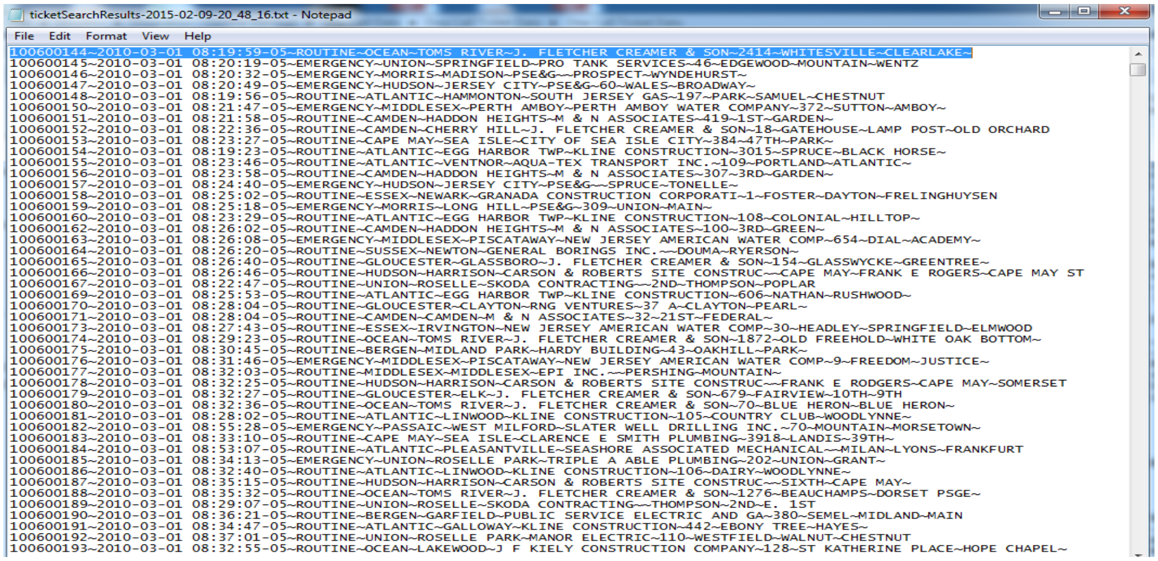
20) Cause of Damage

This is a large data set with approximately 2 Million entries. Unfortunately, some attributes within the data set were quite sparse because of the varying recording procedures among the four gas distributors. For the most part, important information was recorded; however, missing values are still noticeable within core data. Most of the variance in recorded values depends on the responsible party for updating records in master data sheets. While the recording system is most effective if all values are known, excavator names are not always known: because work was performed without requesting locate services and resulting damages were discovered after-the-fact. Another reoccurring error pertains to the accurate entry of data values within the master sheet. Frequently entries were recorded incorrectly into the wrong data cell; this caused a loss of information since the incorrectly recorded instances occupied cells of other required



```
100600194~2010-03-01 08:37:52-05~ROUTINE~GLOUCESTER~GLASSBORO~J. FLETCHER CREAMER & SON~60~MEADOWS~MECOUCH~DEPTFORD
100600195~2010-03-01 08:38:12-05~ROUTINE~OCEAN~TOMS RIVER~J. FLETCHER CREAMER & SON~1690~TODD~CHURCH~
100600196~2010-03-01 08:36:14-05~ROUTINE~UNION~ROSELLE~SKODA CONTRACTING~~SHERIDAN~2ND~E. 1ST
100600197~2010-03-01 08:38:43-05~EMERGENCY~UNION~ROSELLE~NEW JERSEY AMERICAN WATER COMP~405~1ST~HOPE~
100600198~2010-03-01 08:38:45-05~EMERGENCY~BURLINGTON~RIVERTON~PSE&G~527~MAIN~BROAD~HOWARD
100600199~2010-03-01 08:39:17-05~EMERGENCY~MONMOUTH~EATONTOWN~NEW JERSEY AMERICAN WATER COMP~15~ELIZABETH~BROAD~
100600200~2010-03-01 08:36:59-05~ROUTINE~CAPE MAY~LOWER TWP~KLINE CONSTRUCTION~12~SPRUCEWOOD~WILDWOOD~
100600201~2010-03-01 08:40:57-05~EMERGENCY~GLOUCESTER~CLAYTON~SOUTH JERSEY GAS~134~DENNIS~MATHEWS~
100600202~2010-03-01 08:41:40-05~ROUTINE~BURLINGTON~EVESHAM~J. FLETCHER CREAMER & SON~11~EUSTACE~EUSTACE~EVERINGTON
100600203~2010-03-01 08:55:27-05~ROUTINE~ESSEX~NUTLEY~N W SIGN INDUSTRIES~345~CENTRE~RAVINE~
100600204~2010-03-01 08:42:04-05~EMERGENCY~CAPE MAY~LOWER TWP~ATLANTIC CAPE WELL DRILLING~320~TAHOE~BAY~
100600205~2010-03-01 08:38:05-05~ROUTINE~MONMOUTH~FREEHOLD BORO~J F KIELY CONSTRUCTION COMPANY~~KIAWAH~HULL~
100600206~2010-03-01 08:42:26-05~ROUTINE~OCEAN~TOMS RIVER~J. FLETCHER CREAMER & SON~1069~BELL~PEPPERTREE~
100600207~2010-03-01 08:42:39-05~EMERGENCY~CAMDEN~CHERRY HILL~NEW JERSEY AMERICAN WATER COMP~~MERCER~SHERWOOD~
100600208~2010-03-01 08:42:25-05~ROUTINE~MONMOUTH~FREEHOLD BORO~J F KIELY CONSTRUCTION COMPANY~0~COTTAGE PLACE~HULL AVENUE~
100600209~2010-03-01 08:50:03-05~EMERGENCY~SALEM~SALEM~SOUTH JERSEY GAS~211~SINNICKSON~WALNUT~
100600210~2010-03-01 08:43:55-05~EMERGENCY~CAMDEN~MERCHANTVILLE~HENKELS & MCCOY~~VOLAN~SPRINGFIELD~
100600211~2010-03-01 08:40:37-05~ROUTINE~CAPE MAY~UPPER TWP~KLINE CONSTRUCTION~19~KATHERINE~SHORE~
100600212~2010-03-01 08:34:51-05~ROUTINE~MONMOUTH~MARLBORO~J FLETCHER CREAMER & SON~225~NAVAJO~LAREDO~
100600213~2010-03-01 08:44:28-05~ROUTINE~UNION~SUMMIT~ARMSTRONG INC~97~HOBART~WHITTREDGE~SPRINGFIELD
100600214~2010-03-01 08:44:51-05~EMERGENCY~UNION~CRANFORD~NEW JERSEY AMERICAN WATER COMP~1~STRATFORD~LINCOLN~
100600215~2010-03-01 08:44:52-05~ROUTINE~BURLINGTON~MEDFORD~J. FLETCHER CREAMER & SON~16~HIGH POINT~PEPPERBUSH~
100600216~2010-03-01 08:44:59-05~ROUTINE~BURLINGTON~MOUNT LAUREL~PHILADELPHIA SIGN COMPANY~12000~HORIZON~SPRINGDALE~
100600217~2010-03-01 08:45:09-05~ROUTINE~OCEAN~BAY HEAD~TOM OGRADY~513~MAIN~HOWE~MOUNT
100600218~2010-03-01 08:45:13-05~EMERGENCY~PASSAIC~PATERSON~MILLER PIPE LINE~425~CROSBY~ALBION~REDWOOD
100600219~2010-03-01 08:45:21-05~ROUTINE~BURLINGTON~EVESHAM~J. FLETCHER CREAMER & SON~2~KATHRYN~ELMWOOD~
100600220~2010-03-01 08:43:57-05~ROUTINE~MONMOUTH~FREEHOLD BORO~J F KIELY CONSTRUCTION COMPANY~0~SOUTH STREET~BRINKERHOFF AVENUE~
100600221~2010-03-01 08:45:49-05~ROUTINE~MONMOUTH~LITTLE SILVER~FENCES BY DENTE~55~RUSTIC~RUMSON~BRANCH
100600222~2010-03-01 08:44:10-05~ROUTINE~CAPE MAY~STONE HARBOR~KLINE CONSTRUCTION~10~92ND~1ST~
100600223~2010-03-01 08:44:21-05~ROUTINE~MONMOUTH~TINTON FALLS~J FLETCHER CREAMER & SON~13~JONATHAN~HERITAGE~
100600224~2010-03-01 08:45:24-05~ROUTINE~MONMOUTH~FREEHOLD BORO~J F KIELY CONSTRUCTION COMPANY~~KIAWAH~PARK~
100600225~2010-03-01 08:45:41-05~ROUTINE~ATLANTIC~SOMERS POINT~KVP DEVELOPERS~1~NEW YORK~BAY~SHORE
100600227~2010-03-01 08:47:02-05~ROUTINE~CAPE MAY~SEA ISLE~KLINE CONSTRUCTION~4~79TH~PLEASURE~
100600228~2010-03-01 08:49:51-05~ROUTINE~MERCER~HAMILTON TWP~C.J.S. INVESTMENTS~~YARDVILLE ALLENTOWN~TATTLETOWN~
100600229~2010-03-01 08:50:19-05~EMERGENCY~CAPE MAY~UPPER TWP~ATLANTIC CAPE WELL DRILLING~1438~STAGECOACH~WILLIAMS~
100600231~2010-03-01 08:49:42-05~ROUTINE~CAPE MAY~SEA ISLE~KLINE CONSTRUCTION~4509~LANDIS~45TH~
100600232~2010-03-01 08:52:28-05~ROUTINE~MORRIS~DOVER~DAVID B JOHNSON~265~US HWY 46~SCHLEY~
100600233~2010-03-01 08:52:35-05~EMERGENCY~MERCER~HAMILTON TWP~PUBLIC SERVICE ELECTRIC & GAS~156~ELIZABETH~BOW HILL~
100600234~2010-03-01 08:52:49-05~ROUTINE~MORRIS~FLORHAM PARK~GPU~15~WASHINGTON~BRIARWOOD~
100600235~2010-03-01 08:52:52-05~EMERGENCY~PASSAIC~PATERSON~PASSAIC VALLEY WATER COMMISSIO~171~PUTNAM~ROSA PARKS~
100600236~2010-03-01 08:53:10-05~ROUTINE~BURLINGTON~MEDFORD~J. FLETCHER CREAMER & SON~26~KNOTTY OAK~TALLOWOOD~
100600237~2010-03-01 08:53:53-05~EMERGENCY~MONMOUTH~HAZLET~SHORELANDS WATER COMPANY~2150~FLORENCE~HWY 36~
100600238~2010-03-01 08:52:06-05~ROUTINE~CAPE MAY~LOWER TWP~KLINE CONSTRUCTION~412~2ND~BROADWAY~
100600240~2010-03-01 08:54:36-05~ROUTINE~OCEAN~BRICK~CONCORD ENVIRONMENTAL SERVICE~1254~CONCORD~HERBERTSVILLE~
100600241~2010-03-01 08:54:43-05~EMERGENCY~ESSEX~CEDAR GROVE~PSE&G~~OZONE~COMMERCE~
100600243~2010-03-01 08:55:57-05~ROUTINE~GLOUCESTER~WEST DEPTFORD~J. FLETCHER CREAMER & SON~725~CARTER HILL~OLLERTON~
```

Figure 29: Received Raw Data Sample 1


ticketSearchResults-2015-02-09-20_48_16.txt - Notepad
File Edit Format View Help

```
100600144~2010-03-01 08:19:59-05~ROUTINE~OCEAN~TOMS RIVER~J. FLETCHER CREAMER & SON~2414~WHITESVILLE~CLEARLAKE~
100600145~2010-03-01 08:20:19-05~EMERGENCY~UNION~SPRINGFIELD~PRO TANK SERVICES~46~EDGEWOOD~MOUNTAIN~WENTZ
100600146~2010-03-01 08:20:32-05~EMERGENCY~MORRIS~MADISON~PSE&G~~PROSPECT~WYNDEHURST~
100600147~2010-03-01 08:20:49-05~EMERGENCY~HUDSON~JERSEY CITY~PSE&G~60~WALES~BROADWAY~
100600148~2010-03-01 08:19:56-05~ROUTINE~ATLANTIC~HAMMONTON~SOUTH JERSEY GAS~197~PARK~SAMUEL~CHESTNUT
100600150~2010-03-01 08:21:47-05~EMERGENCY~MIDDLESEX~PERTH AMBOY~PERTH AMBOY WATER COMPANY~372~SUTTON~AMBOY~
100600151~2010-03-01 08:21:58-05~ROUTINE~CAMDEN~HADDON HEIGHTS~M & N ASSOCIATES~419~1ST~GARDEN~
100600152~2010-03-01 08:22:36-05~ROUTINE~CAMDEN~CHERRY HILL~J. FLETCHER CREAMER & SON~18~GATEHOUSE~LAMP POST~OLD ORCHARD
100600153~2010-03-01 08:23:27-05~ROUTINE~CAPE MAY~SEA ISLE~CITY OF SEA ISLE CITY~384~47TH~PARK~
100600154~2010-03-01 08:19:23-05~ROUTINE~ATLANTIC~EGG HARBOR TWP~KLINE CONSTRUCTION~3015~SPRUCE~BLACK HORSE~
100600155~2010-03-01 08:23:46-05~ROUTINE~ATLANTIC~VENTNOR~AQUA-TEX TRANSPORT INC.~109~PORTLAND~ATLANTIC~
100600156~2010-03-01 08:23:58-05~ROUTINE~CAMDEN~HADDON HEIGHTS~M & N ASSOCIATES~307~3RD~GARDEN~
100600157~2010-03-01 08:24:40-05~EMERGENCY~HUDSON~JERSEY CITY~PSE&G~~SPRUCE~TONELLE~
100600158~2010-03-01 08:25:02-05~ROUTINE~ESSEX~NEWARK~GRANADA CONSTRUCTION CORPORATI~1~FOSTER~DAYTON~FRELINGHUYSEN
100600159~2010-03-01 08:25:18-05~EMERGENCY~MORRIS~LONG HILL~PSE&G~309~UNION~MAIN~
100600160~2010-03-01 08:23:29-05~ROUTINE~ATLANTIC~EGG HARBOR TWP~KLINE CONSTRUCTION~108~COLONIAL~HILLTOP~
100600162~2010-03-01 08:26:02-05~ROUTINE~CAMDEN~HADDON HEIGHTS~M & N ASSOCIATES~100~3RD~GREEN~
100600163~2010-03-01 08:26:08-05~EMERGENCY~MIDDLESEX~PISCATAWAY~NEW JERSEY AMERICAN WATER COMP~654~DIAL~ACADEMY~
100600164~2010-03-01 08:26:20-05~ROUTINE~SUSSEX~NEWTON~GENERAL BORINGS INC.~~DOUMA~RYERSON~
100600165~2010-03-01 08:26:40-05~ROUTINE~GLOUCESTER~GLASSBORO~J. FLETCHER CREAMER & SON~154~GLASSWYCKE~GREENTREE~
100600166~2010-03-01 08:26:46-05~ROUTINE~HUDSON~HARRISON~CARSON & ROBERTS SITE CONSTRUC~~CAPE MAY~FRANK E ROGERS~CAPE MAY ST
100600167~2010-03-01 08:22:47-05~ROUTINE~UNION~ROSELLE~SKODA CONTRACTING~~2ND~THOMPSON~POPLAR
100600169~2010-03-01 08:25:53-05~ROUTINE~ATLANTIC~EGG HARBOR TWP~KLINE CONSTRUCTION~606~NATHAN~RUSHWOOD~
100600170~2010-03-01 08:28:04-05~ROUTINE~GLOUCESTER~CLAYTON~RNG VENTURES~37 A~CLAYTON~PEARL~
100600171~2010-03-01 08:28:04-05~ROUTINE~CAMDEN~CAMDEN~M & N ASSOCIATES~32~21ST~FEDERAL~
100600173~2010-03-01 08:27:43-05~ROUTINE~ESSEX~IRVINGTON~NEW JERSEY AMERICAN WATER COMP~30~HEADLEY~SPRINGFIELD~ELMWOOD
100600174~2010-03-01 08:29:23-05~ROUTINE~OCEAN~TOMS RIVER~J. FLETCHER CREAMER & SON~1872~OLD FREEHOLD~WHITE OAK BOTTOM~
100600175~2010-03-01 08:30:45-05~ROUTINE~BERGEN~MIDLAND PARK~HARDY BUILDING~43~OAKHILL~PARK~
100600176~2010-03-01 08:31:46-05~EMERGENCY~MIDDLESEX~PISCATAWAY~NEW JERSEY AMERICAN WATER COMP~9~FREEDOM~JUSTICE~
100600177~2010-03-01 08:32:03-05~ROUTINE~MIDDLESEX~MIDDLESEX~EPI INC.~~PERSHING~MOUNTAIN~
100600178~2010-03-01 08:32:25-05~ROUTINE~HUDSON~HARRISON~CARSON & ROBERTS SITE CONSTRUC~~FRANK E RODGERS~CAPE MAY~SOMERSET
100600179~2010-03-01 08:32:27-05~ROUTINE~GLOUCESTER~ELK~J. FLETCHER CREAMER & SON~679~FAIRVIEW~10TH~9TH
100600180~2010-03-01 08:32:36-05~ROUTINE~OCEAN~TOMS RIVER~J. FLETCHER CREAMER & SON~70~BLUE HERON~BLUE HERON~
100600181~2010-03-01 08:28:02-05~ROUTINE~ATLANTIC~LINWOOD~KLINE CONSTRUCTION~105~COUNTRY CLUB~WOODLYNNE~
100600182~2010-03-01 08:55:28-05~EMERGENCY~PASSAIC~WEST MILFORD~SLATER WELL DRILLING INC.~70~MOUNTAIN~MORSETOWN~
100600183~2010-03-01 08:33:10-05~ROUTINE~CAPE MAY~SEA ISLE~CLARENCE E SMITH PLUMBING~3918~LANDIS~39TH~
100600184~2010-03-01 08:53:07-05~ROUTINE~ATLANTIC~PLEASANTVILLE~SEASHORE ASSOCIATED MECHANICAL~~MILAN~LYONS~FRANKFURT
100600185~2010-03-01 08:34:13-05~EMERGENCY~UNION~ROSELLE PARK~TRIPLE A ABLE PLUMBING~202~UNION~GRANT~
100600186~2010-03-01 08:32:40-05~ROUTINE~ATLANTIC~LINWOOD~KLINE CONSTRUCTION~106~DAIRY~WOODLYNNE~
100600187~2010-03-01 08:35:15-05~ROUTINE~HUDSON~HARRISON~CARSON & ROBERTS SITE CONSTRUC~~SIXTH~CAPE MAY~
100600188~2010-03-01 08:35:32-05~ROUTINE~OCEAN~TOMS RIVER~J. FLETCHER CREAMER & SON~1276~BEAUCHAMPS~DORSET PSGE~
100600189~2010-03-01 08:29:07-05~ROUTINE~UNION~ROSELLE~SKODA CONTRACTING~~THOMPSON~2ND~E. 1ST
100600190~2010-03-01 08:36:21-05~ROUTINE~BERGEN~GARFIELD~PUBLIC SERVICE ELECTRIC AND GA~380~SEMEL~MIDLAND~MAIN
100600191~2010-03-01 08:34:47-05~ROUTINE~ATLANTIC~GALLOWAY~KLINE CONSTRUCTION~442~EBONY TREE~HAYES~
100600192~2010-03-01 08:37:01-05~ROUTINE~UNION~ROSELLE PARK~MANOR ELECTRIC~110~WESTFIELD~WALNUT~CHESTNUT
100600193~2010-03-01 08:32:55-05~ROUTINE~OCEAN~LAKEWOOD~J F KIELY CONSTRUCTION COMPANY~128~ST KATHERINE PLACE~HOPE CHAPEL~
```

Figure 30: Received Raw Data Sample 2

1.23 Data Pre-processing

The data preprocessing steps were performed to better prepare the data for Bayesian model and probability percentage for the risk factors. Since a large portion of the data set contained missing values, it was important to focus on correcting these instances. When examining the data, it was evident that some gas distributors kept extra records. Often there were variables that were recorded for only for a single company. That means that some entries would exist for one attribute. In other words, 25% of the attribute would contain values while the other 75% were unknown. Having such sparse data would have complicated data analysis and, more importantly, risk analysis methods. In that event, the first data preprocessing step performed eliminated attributes that were deemed unnecessary for further computations. Eliminating attributes cleaned the data set up and assisted with the removal of a large number of missing values that existed. Furthermore, eliminating attributes actively reduces the dimensionality of the data set and allows for increased model performance with a reduced computation time. Data set reduction resulted in the following remaining attributes:

1) Incident Date
2) Underground Facility Operator
3) Latitude Coordinate
4) Longitude Coordinate

One of the goals of this study was to develop risk model to predict future risk involved on the gas pipe damage process which involve all parties. Thus, the distribution of 4 groups and the attributes was necessary to get precise results. It was concluded that additional attributes would be required in order to effectively accomplish our research goal.

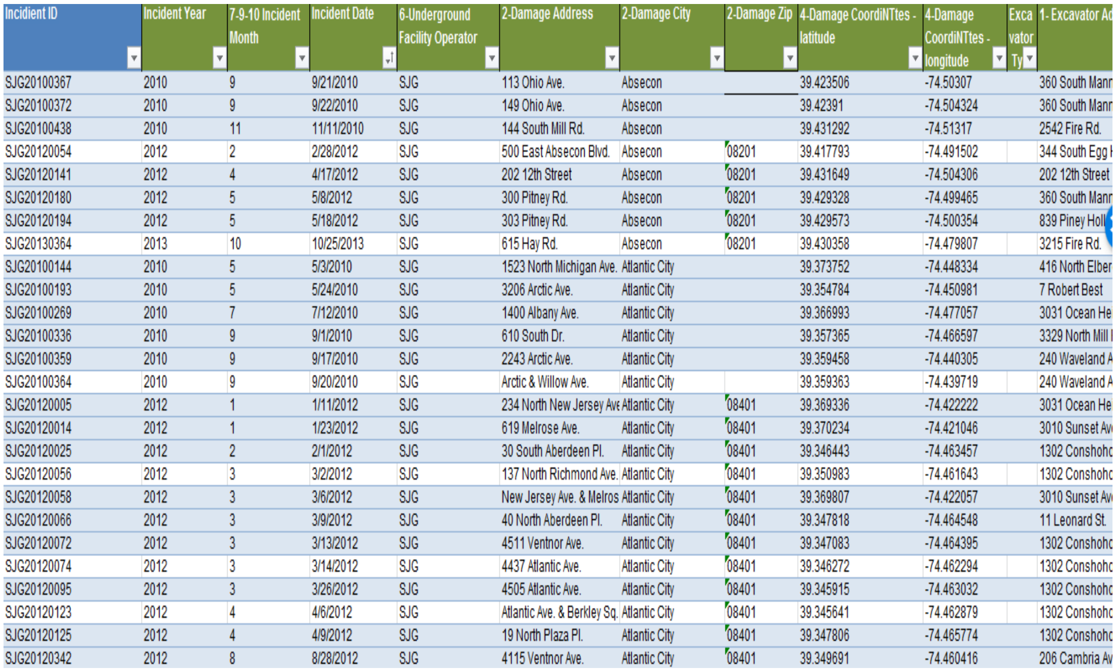
This concept was expanded to apply to nominal attributes which resulted in the following set of attributes:

1) Damage Zip code
2) Locate Type Ticket
3) Latitude Coordinate
4) Number of Excavating Request on the Same Area
5) Excavation Depth
6) Number of outgoing Calls
7) Number of Passed damages with 10-mile radius
8) Damage Description

In addition, the original nominal attributes were left in the raw data for easier identification, but were removed for any risk model application.

In addition, the data was distributed over 5 sheets in excel as 2010, 2011, 2012, 2013, and 2014. Then the data cleaned by some functions built in Excel. Then, some attribute was selected. Then, the complete address was combined together including street number, name, city, state, and zip code. Then, excel functions were developed to cross reference the data set and produce

two categories damaged data, and non-damaged data as in Figure 31.

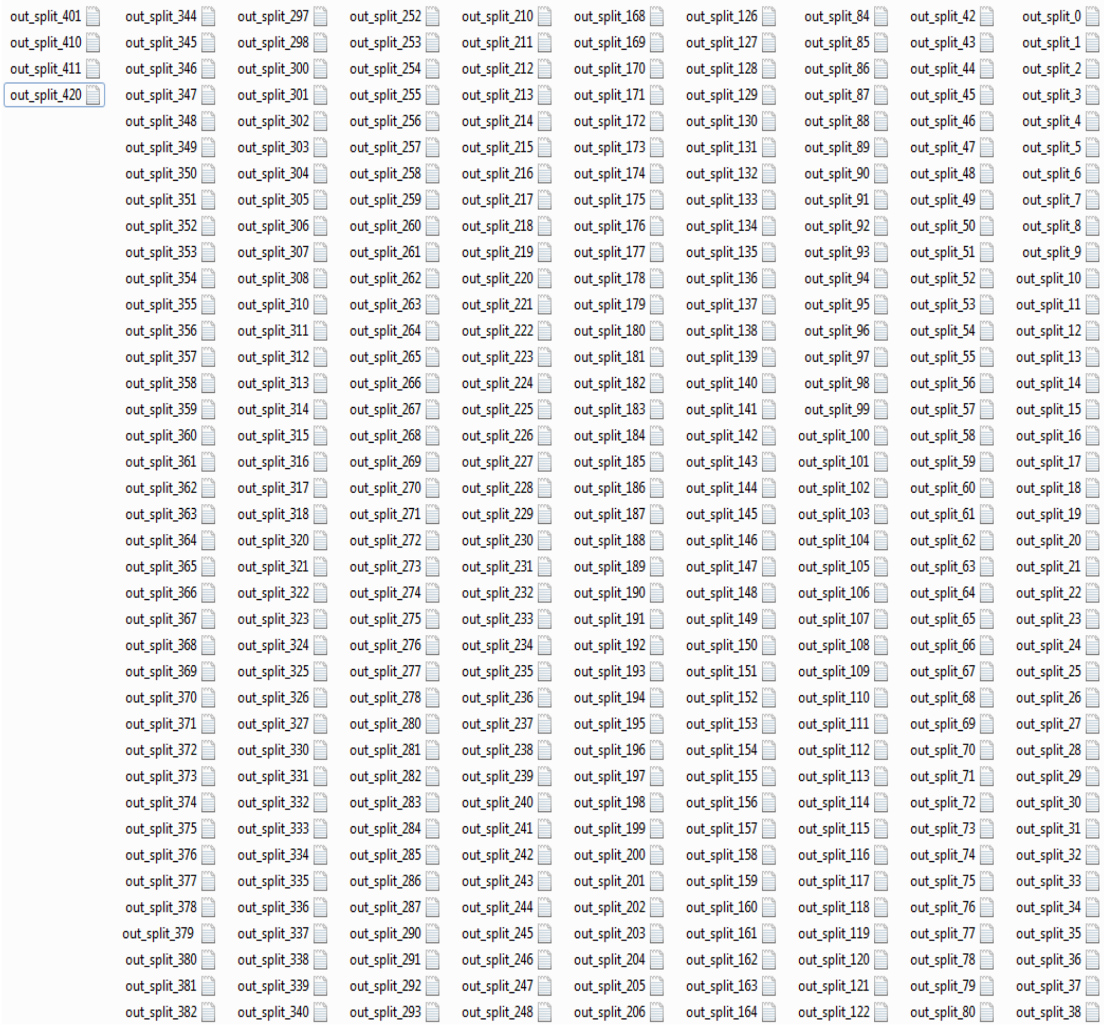


Incident ID	Incident Year	7-9-10 Incident Month	Incident Date	6-Underground Facility Operator	2-Damage Address	2-Damage City	2-Damage Zip	4-Damage CoordiNTtes - latitude	4-Damage CoordiNTtes - longitude	Exca vator Ty	1- Excavator Ad
SJG20100367	2010	9	9/21/2010	SJG	113 Ohio Ave.	Absecon		39.423506	-74.50307		360 South Mann
SJG20100372	2010	9	9/22/2010	SJG	149 Ohio Ave.	Absecon		39.42391	-74.504324		360 South Mann
SJG20100438	2010	11	11/11/2010	SJG	144 South Mill Rd.	Absecon		39.431292	-74.51317		2542 Fire Rd.
SJG20120054	2012	2	2/28/2012	SJG	500 East Absecon Blvd.	Absecon	08201	39.417793	-74.491502		344 South Egg I
SJG20120141	2012	4	4/17/2012	SJG	202 12th Street	Absecon	08201	39.431649	-74.504306		202 12th Street
SJG20120180	2012	5	5/8/2012	SJG	300 Pitney Rd.	Absecon	08201	39.429328	-74.499465		360 South Mann
SJG20120194	2012	5	5/18/2012	SJG	303 Pitney Rd.	Absecon	08201	39.429573	-74.500354		839 Piney Holl
SJG20130364	2013	10	10/25/2013	SJG	615 Hay Rd.	Absecon	08201	39.430358	-74.479807		3215 Fire Rd.
SJG20100144	2010	5	5/3/2010	SJG	1523 North Michigan Ave.	Atlantic City		39.373752	-74.448334		416 North Elber
SJG20100193	2010	5	5/24/2010	SJG	3206 Arctic Ave.	Atlantic City		39.354784	-74.450981		7 Robert Best
SJG20100269	2010	7	7/12/2010	SJG	1400 Albany Ave.	Atlantic City		39.366993	-74.477057		3031 Ocean He
SJG20100336	2010	9	9/1/2010	SJG	610 South Dr.	Atlantic City		39.357365	-74.466597		3329 North Mill I
SJG20100359	2010	9	9/17/2010	SJG	2243 Arctic Ave.	Atlantic City		39.359458	-74.440305		240 Waveland A
SJG20100364	2010	9	9/20/2010	SJG	Arctic & Willow Ave.	Atlantic City		39.359363	-74.439719		240 Waveland A
SJG20120005	2012	1	1/11/2012	SJG	234 North New Jersey Ave	Atlantic City	08401	39.369336	-74.422222		3031 Ocean He
SJG20120014	2012	1	1/23/2012	SJG	619 Melrose Ave.	Atlantic City	08401	39.370234	-74.421046		3010 Sunset Av
SJG20120025	2012	2	2/1/2012	SJG	30 South Aberdeen Pl.	Atlantic City	08401	39.346443	-74.463457		1302 Conshohc
SJG20120056	2012	3	3/2/2012	SJG	137 North Richmond Ave.	Atlantic City	08401	39.350983	-74.461643		1302 Conshohc
SJG20120058	2012	3	3/6/2012	SJG	New Jersey Ave. & Melros	Atlantic City	08401	39.369807	-74.422057		3010 Sunset Av
SJG20120066	2012	3	3/9/2012	SJG	40 North Aberdeen Pl.	Atlantic City	08401	39.347818	-74.464548		11 Leonard St.
SJG20120072	2012	3	3/13/2012	SJG	4511 Ventnor Ave.	Atlantic City	08401	39.347083	-74.464395		1302 Conshohc
SJG20120074	2012	3	3/14/2012	SJG	4437 Atlantic Ave.	Atlantic City	08401	39.346272	-74.462294		1302 Conshohc
SJG20120095	2012	3	3/26/2012	SJG	4505 Atlantic Ave.	Atlantic City	08401	39.345915	-74.463032		1302 Conshohc
SJG20120123	2012	4	4/6/2012	SJG	Atlantic Ave. & Berkley Sq.	Atlantic City	08401	39.345641	-74.462879		1302 Conshohc
SJG20120125	2012	4	4/9/2012	SJG	19 North Plaza Pl.	Atlantic City	08401	39.347806	-74.465774		1302 Conshohc
SJG20120342	2012	8	8/28/2012	SJG	4115 Ventnor Ave.	Atlantic City	08401	39.349691	-74.460416		206 Cambria Av

Figure 31: Organized data set sample

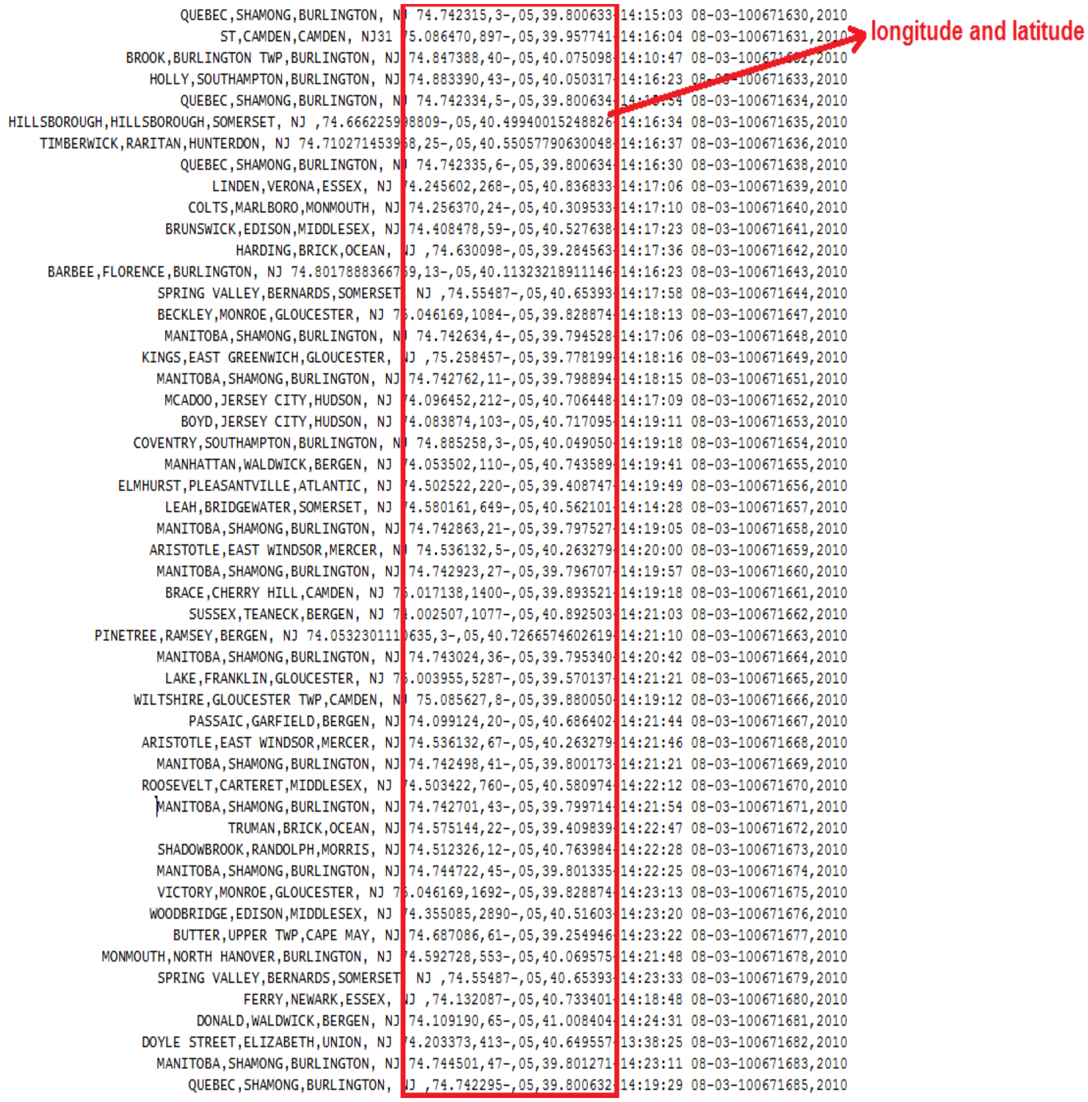
1.24 Geo-code data, Latitude, and Longitude.

The collected undamaged data did not have the Latitude, and Longitude. Thus, it was not possible to plot large pile of data without Latitude, and Longitude. After determined the Latitude, and Longitude to all locations, the data was split into 420 files figure 32, each file contains complete address, and Lat/Long figure 33. Then, all files combined into one Excel file so we can import it to QGIS, OR ArcGIS later.



```
out_split_401
out_split_410
out_split_411
out_split_420
out_split_344
out_split_345
out_split_346
out_split_347
out_split_348
out_split_349
out_split_350
out_split_351
out_split_352
out_split_353
out_split_354
out_split_355
out_split_356
out_split_357
out_split_358
out_split_359
out_split_360
out_split_361
out_split_362
out_split_363
out_split_364
out_split_365
out_split_366
out_split_367
out_split_368
out_split_369
out_split_370
out_split_371
out_split_372
out_split_373
out_split_374
out_split_375
out_split_376
out_split_377
out_split_378
out_split_379
out_split_380
out_split_381
out_split_382
out_split_297
out_split_298
out_split_300
out_split_301
out_split_302
out_split_303
out_split_304
out_split_305
out_split_306
out_split_307
out_split_308
out_split_310
out_split_311
out_split_312
out_split_313
out_split_314
out_split_315
out_split_316
out_split_317
out_split_318
out_split_320
out_split_321
out_split_322
out_split_323
out_split_324
out_split_325
out_split_326
out_split_327
out_split_330
out_split_331
out_split_332
out_split_333
out_split_334
out_split_335
out_split_336
out_split_337
out_split_338
out_split_339
out_split_340
out_split_252
out_split_253
out_split_254
out_split_255
out_split_256
out_split_257
out_split_258
out_split_259
out_split_260
out_split_261
out_split_262
out_split_263
out_split_264
out_split_265
out_split_266
out_split_267
out_split_268
out_split_269
out_split_270
out_split_271
out_split_272
out_split_273
out_split_274
out_split_275
out_split_276
out_split_277
out_split_278
out_split_280
out_split_281
out_split_282
out_split_283
out_split_284
out_split_285
out_split_286
out_split_287
out_split_290
out_split_291
out_split_292
out_split_293
out_split_210
out_split_211
out_split_212
out_split_213
out_split_214
out_split_215
out_split_216
out_split_217
out_split_218
out_split_219
out_split_220
out_split_221
out_split_222
out_split_223
out_split_224
out_split_225
out_split_226
out_split_227
out_split_228
out_split_229
out_split_230
out_split_231
out_split_232
out_split_233
out_split_234
out_split_235
out_split_236
out_split_237
out_split_238
out_split_239
out_split_240
out_split_241
out_split_242
out_split_243
out_split_244
out_split_245
out_split_246
out_split_247
out_split_248
out_split_168
out_split_169
out_split_170
out_split_171
out_split_172
out_split_173
out_split_174
out_split_175
out_split_176
out_split_177
out_split_178
out_split_179
out_split_180
out_split_181
out_split_182
out_split_183
out_split_184
out_split_185
out_split_186
out_split_187
out_split_188
out_split_189
out_split_190
out_split_191
out_split_192
out_split_193
out_split_194
out_split_195
out_split_196
out_split_197
out_split_198
out_split_199
out_split_200
out_split_201
out_split_202
out_split_203
out_split_204
out_split_205
out_split_206
out_split_126
out_split_127
out_split_128
out_split_129
out_split_130
out_split_131
out_split_132
out_split_133
out_split_134
out_split_135
out_split_136
out_split_137
out_split_138
out_split_139
out_split_140
out_split_141
out_split_142
out_split_143
out_split_144
out_split_145
out_split_146
out_split_147
out_split_148
out_split_149
out_split_150
out_split_151
out_split_152
out_split_153
out_split_154
out_split_155
out_split_156
out_split_157
out_split_158
out_split_159
out_split_160
out_split_161
out_split_162
out_split_163
out_split_164
out_split_84
out_split_85
out_split_86
out_split_87
out_split_88
out_split_89
out_split_90
out_split_91
out_split_92
out_split_93
out_split_94
out_split_95
out_split_96
out_split_97
out_split_98
out_split_99
out_split_100
out_split_101
out_split_102
out_split_103
out_split_104
out_split_105
out_split_106
out_split_107
out_split_108
out_split_109
out_split_110
out_split_111
out_split_112
out_split_113
out_split_114
out_split_115
out_split_116
out_split_117
out_split_118
out_split_119
out_split_120
out_split_121
out_split_122
out_split_42
out_split_43
out_split_44
out_split_45
out_split_46
out_split_47
out_split_48
out_split_49
out_split_50
out_split_51
out_split_52
out_split_53
out_split_54
out_split_55
out_split_56
out_split_57
out_split_58
out_split_59
out_split_60
out_split_61
out_split_62
out_split_63
out_split_64
out_split_65
out_split_66
out_split_67
out_split_68
out_split_69
out_split_70
out_split_71
out_split_72
out_split_73
out_split_74
out_split_75
out_split_76
out_split_77
out_split_78
out_split_79
out_split_80
out_split_0
out_split_1
out_split_2
out_split_3
out_split_4
out_split_5
out_split_6
out_split_7
out_split_8
out_split_9
out_split_10
out_split_11
out_split_12
out_split_13
out_split_14
out_split_15
out_split_16
out_split_17
out_split_18
out_split_19
out_split_20
out_split_21
out_split_22
out_split_23
out_split_24
out_split_25
out_split_26
out_split_27
out_split_28
out_split_29
out_split_30
out_split_31
out_split_32
out_split_33
out_split_34
out_split_35
out_split_36
out_split_37
out_split_38
```

Figure 32 : Shows 420 files of undamaged data



```
                 QUEBEC,SHAMONG,BURLINGTON, NJ 74.742315,3-,05,39.800633 14:15:03 08-03-100671630,2010
                           ST,CAMDEN,CAMDEN, NJ31 75.086470,897-,05,39.957741 14:16:04 08-03-100671631,2010
       BROOK,BURLINGTON TWP,BURLINGTON, NJ 74.847388,40-,05,40.075098 14:10:47 08-03-100671632,2010
          HOLLY,SOUTHAMPTON,BURLINGTON, NJ 74.883390,43-,05,40.050317 14:16:23 08-03-100671633,2010
                 QUEBEC,SHAMONG,BURLINGTON, NJ 74.742334,5-,05,39.800634 14:16:54 08-03-100671634,2010
HILLSBOROUGH,HILLSBOROUGH,SOMERSET, NJ ,74.66622598809-,05,40.49940015248826 14:16:34 08-03-100671635,2010
    TIMBERWICK,RARITAN,HUNTERDON, NJ 74.71027145398,25-,05,40.55057790630048 14:16:37 08-03-100671636,2010
                 QUEBEC,SHAMONG,BURLINGTON, NJ 74.742335,6-,05,39.800634 14:16:30 08-03-100671638,2010
                        LINDEN,VERONA,ESSEX, NJ 74.245602,268-,05,40.836833 14:17:06 08-03-100671639,2010
               COLTS,MARLBORO,MONMOUTH, NJ 74.256370,24-,05,40.309533 14:17:10 08-03-100671640,2010
            BRUNSWICK,EDISON,MIDDLESEX, NJ 74.408478,59-,05,40.527638 14:17:23 08-03-100671641,2010
                      HARDING,BRICK,OCEAN, NJ ,74.630098-,05,39.284563 14:17:36 08-03-100671642,2010
   BARBEE,FLORENCE,BURLINGTON, NJ 74.80178883667б9,13-,05,40.11323218911146 14:16:23 08-03-100671643,2010
       SPRING VALLEY,BERNARDS,SOMERSET  NJ ,74.55487-,05,40.65393 14:17:58 08-03-100671644,2010
          BECKLEY,MONROE,GLOUCESTER, NJ 75.046169,1084-,05,39.828874 14:18:13 08-03-100671647,2010
            MANITOBA,SHAMONG,BURLINGTON, NJ 74.742634,4-,05,39.794528 14:17:06 08-03-100671648,2010
       KINGS,EAST GREENWICH,GLOUCESTER, NJ ,75.258457-,05,39.778199 14:18:16 08-03-100671649,2010
           MANITOBA,SHAMONG,BURLINGTON, NJ 74.742762,11-,05,39.798894 14:18:15 08-03-100671651,2010
            MCADOO,JERSEY CITY,HUDSON, NJ 74.096452,212-,05,40.706448 14:17:09 08-03-100671652,2010
                BOYD,JERSEY CITY,HUDSON, NJ 74.083874,103-,05,40.717095 14:19:11 08-03-100671653,2010
      COVENTRY,SOUTHAMPTON,BURLINGTON, NJ 74.885258,3-,05,40.049050 14:19:18 08-03-100671654,2010
            MANHATTAN,WALDWICK,BERGEN, NJ 74.053502,110-,05,40.743589 14:19:41 08-03-100671655,2010
     ELMHURST,PLEASANTVILLE,ATLANTIC, NJ 74.502522,220-,05,39.408747 14:19:49 08-03-100671656,2010
            LEAH,BRIDGEWATER,SOMERSET, NJ 74.580161,649-,05,40.562101 14:14:28 08-03-100671657,2010
           MANITOBA,SHAMONG,BURLINGTON, NJ 74.742863,21-,05,39.797527 14:19:05 08-03-100671658,2010
          ARISTOTLE,EAST WINDSOR,MERCER, NJ 74.536132,5-,05,40.263279 14:20:00 08-03-100671659,2010
           MANITOBA,SHAMONG,BURLINGTON, NJ 74.742923,27-,05,39.796707 14:19:57 08-03-100671660,2010
            BRACE,CHERRY HILL,CAMDEN, NJ 75.017138,1400-,05,39.893521 14:19:18 08-03-100671661,2010
               SUSSEX,TEANECK,BERGEN, NJ 74.002507,1077-,05,40.892503 14:21:03 08-03-100671662,2010
   PINETREE,RAMSEY,BERGEN, NJ 74.05323011100635,3-,05,40.7266574602619 14:21:10 08-03-100671663,2010
           MANITOBA,SHAMONG,BURLINGTON, NJ 74.743024,36-,05,39.795340 14:20:42 08-03-100671664,2010
            LAKE,FRANKLIN,GLOUCESTER, NJ 75.003955,5287-,05,39.570137 14:21:21 08-03-100671665,2010
      WILTSHIRE,GLOUCESTER TWP,CAMDEN, NJ 75.085627,8-,05,39.880050 14:19:12 08-03-100671666,2010
               PASSAIC,GARFIELD,BERGEN, NJ 74.099124,20-,05,40.686402 14:21:44 08-03-100671667,2010
         ARISTOTLE,EAST WINDSOR,MERCER, NJ 74.536132,67-,05,40.263279 14:21:46 08-03-100671668,2010
           MANITOBA,SHAMONG,BURLINGTON, NJ 74.742498,41-,05,39.800173 14:21:21 08-03-100671669,2010
         ROOSEVELT,CARTERET,MIDDLESEX, NJ 74.503422,760-,05,40.580974 14:22:12 08-03-100671670,2010
           MANITOBA,SHAMONG,BURLINGTON, NJ 74.742701,43-,05,39.799714 14:21:54 08-03-100671671,2010
                    TRUMAN,BRICK,OCEAN, NJ 74.575144,22-,05,39.409839 14:22:47 08-03-100671672,2010
           SHADOWBROOK,RANDOLPH,MORRIS, NJ 74.512326,12-,05,40.763984 14:22:28 08-03-100671673,2010
           MANITOBA,SHAMONG,BURLINGTON, NJ 74.744722,45-,05,39.801335 14:22:25 08-03-100671674,2010
          VICTORY,MONROE,GLOUCESTER, NJ 75.046169,1692-,05,39.828874 14:23:13 08-03-100671675,2010
          WOODBRIDGE,EDISON,MIDDLESEX, NJ 74.355085,2890-,05,40.51603 14:23:20 08-03-100671676,2010
              BUTTER,UPPER TWP,CAPE MAY, NJ 74.687086,61-,05,39.254946 14:23:22 08-03-100671677,2010
  MONMOUTH,NORTH HANOVER,BURLINGTON, NJ 74.592728,553-,05,40.069575 14:21:48 08-03-100671678,2010
       SPRING VALLEY,BERNARDS,SOMERSET  NJ ,74.55487-,05,40.65393 14:23:33 08-03-100671679,2010
                      FERRY,NEWARK,ESSEX, NJ ,74.132087-,05,40.733401 14:18:48 08-03-100671680,2010
                 DONALD,WALDWICK,BERGEN, NJ 74.109190,65-,05,41.008404 14:24:31 08-03-100671681,2010
         DOYLE STREET,ELIZABETH,UNION, NJ 74.203373,413-,05,40.649557 13:38:25 08-03-100671682,2010
           MANITOBA,SHAMONG,BURLINGTON, NJ 74.744501,47-,05,39.801271 14:23:11 08-03-100671683,2010
                QUEBEC,SHAMONG,BURLINGTON, NJ ,74.742295-,05,39.800632 14:19:29 08-03-100671685,2010
```

Figure 33 file shows latitude/Longitude of UG undamaged data

1.25 Performing Exploratory Analysis

There is mutable software platforms to perform the task of plotting locations. However, not all of them will handle large number of data, and requirement of spatial analysis including heat map. Thus, for the purpose of this research will use ArcGIS to plot, an analyze 775,000 locations of undamaged UG gas pipe data. In addition, Fusion table was used to perform preliminary plotting.

All geocoded data was saved and combined to one CSV file. Then the CSV imported to ArcGIS through multiple steps including creating new project, creating new map, setting up layers, importing data to the map, and plotting .See figures 34, 35.

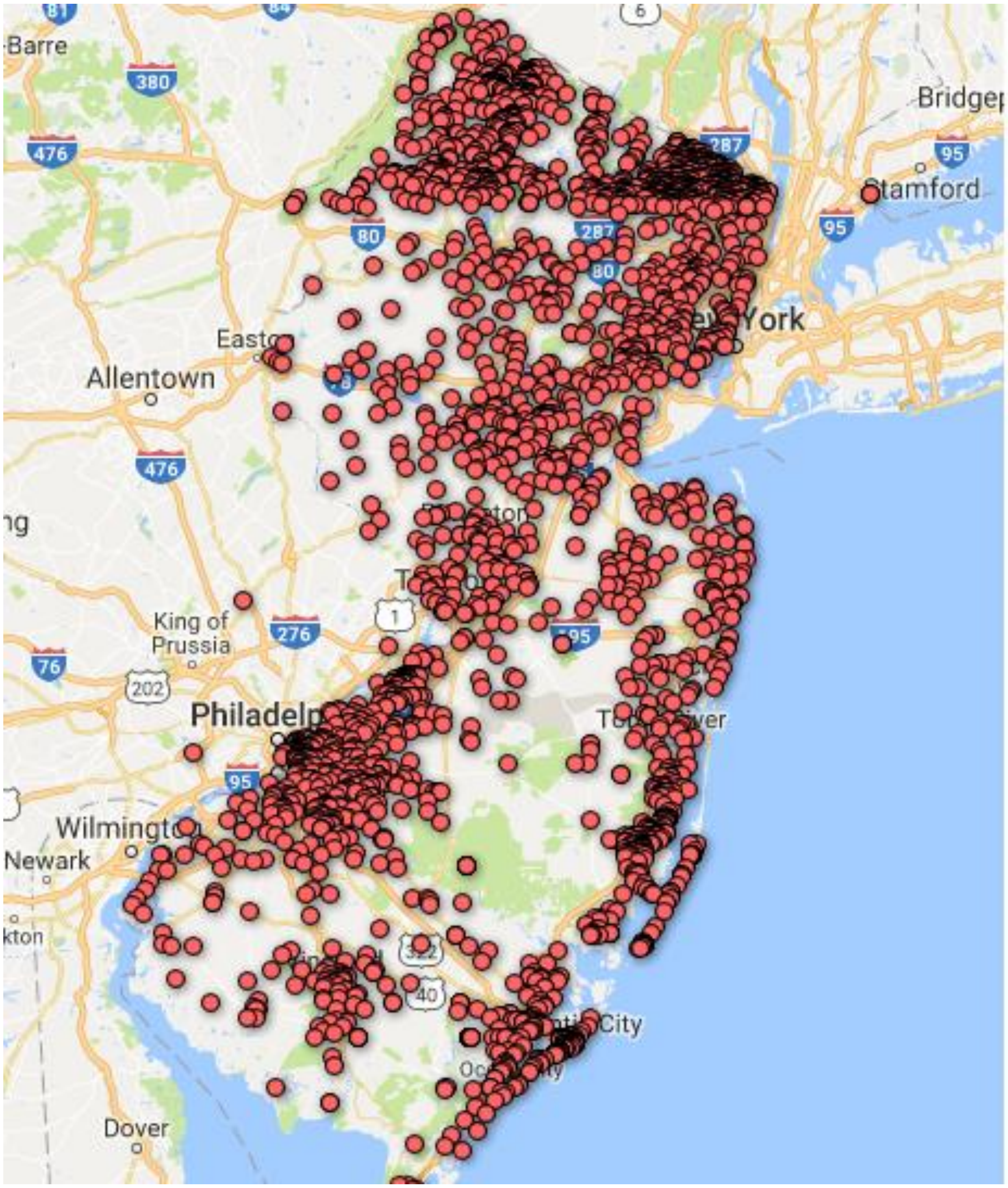


Figure 34 Shows plotted 775,000 locations using Fusion Table.

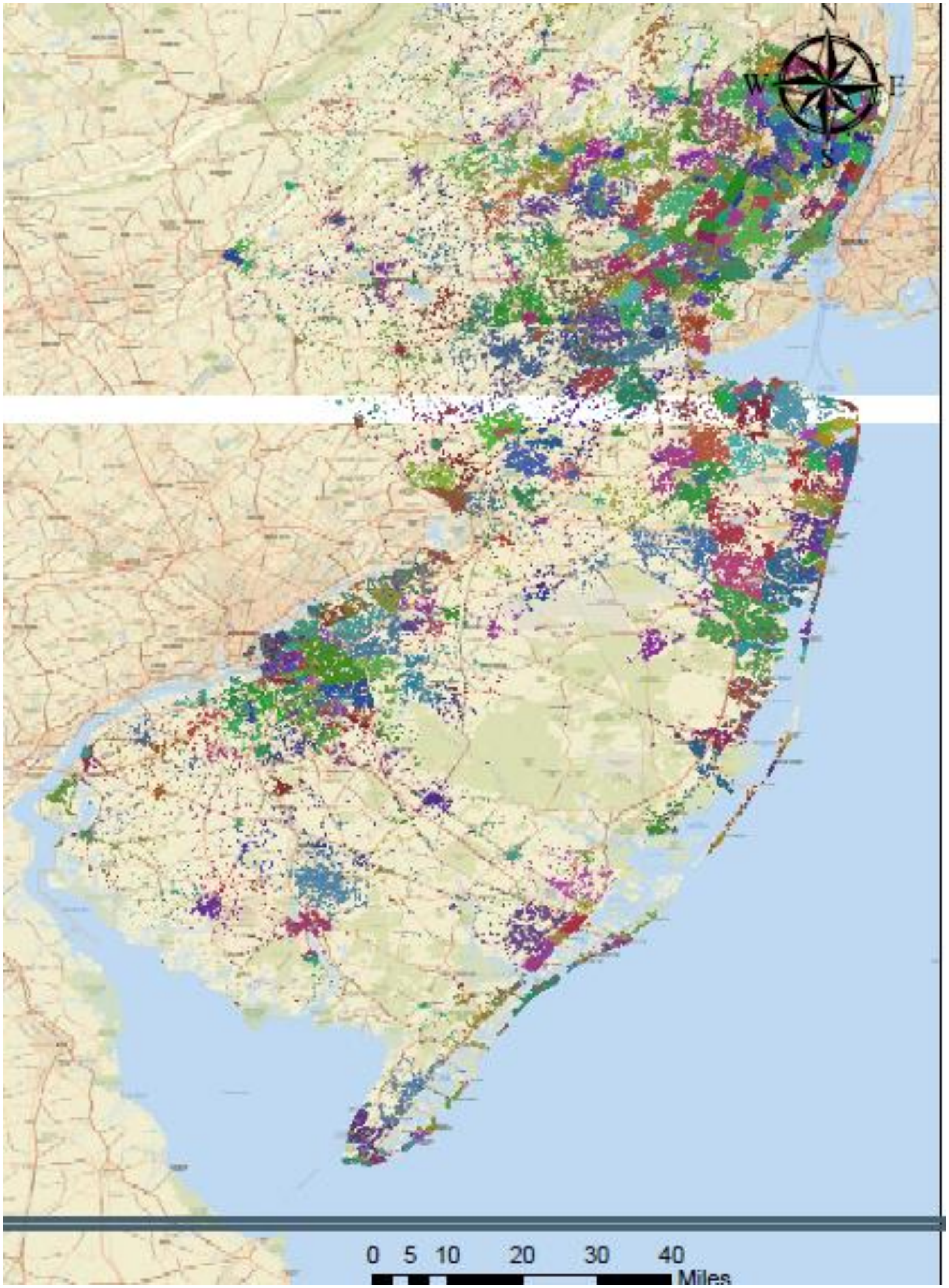
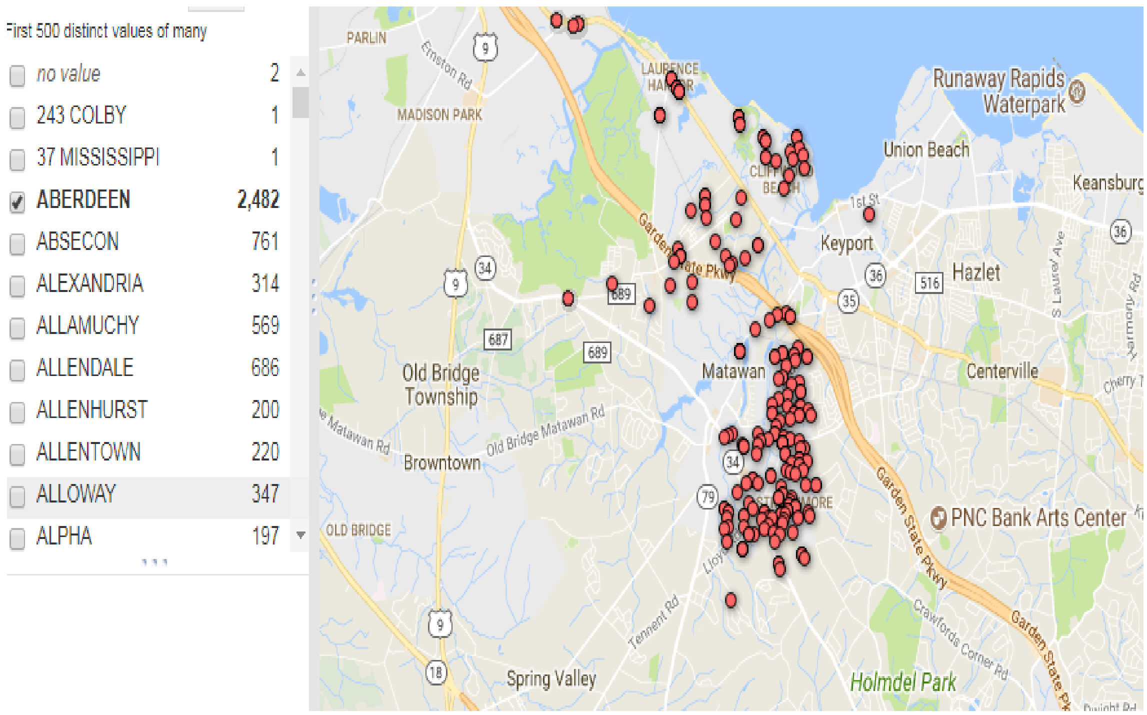


Figure 35 Shows plotted 775,000 locations using ArcGIS

The system was developed to visually communicate and allow analysis of underground (damaged) gas pipes related to utility companies. One of the distinctive objectives of this method is the use of 'heat maps' as a visual means for communicating the spatial density of UG gas percentage of damage.

Distance spatial analysis was performed on Aberdeen city. Figure 36 shows 2,482 of undamaged UG data plotted, then heat map was created to determine the densest/cluster data center figure 37. After the center location been specified on the heat map, three distances were plotted on the map, 5 miles, 10 miles, and 15 miles. All data was imported to Fusion Tables for better visualization Figure 38.



First 500 distinct values of many

□ *no value*	2
□ 243 COLBY	1
□ 37 MISSISSIPPI	1
☑ ABERDEEN	2,482
□ ABSECON	761
□ ALEXANDRIA	314
□ ALLAMUCHY	569
□ ALLENDALE	686
□ ALLENHURST	200
□ ALLENTOWN	220
□ ALLOWAY	347
□ ALPHA	197

Figure 36 Plotted Undamaged UG gas pipes in Aberdeen city

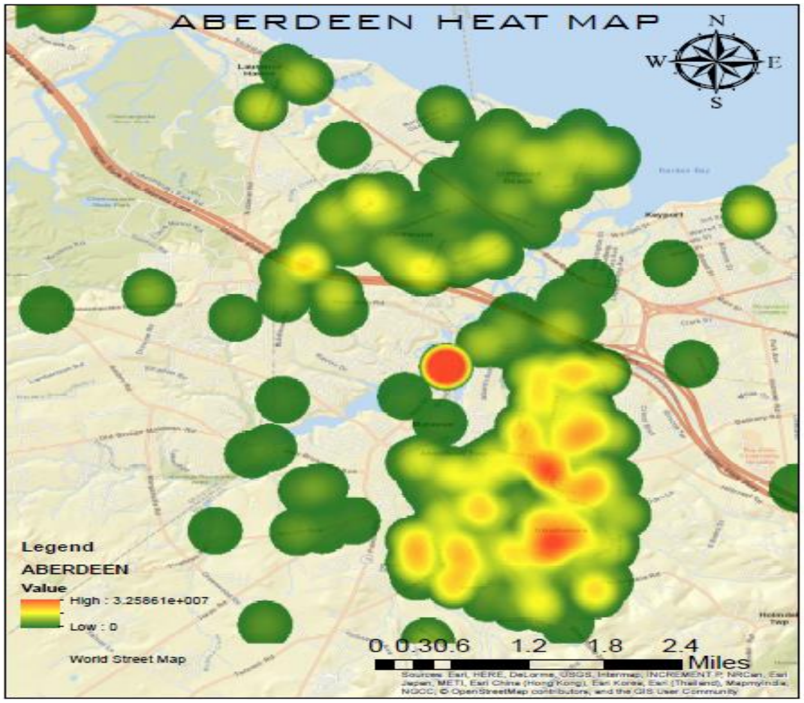


Figure 37 Shows Heat map to Undamaged UG gas pipes to Aberdeen City by ArcGIS

1.26 Data Visualization

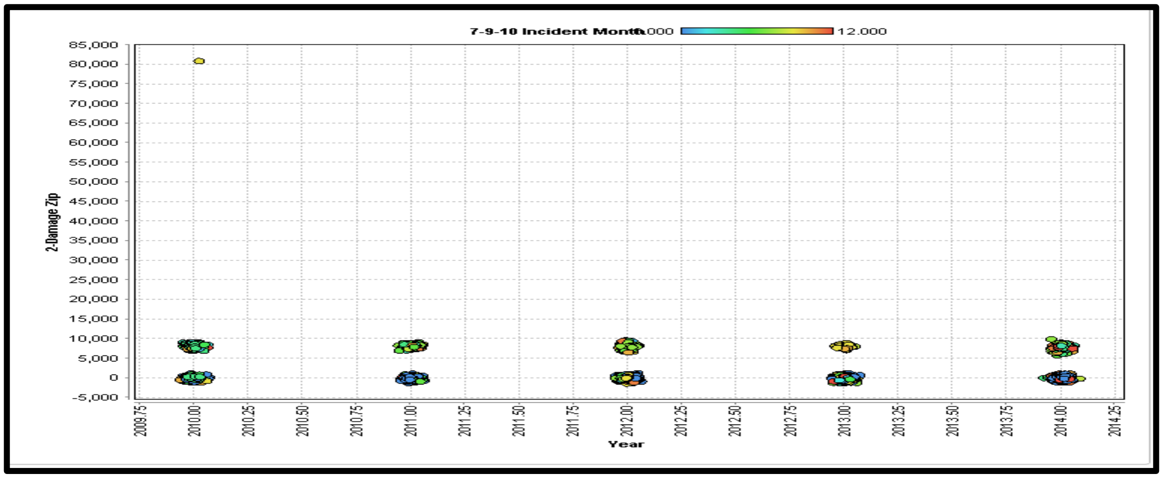


Figure 38: Damaged Data plotting Zip code & Year (phase1)

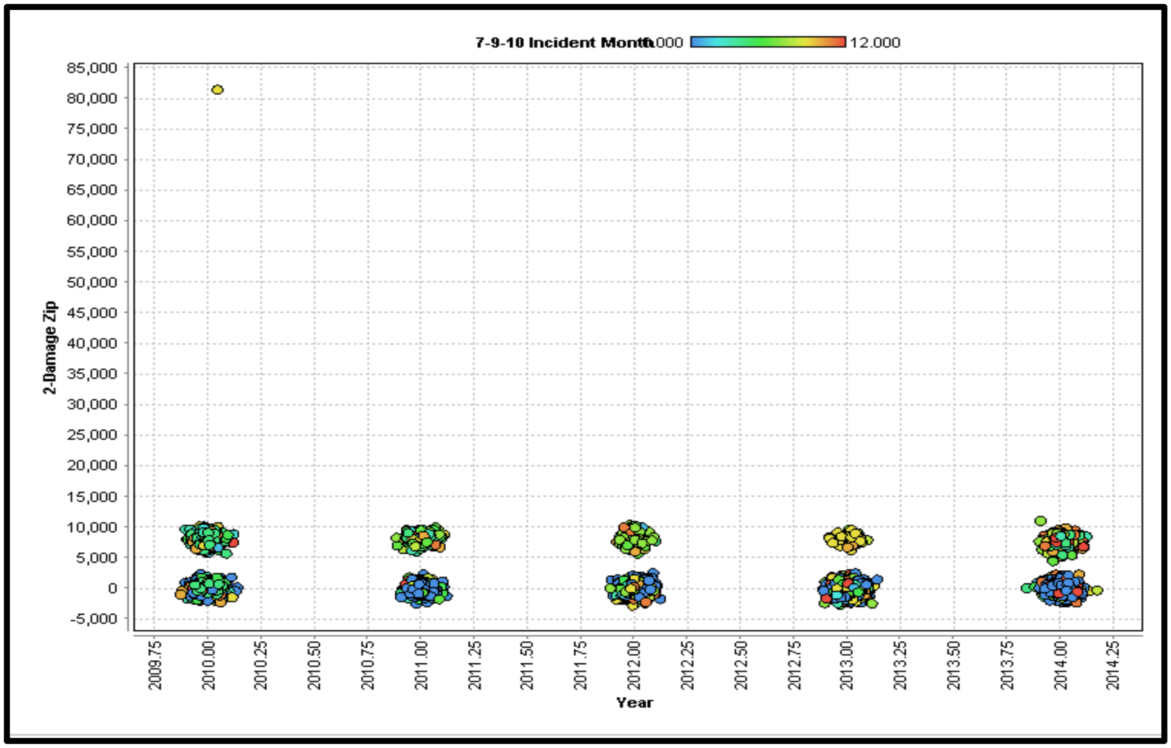


Figure 39: Damaged Data plotting Zip code & Year (phas2)

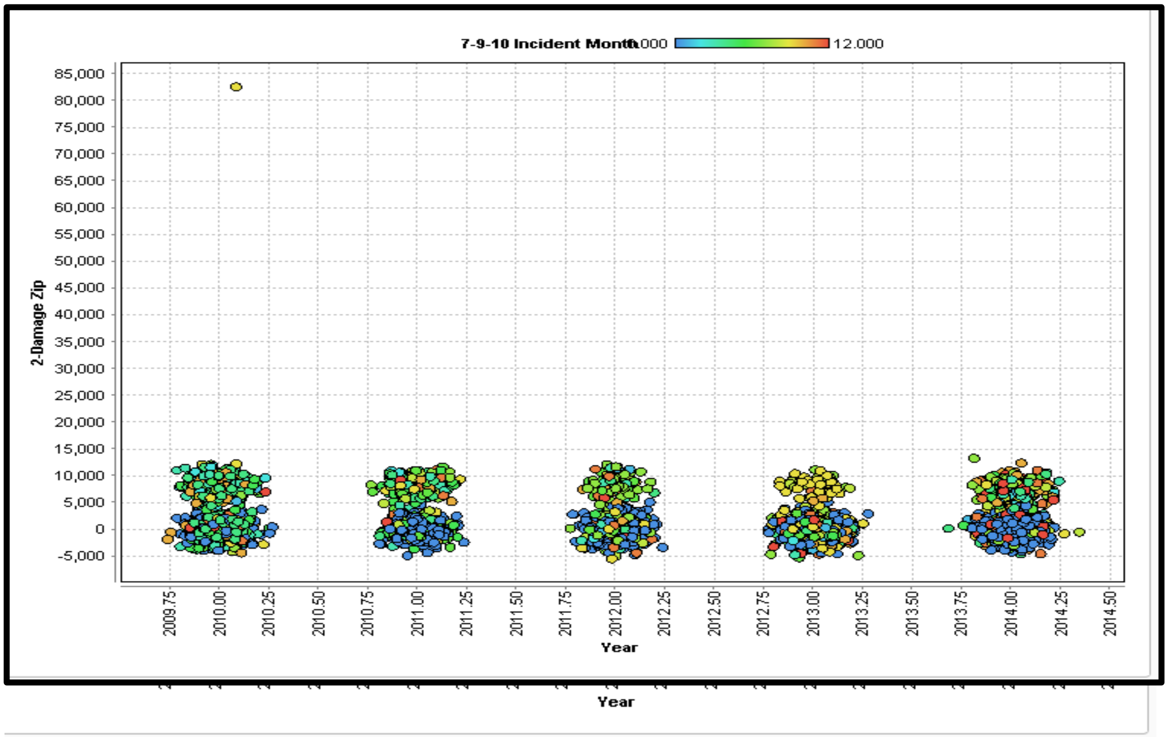


Figure 40: Damaged Data plotting Zip code & Year (phase 3)

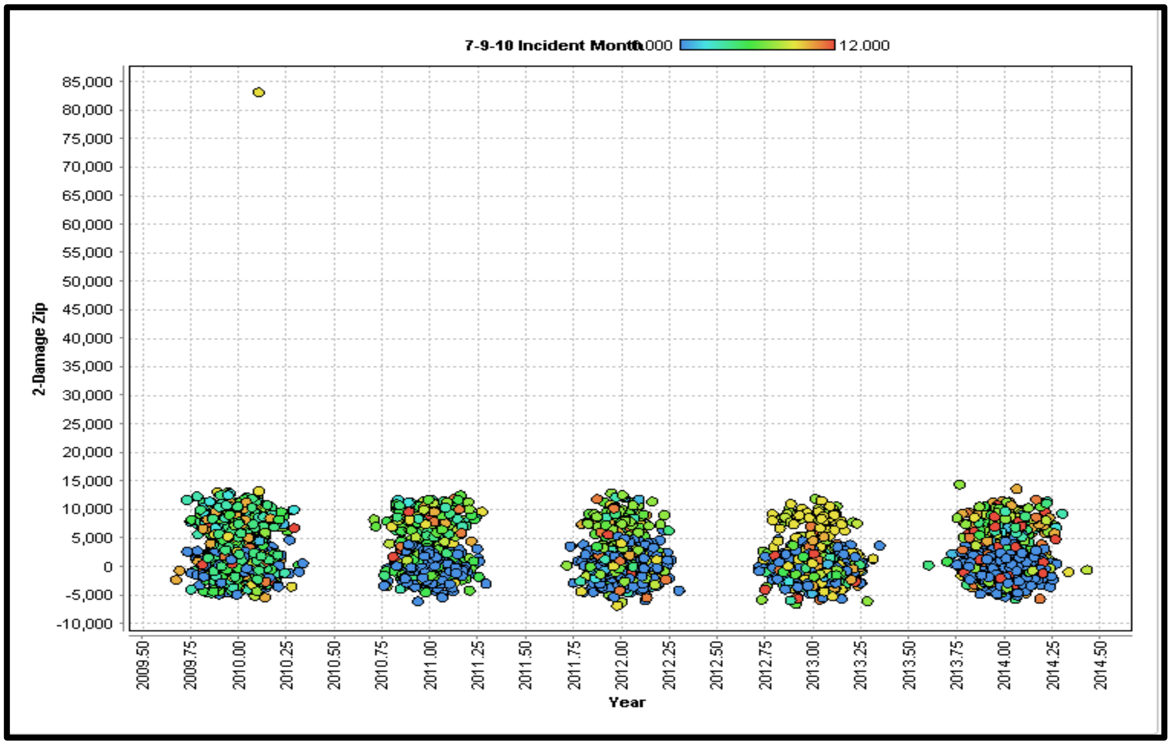


Figure 41: Damaged Data plotting Zip code & Year (phase4)

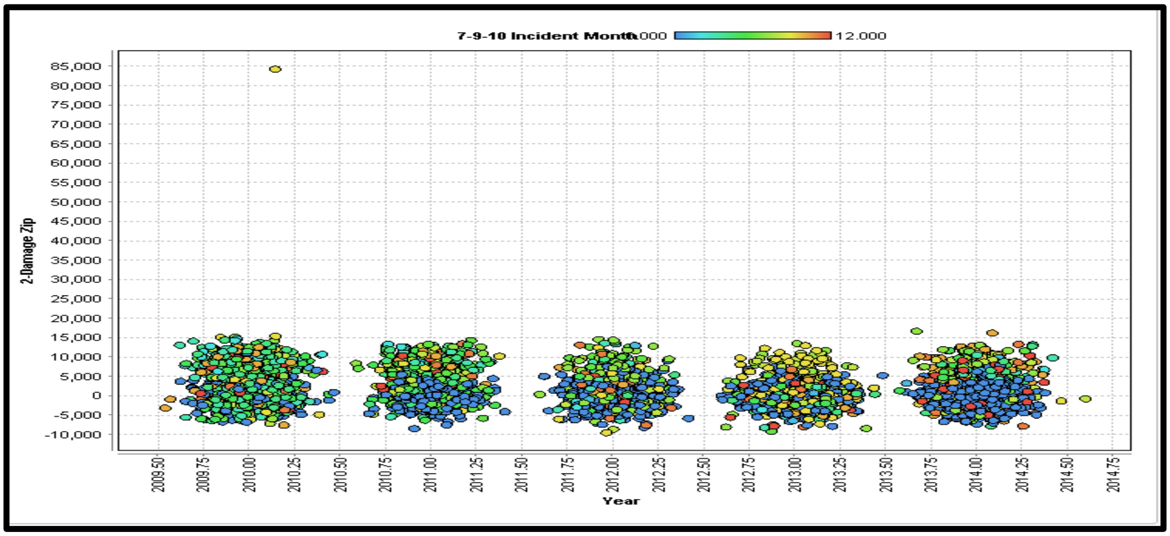


Figure 42: Damaged Data plotting Zip code & Year (phase 5)

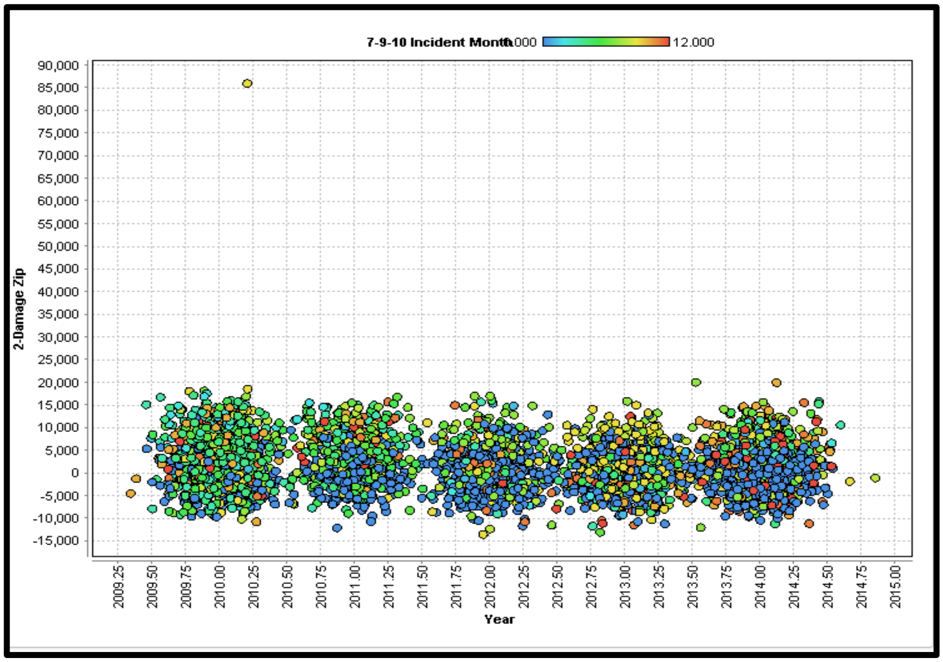


Figure 43: Damaged Data plotting Zip code & Year (phase 6)

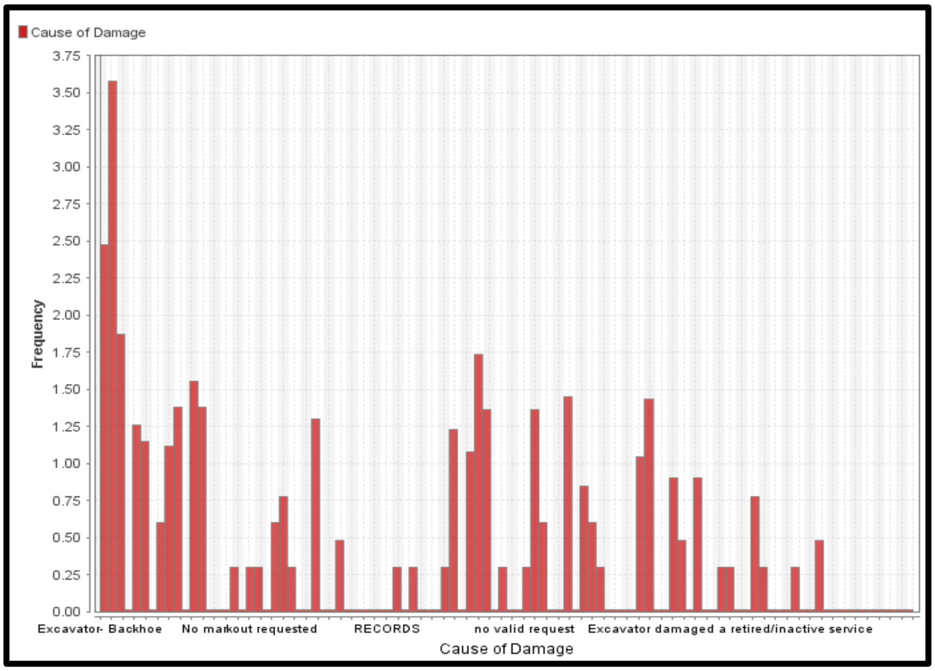


Figure 44 Classification by Cause of Damage

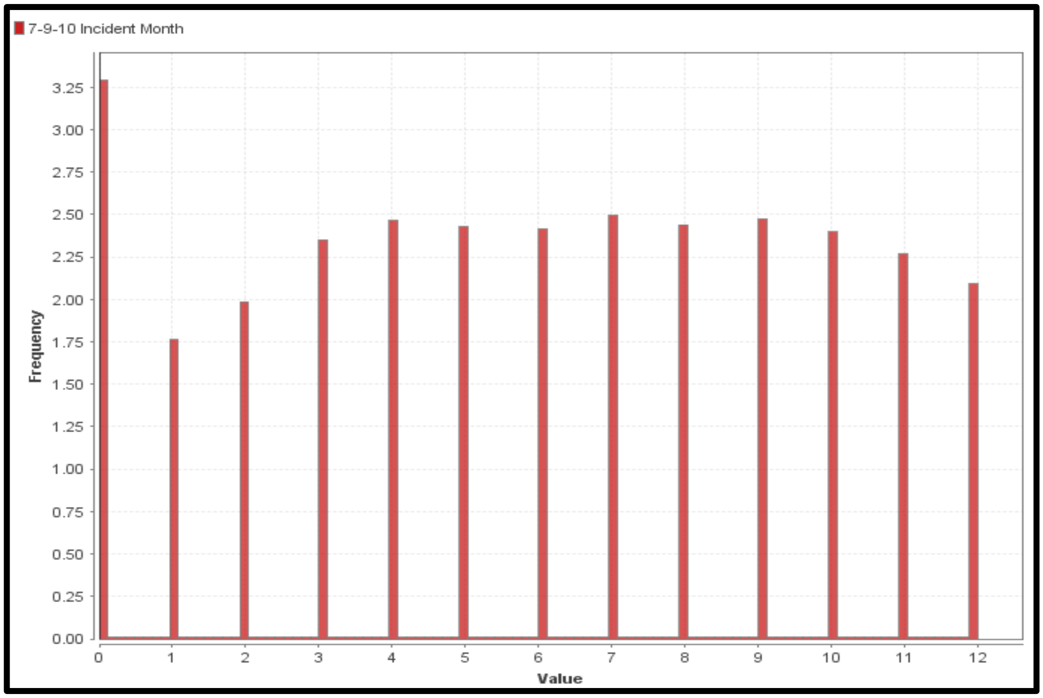


Figure 46: Damage Frequency by Month

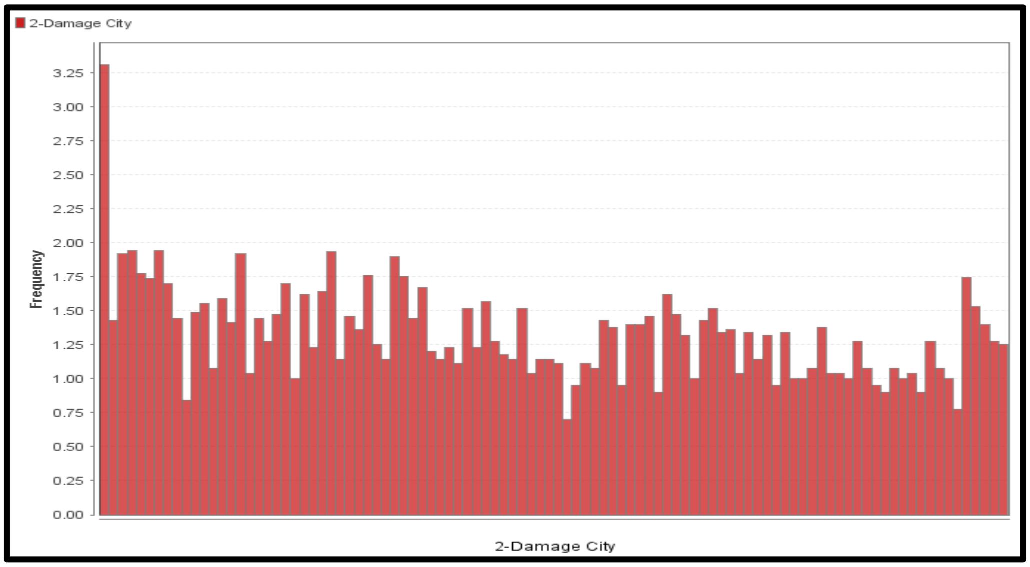


Figure 45: Damage Frequency by City

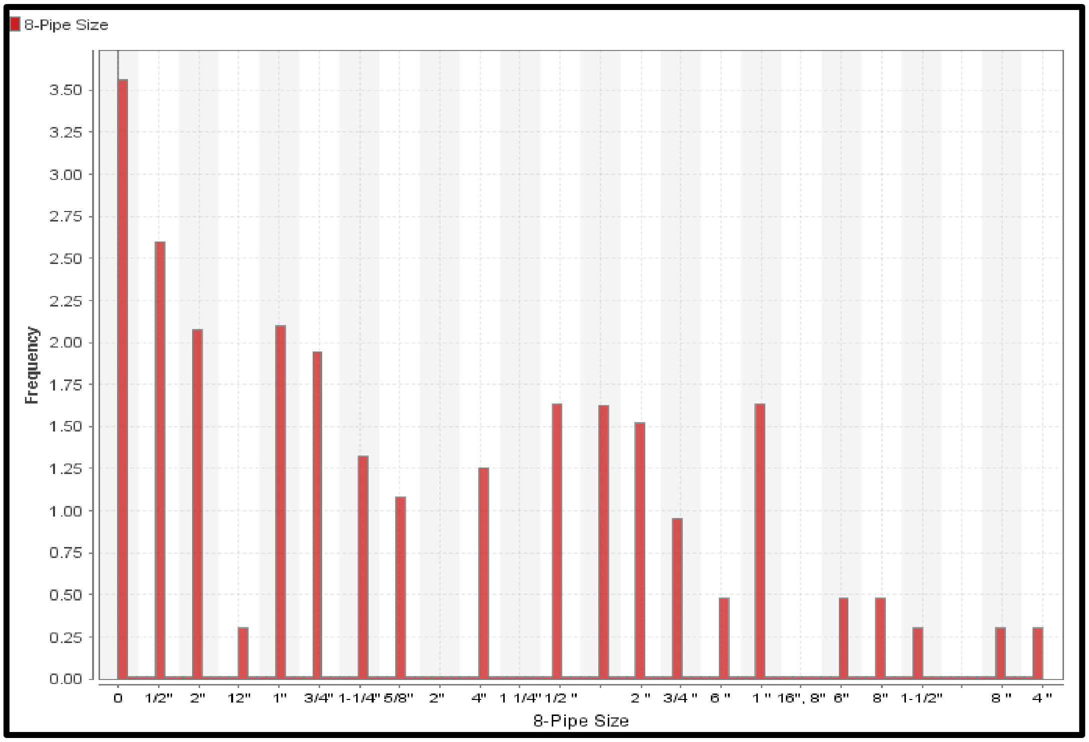


Figure 47: Damage classification by Pipe Size

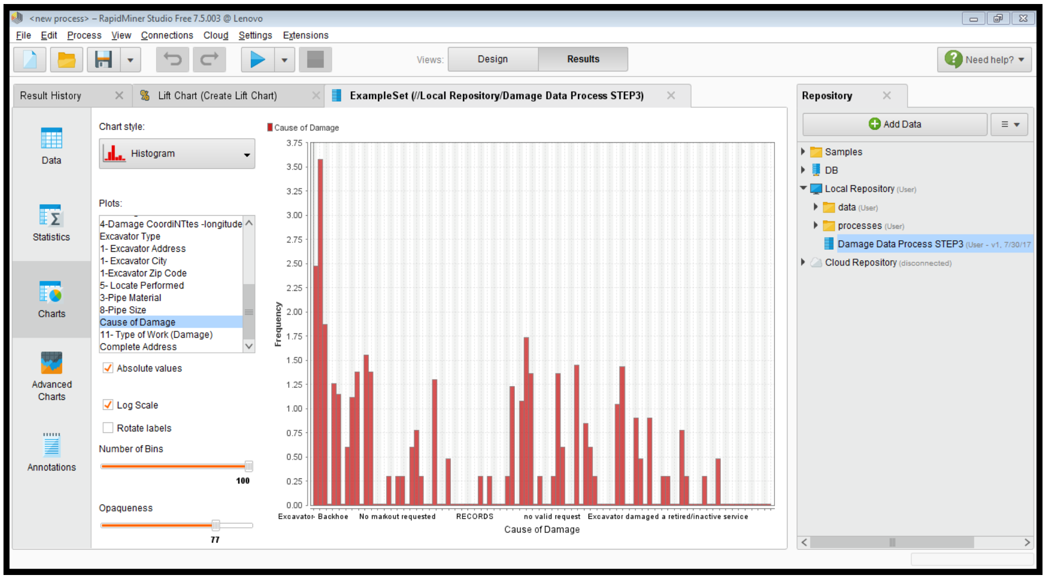


Chart style: Histogram

Plots:
4-Damage CoordiNTtes -longitude
Excavator Type
1- Excavator Address
1- Excavator City
1-Excavator Zip Code
5- Locate Performed
3-Pipe Material
8-Pipe Size
Cause of Damage
11- Type of Work (Damage)
Complete Address

☑ Absolute values

☑ Log Scale

□ Rotate labels

Number of Bins: 100

Opaqueness: 77

Figure 48:Damage Classification by Cause of Damage

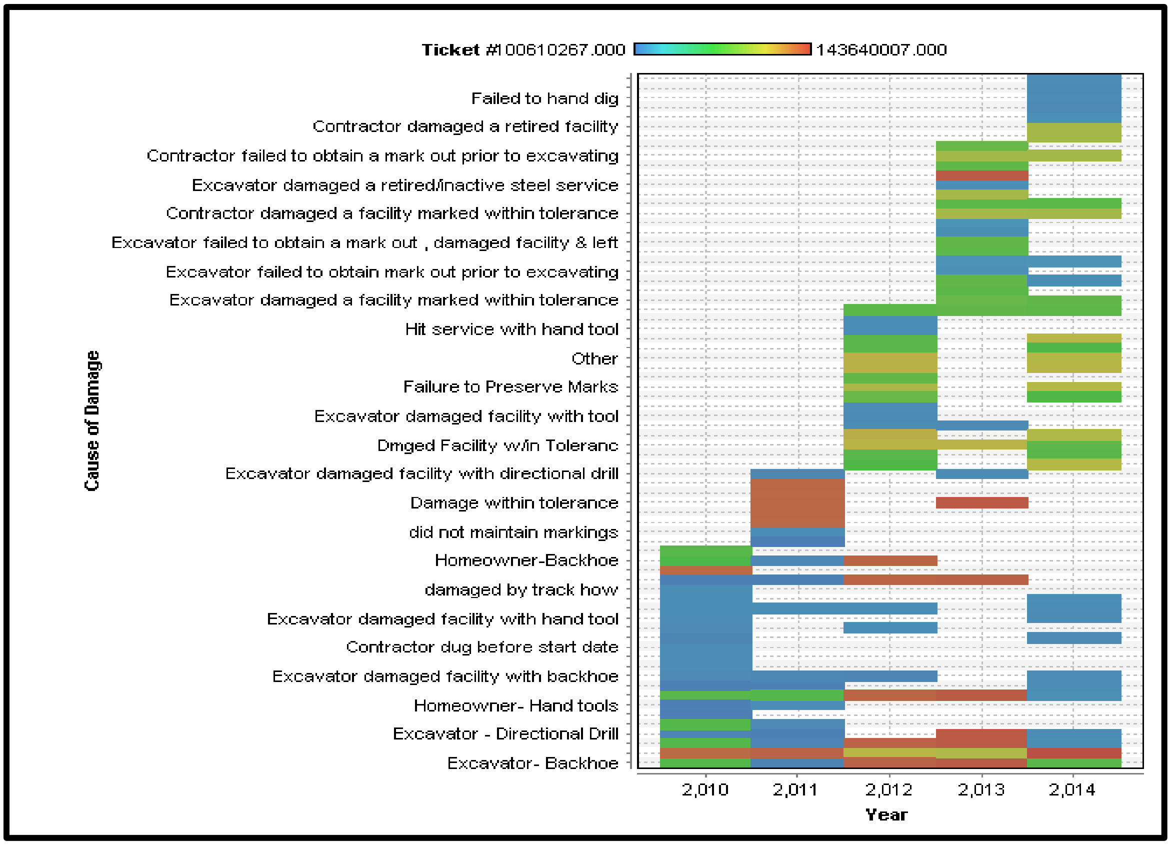


Figure 49:Reason for Damage

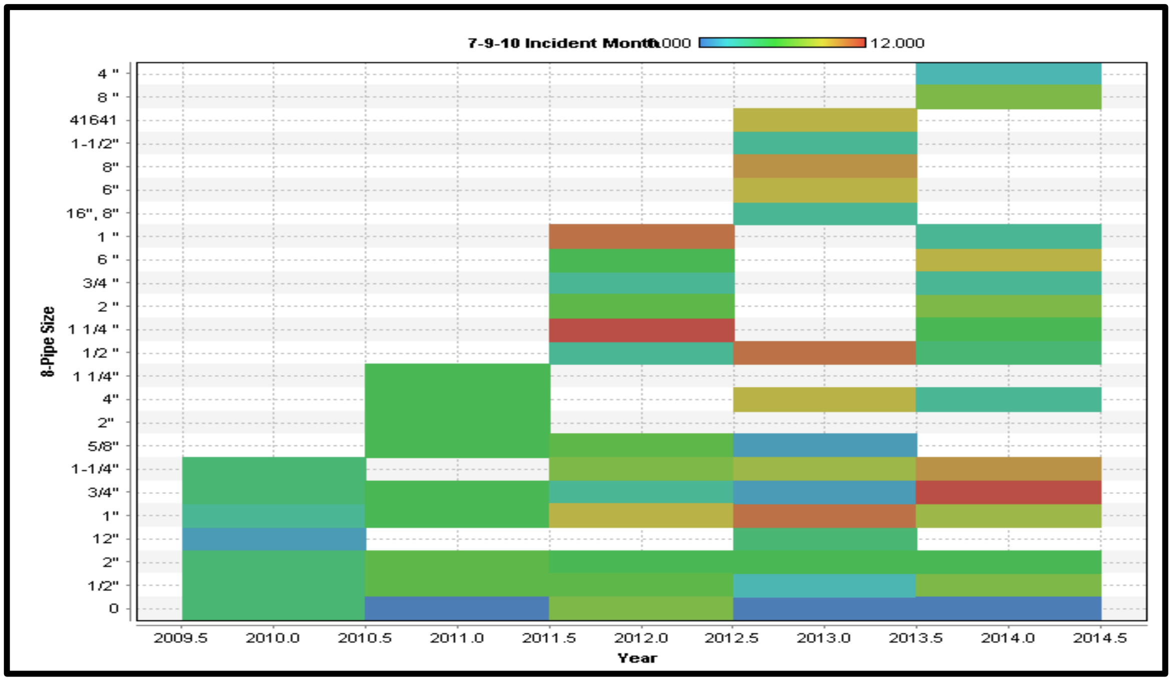


Figure 50: Relationship between Pipe Size & Year

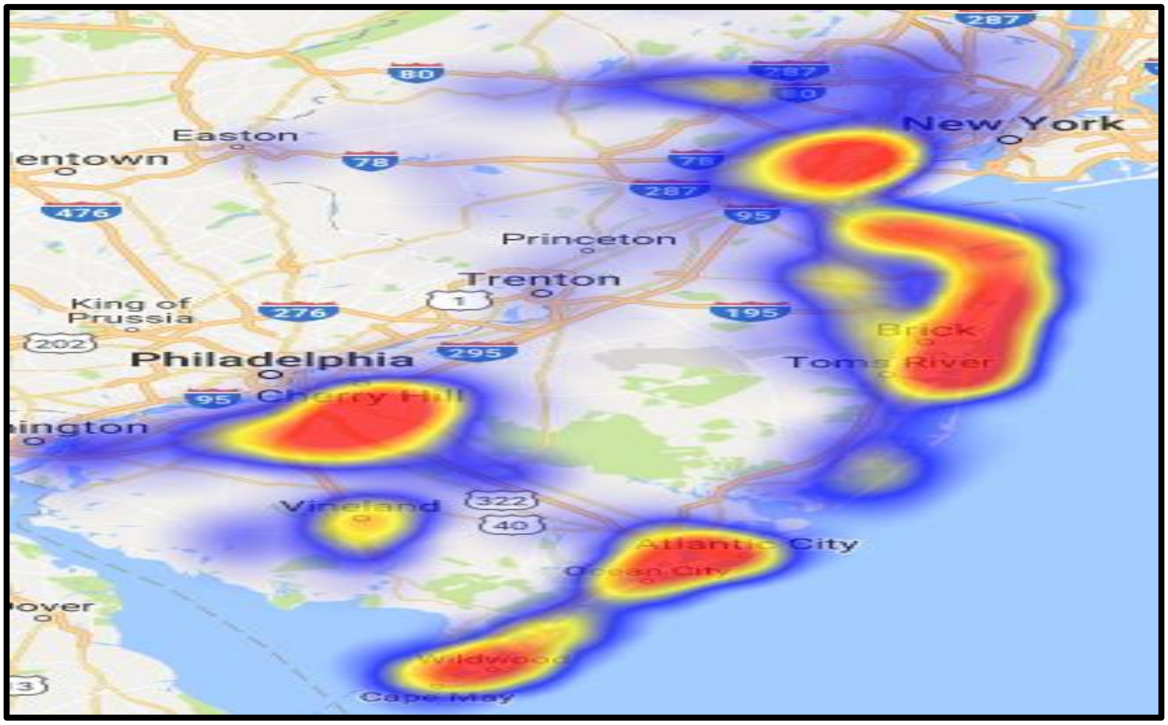


Figure 51:Plotted damages per Year

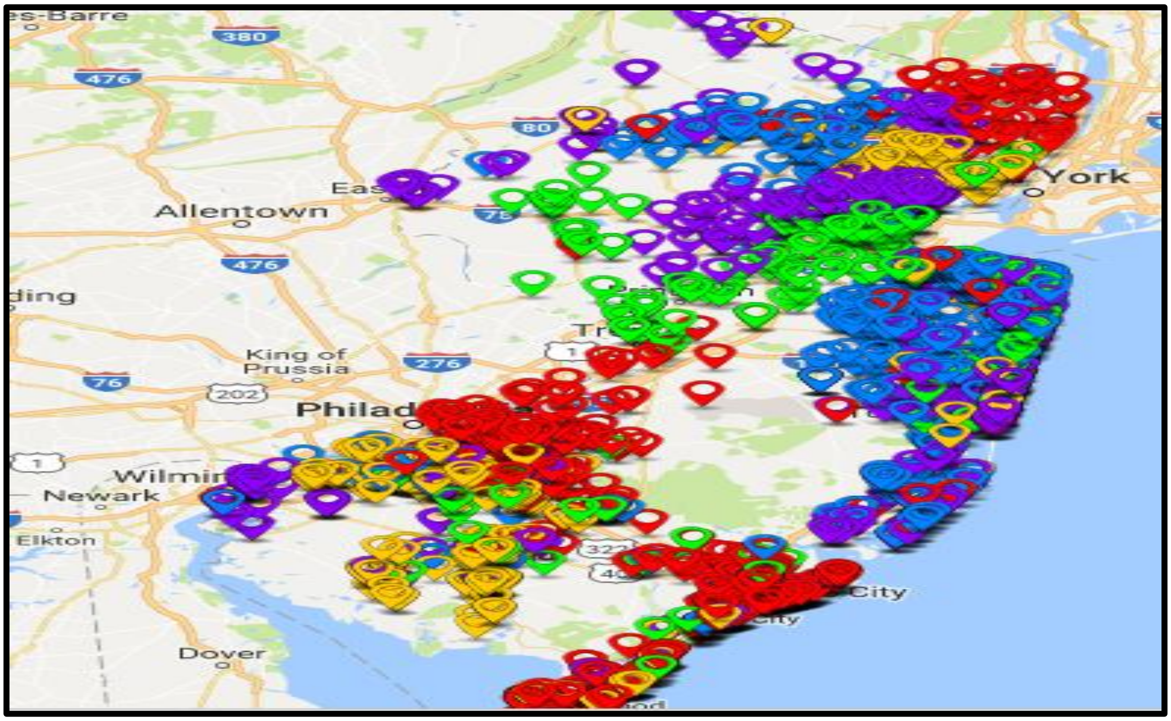


Figure 52: Heat Maps Shows Affected Areas

CHAPTER FIVE: IDENTIFICATION OF DOMENET RISK FACTORS BY USING LOGISTIC REGRESSION PREDICTIVE MODEL

1.27 Introduction

This chapter describes developing a predictive model by using machine learning algorithms to predict the gas pipe damages and related important risk factors. This chapter consists of Introduction, Data Description, and data Preparation, Modeling Methodology, Machine Learning Algorithms used in the analysis, Processing The Predictive Model, Model Testing, and Validation, Discussion of Results. Thischapter will answer the following research question:

Design risk effective model to predict future important risk factors in UG gas pipe damages by using Machine Learning Algorithms by studying the past underground gas line damages in urban congested cities?

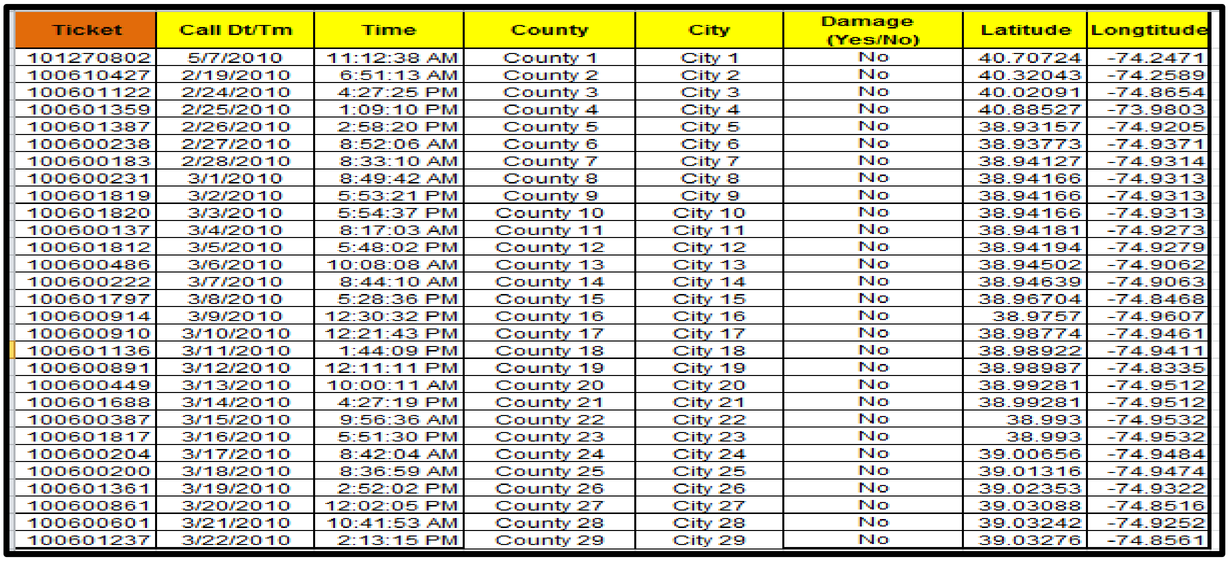
Gas pipe data is used to develop the risk predictive model and verify the validation and testing for the model. Four Machine Learning Algorithm were used in developing the predictive model, Logistic Regression, Support Vector Machine, Random Forest, and k-nearest neighbor’s algorithm. The reason for using multiple algorithm because each and algorithm has its unique features in terms of inputs, analysis, and results. In addition, later on this chapter the risk factors were categorized based on the outputs. Statistical analysis was introduced as well to further analyze the important factors and how they interact with each other.

1.28 Preparing Data Features

The Data was all assembled in one file, as can be seen in figure 53. Each raw has a ticket number, and the date of the incident; either this incident was regular maintenance, or damage incident. The third column represents the time of the incident available for both damaged and undamaged data. The place (i.e., city and county) of the ticket was identified and added to the data. Cross referencing was performed for all the damaged data, every ticket request was checked if it was damaged or undamaged: for damaged (YES) was used, and for undamaged (NO) was used. In more details, (YES) means the one call center was called and the service was requested. However, the underground gas pipe was damaged. (NO) means the agency was called, service was requested, and the maintenance was performed. However, the underground gas pipe was not damaged.

In addition, the damaged tickets were received in different files than the undamaged data. Thus cross referencing, and compiling the data in one file was important task to perform before starting the data preparation.

As mentioned in the previous section, the damaged data was received with latitude, and longitude. However, undamaged data did not have latitude, and longitude. Thus, Server was built to determine the Latitude, and Longitude for the undamaged data, which is important feature in deriving new attributes, and specify the exact location for the underground gas pipeline. See figure 53 for columns ticket #, data, time, county, city, damage (YES/NO), Longitude, and Latitude.

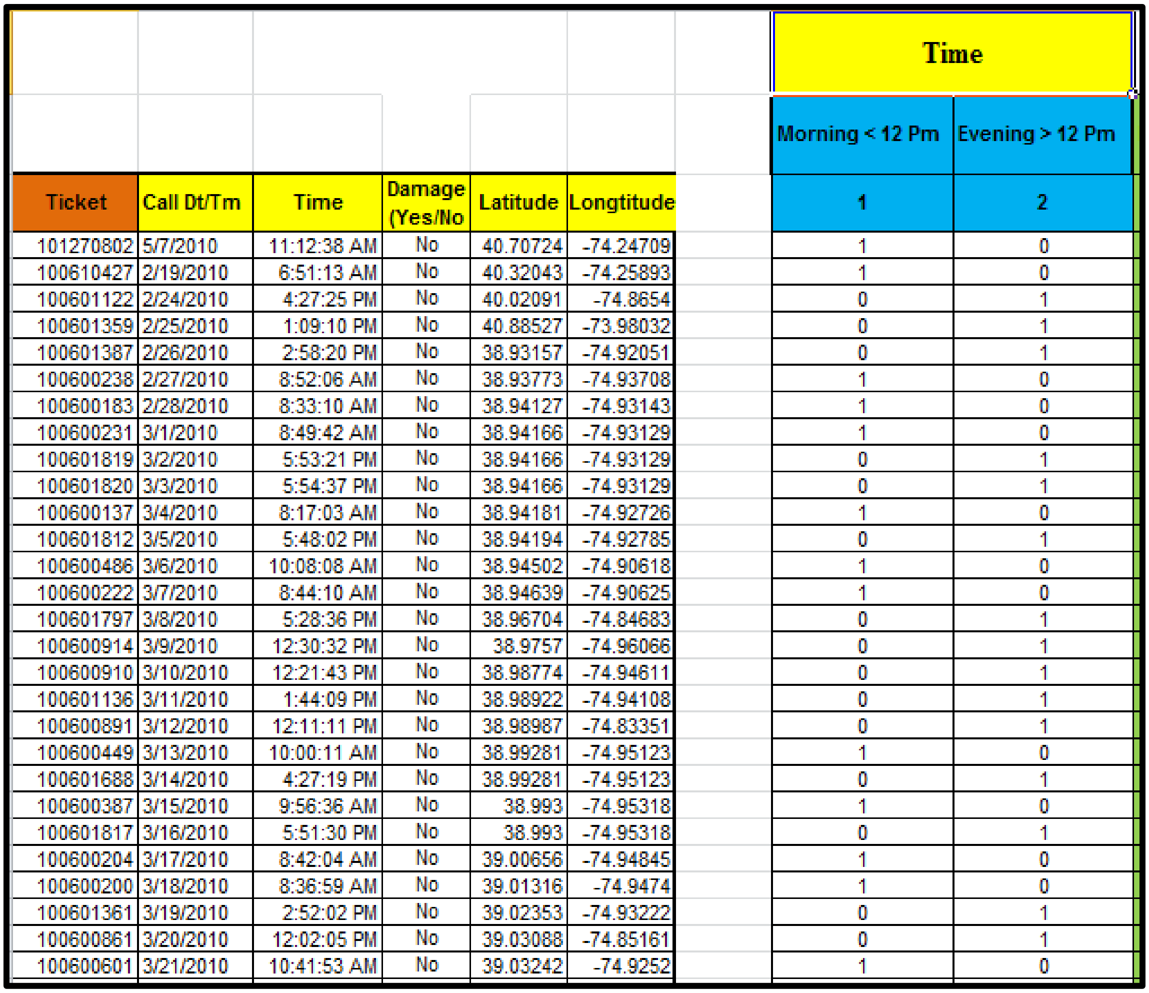


Ticket	Call Dt/Tm	Time	County	City	Damage (Yes/No)	Latitude	Longtitude
101270802	5/7/2010	11:12:38 AM	County 1	City 1	No	40.70724	-74.2471
100610427	2/19/2010	6:51:13 AM	County 2	City 2	No	40.32043	-74.2589
100601122	2/24/2010	4:27:25 PM	County 3	City 3	No	40.02091	-74.8654
100601359	2/25/2010	1:09:10 PM	County 4	City 4	No	40.88527	-73.9803
100601387	2/26/2010	2:58:20 PM	County 5	City 5	No	38.93157	-74.9205
100600238	2/27/2010	8:52:06 AM	County 6	City 6	No	38.93773	-74.9371
100600183	2/28/2010	8:33:10 AM	County 7	City 7	No	38.94127	-74.9314
100600231	3/1/2010	8:49:42 AM	County 8	City 8	No	38.94166	-74.9313
100601819	3/2/2010	5:53:21 PM	County 9	City 9	No	38.94166	-74.9313
100601820	3/3/2010	5:54:37 PM	County 10	City 10	No	38.94166	-74.9313
100600137	3/4/2010	8:17:03 AM	County 11	City 11	No	38.94181	-74.9273
100601812	3/5/2010	5:48:02 PM	County 12	City 12	No	38.94194	-74.9279
100600486	3/6/2010	10:08:08 AM	County 13	City 13	No	38.94502	-74.9062
100600222	3/7/2010	8:44:10 AM	County 14	City 14	No	38.94639	-74.9063
100601797	3/8/2010	5:28:36 PM	County 15	City 15	No	38.96704	-74.8468
100600914	3/9/2010	12:30:32 PM	County 16	City 16	No	38.9757	-74.9607
100600910	3/10/2010	12:21:43 PM	County 17	City 17	No	38.98774	-74.9461
100601136	3/11/2010	1:44:09 PM	County 18	City 18	No	38.98922	-74.9411
100600891	3/12/2010	12:11:11 PM	County 19	City 19	No	38.98987	-74.8335
100600449	3/13/2010	10:00:11 AM	County 20	City 20	No	38.99281	-74.9512
100601688	3/14/2010	4:27:19 PM	County 21	City 21	No	38.99281	-74.9512
100600387	3/15/2010	9:56:36 AM	County 22	City 22	No	38.993	-74.9532
100601817	3/16/2010	5:51:30 PM	County 23	City 23	No	38.993	-74.9532
100600204	3/17/2010	8:42:04 AM	County 24	City 24	No	39.00656	-74.9484
100600200	3/18/2010	8:36:59 AM	County 25	City 25	No	39.01316	-74.9474
100601361	3/19/2010	2:52:02 PM	County 26	City 26	No	39.02353	-74.9322
100600861	3/20/2010	12:02:05 PM	County 27	City 27	No	39.03088	-74.8516
100600601	3/21/2010	10:41:53 AM	County 28	City 28	No	39.03242	-74.9252
100601237	3/22/2010	2:13:15 PM	County 29	City 29	No	39.03276	-74.8561

Figure 53: Assembeled Data (Damage& undamage)

1.29 Developing New Attributes

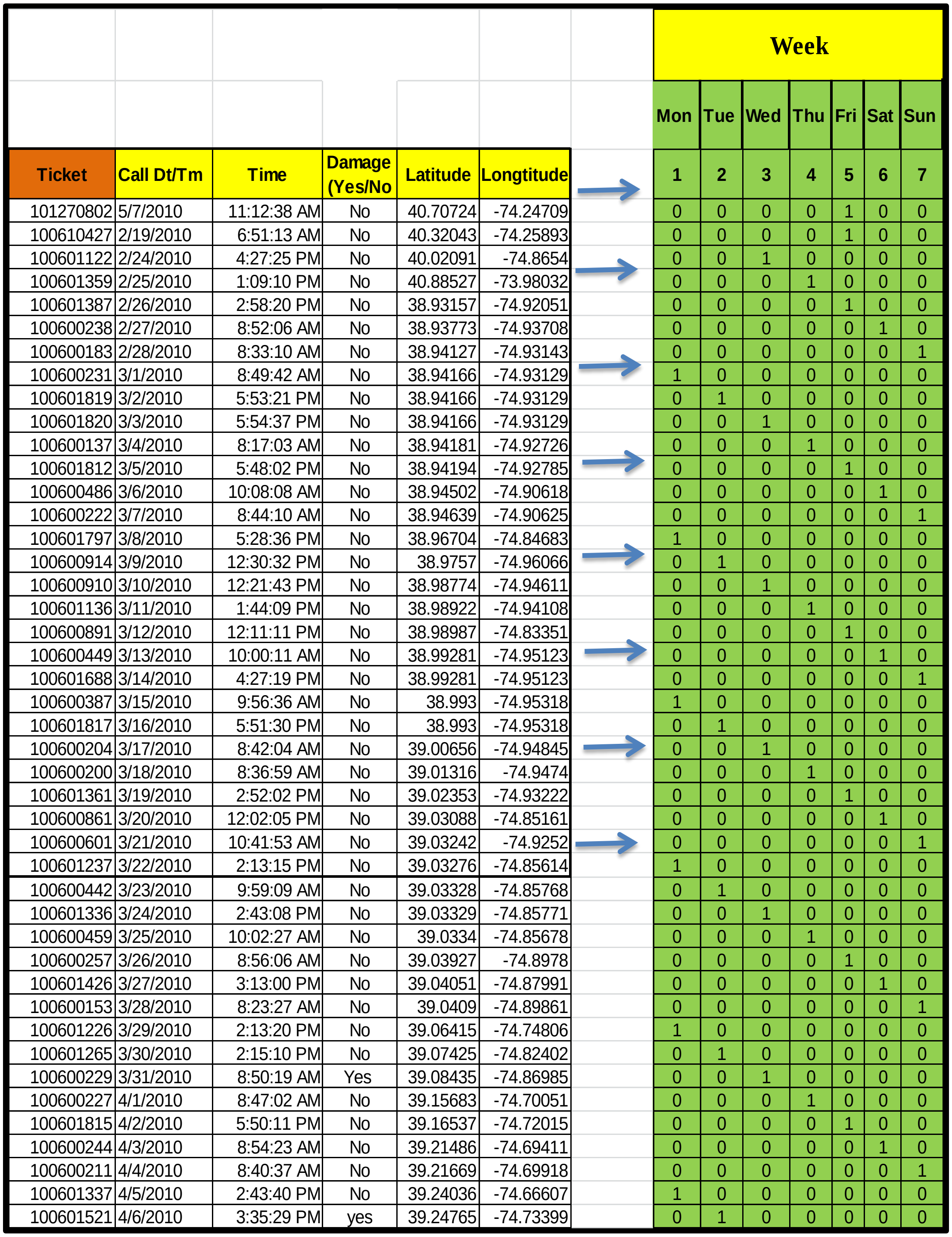
1- Time (AM, or PM), the time of incident was categorized into two categories. Any time before 12 AM was categorized as equal to 1. Any time after 12 PM was categorized as PM equal to 1. Excel F function was used to map out the time into two categories AM, PM. See figure 54.



							Time	
							Morning < 12 Pm	Evening > 12 Pm
Ticket	Call Dt/Tm	Time	Damage (Yes/No	Latitude	Longtitude		1	2
101270802	5/7/2010	11:12:38 AM	No	40.70724	-74.24709		1	0
100610427	2/19/2010	6:51:13 AM	No	40.32043	-74.25893		1	0
100601122	2/24/2010	4:27:25 PM	No	40.02091	-74.8654		0	1
100601359	2/25/2010	1:09:10 PM	No	40.88527	-73.98032		0	1
100601387	2/26/2010	2:58:20 PM	No	38.93157	-74.92051		0	1
100600238	2/27/2010	8:52:06 AM	No	38.93773	-74.93708		1	0
100600183	2/28/2010	8:33:10 AM	No	38.94127	-74.93143		1	0
100600231	3/1/2010	8:49:42 AM	No	38.94166	-74.93129		1	0
100601819	3/2/2010	5:53:21 PM	No	38.94166	-74.93129		0	1
100601820	3/3/2010	5:54:37 PM	No	38.94166	-74.93129		0	1
100600137	3/4/2010	8:17:03 AM	No	38.94181	-74.92726		1	0
100601812	3/5/2010	5:48:02 PM	No	38.94194	-74.92785		0	1
100600486	3/6/2010	10:08:08 AM	No	38.94502	-74.90618		1	0
100600222	3/7/2010	8:44:10 AM	No	38.94639	-74.90625		1	0
100601797	3/8/2010	5:28:36 PM	No	38.96704	-74.84683		0	1
100600914	3/9/2010	12:30:32 PM	No	38.9757	-74.96066		0	1
100600910	3/10/2010	12:21:43 PM	No	38.98774	-74.94611		0	1
100601136	3/11/2010	1:44:09 PM	No	38.98922	-74.94108		0	1
100600891	3/12/2010	12:11:11 PM	No	38.98987	-74.83351		0	1
100600449	3/13/2010	10:00:11 AM	No	38.99281	-74.95123		1	0
100601688	3/14/2010	4:27:19 PM	No	38.99281	-74.95123		0	1
100600387	3/15/2010	9:56:36 AM	No	38.993	-74.95318		1	0
100601817	3/16/2010	5:51:30 PM	No	38.993	-74.95318		0	1
100600204	3/17/2010	8:42:04 AM	No	39.00656	-74.94845		1	0
100600200	3/18/2010	8:36:59 AM	No	39.01316	-74.9474		1	0
100601361	3/19/2010	2:52:02 PM	No	39.02353	-74.93222		0	1
100600861	3/20/2010	12:02:05 PM	No	39.03088	-74.85161		0	1
100600601	3/21/2010	10:41:53 AM	No	39.03242	-74.9252		1	0

Figure 54 Time(AM/PM) Attribute

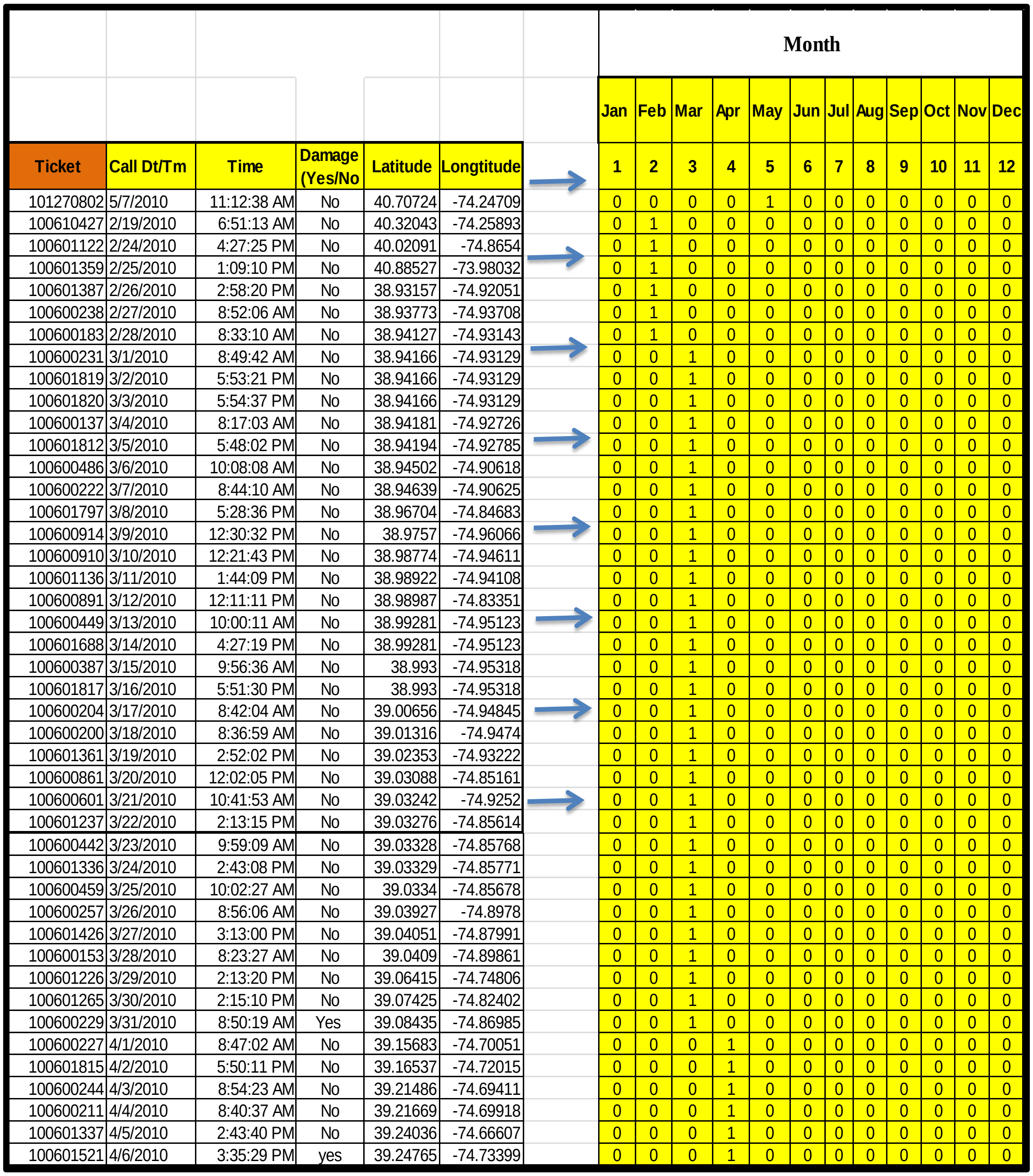
2- Week (Mon, Tue, Wed, Thur, Fri, Sat, Sun), the week was selected as attribute and categorized into 7 categories. So the function will look where is the ticket falling in and place the ticket in the day of the week which happened in. See figure 55.



						Week						
						Mon	Tue	Wed	Thu	Fri	Sat	Sun
Ticket	Call Dt/Tm	Time	Damage (Yes/No	Latitude	Longtitude	1	2	3	4	5	6	7
101270802	5/7/2010	11:12:38 AM	No	40.70724	-74.24709	0	0	0	0	1	0	0
100610427	2/19/2010	6:51:13 AM	No	40.32043	-74.25893	0	0	0	0	1	0	0
100601122	2/24/2010	4:27:25 PM	No	40.02091	-74.8654	0	0	1	0	0	0	0
100601359	2/25/2010	1:09:10 PM	No	40.88527	-73.98032	0	0	0	1	0	0	0
100601387	2/26/2010	2:58:20 PM	No	38.93157	-74.92051	0	0	0	0	1	0	0
100600238	2/27/2010	8:52:06 AM	No	38.93773	-74.93708	0	0	0	0	0	1	0
100600183	2/28/2010	8:33:10 AM	No	38.94127	-74.93143	0	0	0	0	0	0	1
100600231	3/1/2010	8:49:42 AM	No	38.94166	-74.93129	1	0	0	0	0	0	0
100601819	3/2/2010	5:53:21 PM	No	38.94166	-74.93129	0	1	0	0	0	0	0
100601820	3/3/2010	5:54:37 PM	No	38.94166	-74.93129	0	0	1	0	0	0	0
100600137	3/4/2010	8:17:03 AM	No	38.94181	-74.92726	0	0	0	1	0	0	0
100601812	3/5/2010	5:48:02 PM	No	38.94194	-74.92785	0	0	0	0	1	0	0
100600486	3/6/2010	10:08:08 AM	No	38.94502	-74.90618	0	0	0	0	0	1	0
100600222	3/7/2010	8:44:10 AM	No	38.94639	-74.90625	0	0	0	0	0	0	1
100601797	3/8/2010	5:28:36 PM	No	38.96704	-74.84683	1	0	0	0	0	0	0
100600914	3/9/2010	12:30:32 PM	No	38.9757	-74.96066	0	1	0	0	0	0	0
100600910	3/10/2010	12:21:43 PM	No	38.98774	-74.94611	0	0	1	0	0	0	0
100601136	3/11/2010	1:44:09 PM	No	38.98922	-74.94108	0	0	0	1	0	0	0
100600891	3/12/2010	12:11:11 PM	No	38.98987	-74.83351	0	0	0	0	1	0	0
100600449	3/13/2010	10:00:11 AM	No	38.99281	-74.95123	0	0	0	0	0	1	0
100601688	3/14/2010	4:27:19 PM	No	38.99281	-74.95123	0	0	0	0	0	0	1
100600387	3/15/2010	9:56:36 AM	No	38.993	-74.95318	1	0	0	0	0	0	0
100601817	3/16/2010	5:51:30 PM	No	38.993	-74.95318	0	1	0	0	0	0	0
100600204	3/17/2010	8:42:04 AM	No	39.00656	-74.94845	0	0	1	0	0	0	0
100600200	3/18/2010	8:36:59 AM	No	39.01316	-74.9474	0	0	0	1	0	0	0
100601361	3/19/2010	2:52:02 PM	No	39.02353	-74.93222	0	0	0	0	1	0	0
100600861	3/20/2010	12:02:05 PM	No	39.03088	-74.85161	0	0	0	0	0	1	0
100600601	3/21/2010	10:41:53 AM	No	39.03242	-74.9252	0	0	0	0	0	0	1
100601237	3/22/2010	2:13:15 PM	No	39.03276	-74.85614	1	0	0	0	0	0	0
100600442	3/23/2010	9:59:09 AM	No	39.03328	-74.85768	0	1	0	0	0	0	0
100601336	3/24/2010	2:43:08 PM	No	39.03329	-74.85771	0	0	1	0	0	0	0
100600459	3/25/2010	10:02:27 AM	No	39.0334	-74.85678	0	0	0	1	0	0	0
100600257	3/26/2010	8:56:06 AM	No	39.03927	-74.8978	0	0	0	0	1	0	0
100601426	3/27/2010	3:13:00 PM	No	39.04051	-74.87991	0	0	0	0	0	1	0
100600153	3/28/2010	8:23:27 AM	No	39.0409	-74.89861	0	0	0	0	0	0	1
100601226	3/29/2010	2:13:20 PM	No	39.06415	-74.74806	1	0	0	0	0	0	0
100601265	3/30/2010	2:15:10 PM	No	39.07425	-74.82402	0	1	0	0	0	0	0
100600229	3/31/2010	8:50:19 AM	Yes	39.08435	-74.86985	0	0	1	0	0	0	0
100600227	4/1/2010	8:47:02 AM	No	39.15683	-74.70051	0	0	0	1	0	0	0
100601815	4/2/2010	5:50:11 PM	No	39.16537	-74.72015	0	0	0	0	1	0	0
100600244	4/3/2010	8:54:23 AM	No	39.21486	-74.69411	0	0	0	0	0	1	0
100600211	4/4/2010	8:40:37 AM	No	39.21669	-74.69918	0	0	0	0	0	0	1
100601337	4/5/2010	2:43:40 PM	No	39.24036	-74.66607	1	0	0	0	0	0	0
100601521	4/6/2010	3:35:29 PM	yes	39.24765	-74.73399	0	1	0	0	0	0	0

Figure 55 Week (Mon, Tue, Wed, Thu,Fri,Sat,Sun) Attribute

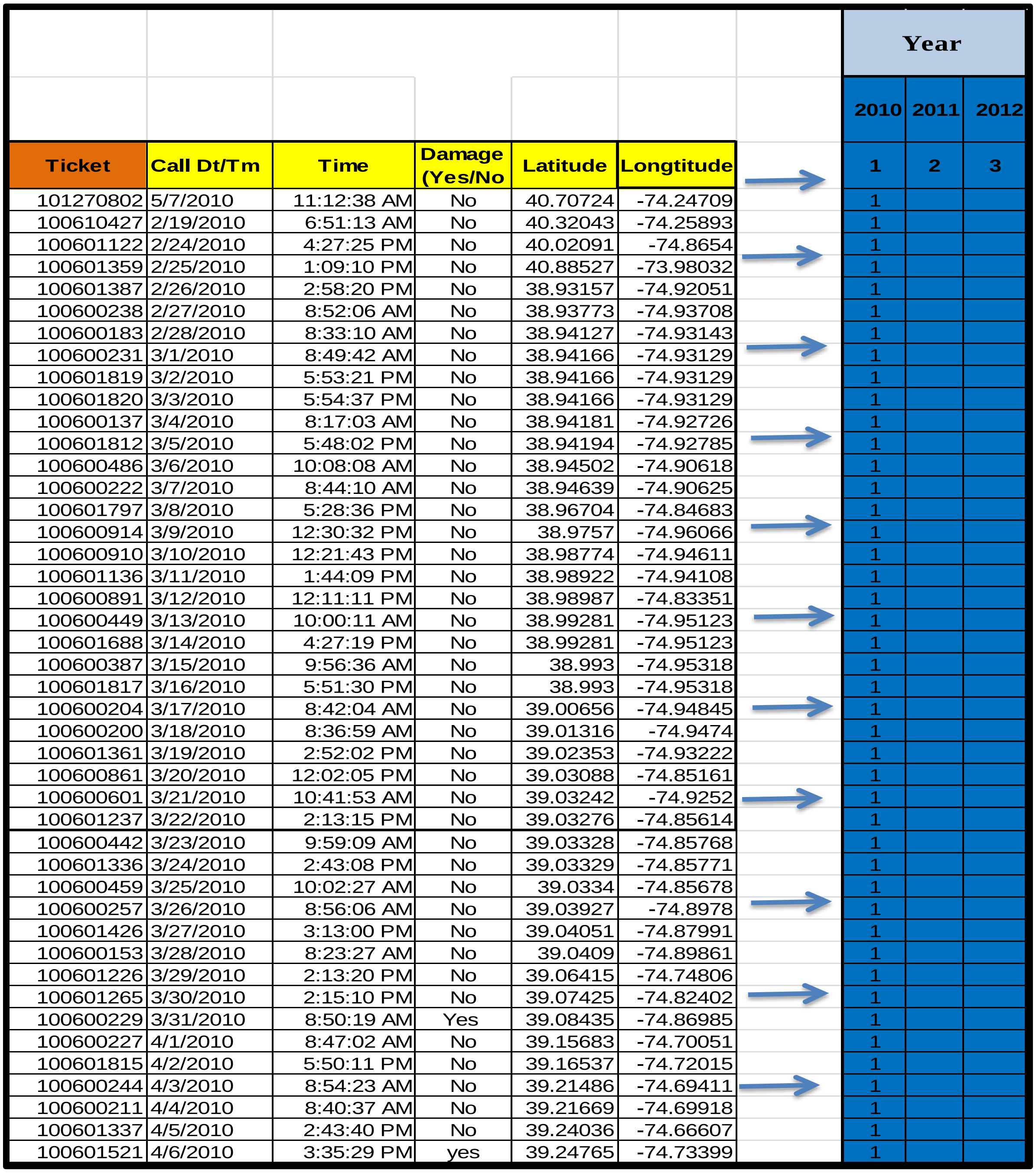
3- Month (Jan, Feb, Mar, April, May, Jun, July, Aug, Sept, Oct, Nov, Dec), the Month was selected as attribute and categorized into 12 categories: so the function will look where is the ticket falling in and place the ticket in the month of the year which happened in. See figure 56.



						Month											
						Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Ticket	Call Dt/Tm	Time	Damage (Yes/No	Latitude	Longtitude	1	2	3	4	5	6	7	8	9	10	11	12
101270802	5/7/2010	11:12:38 AM	No	40.70724	-74.24709	0	0	0	0	1	0	0	0	0	0	0	0
100610427	2/19/2010	6:51:13 AM	No	40.32043	-74.25893	0	1	0	0	0	0	0	0	0	0	0	0
100601122	2/24/2010	4:27:25 PM	No	40.02091	-74.8654	0	1	0	0	0	0	0	0	0	0	0	0
100601359	2/25/2010	1:09:10 PM	No	40.88527	-73.98032	0	1	0	0	0	0	0	0	0	0	0	0
100601387	2/26/2010	2:58:20 PM	No	38.93157	-74.92051	0	1	0	0	0	0	0	0	0	0	0	0
100600238	2/27/2010	8:52:06 AM	No	38.93773	-74.93708	0	1	0	0	0	0	0	0	0	0	0	0
100600183	2/28/2010	8:33:10 AM	No	38.94127	-74.93143	0	1	0	0	0	0	0	0	0	0	0	0
100600231	3/1/2010	8:49:42 AM	No	38.94166	-74.93129	0	0	1	0	0	0	0	0	0	0	0	0
100601819	3/2/2010	5:53:21 PM	No	38.94166	-74.93129	0	0	1	0	0	0	0	0	0	0	0	0
100601820	3/3/2010	5:54:37 PM	No	38.94166	-74.93129	0	0	1	0	0	0	0	0	0	0	0	0
100600137	3/4/2010	8:17:03 AM	No	38.94181	-74.92726	0	0	1	0	0	0	0	0	0	0	0	0
100601812	3/5/2010	5:48:02 PM	No	38.94194	-74.92785	0	0	1	0	0	0	0	0	0	0	0	0
100600486	3/6/2010	10:08:08 AM	No	38.94502	-74.90618	0	0	1	0	0	0	0	0	0	0	0	0
100600222	3/7/2010	8:44:10 AM	No	38.94639	-74.90625	0	0	1	0	0	0	0	0	0	0	0	0
100601797	3/8/2010	5:28:36 PM	No	38.96704	-74.84683	0	0	1	0	0	0	0	0	0	0	0	0
100600914	3/9/2010	12:30:32 PM	No	38.9757	-74.96066	0	0	1	0	0	0	0	0	0	0	0	0
100600910	3/10/2010	12:21:43 PM	No	38.98774	-74.94611	0	0	1	0	0	0	0	0	0	0	0	0
100601136	3/11/2010	1:44:09 PM	No	38.98922	-74.94108	0	0	1	0	0	0	0	0	0	0	0	0
100600891	3/12/2010	12:11:11 PM	No	38.98987	-74.83351	0	0	1	0	0	0	0	0	0	0	0	0
100600449	3/13/2010	10:00:11 AM	No	38.99281	-74.95123	0	0	1	0	0	0	0	0	0	0	0	0
100601688	3/14/2010	4:27:19 PM	No	38.99281	-74.95123	0	0	1	0	0	0	0	0	0	0	0	0
100600387	3/15/2010	9:56:36 AM	No	38.993	-74.95318	0	0	1	0	0	0	0	0	0	0	0	0
100601817	3/16/2010	5:51:30 PM	No	38.993	-74.95318	0	0	1	0	0	0	0	0	0	0	0	0
100600204	3/17/2010	8:42:04 AM	No	39.00656	-74.94845	0	0	1	0	0	0	0	0	0	0	0	0
100600200	3/18/2010	8:36:59 AM	No	39.01316	-74.9474	0	0	1	0	0	0	0	0	0	0	0	0
100601361	3/19/2010	2:52:02 PM	No	39.02353	-74.93222	0	0	1	0	0	0	0	0	0	0	0	0
100600861	3/20/2010	12:02:05 PM	No	39.03088	-74.85161	0	0	1	0	0	0	0	0	0	0	0	0
100600601	3/21/2010	10:41:53 AM	No	39.03242	-74.9252	0	0	1	0	0	0	0	0	0	0	0	0
100601237	3/22/2010	2:13:15 PM	No	39.03276	-74.85614	0	0	1	0	0	0	0	0	0	0	0	0
100600442	3/23/2010	9:59:09 AM	No	39.03328	-74.85768	0	0	1	0	0	0	0	0	0	0	0	0
100601336	3/24/2010	2:43:08 PM	No	39.03329	-74.85771	0	0	1	0	0	0	0	0	0	0	0	0
100600459	3/25/2010	10:02:27 AM	No	39.0334	-74.85678	0	0	1	0	0	0	0	0	0	0	0	0
100600257	3/26/2010	8:56:06 AM	No	39.03927	-74.8978	0	0	1	0	0	0	0	0	0	0	0	0
100601426	3/27/2010	3:13:00 PM	No	39.04051	-74.87991	0	0	1	0	0	0	0	0	0	0	0	0
100600153	3/28/2010	8:23:27 AM	No	39.0409	-74.89861	0	0	1	0	0	0	0	0	0	0	0	0
100601226	3/29/2010	2:13:20 PM	No	39.06415	-74.74806	0	0	1	0	0	0	0	0	0	0	0	0
100601265	3/30/2010	2:15:10 PM	No	39.07425	-74.82402	0	0	1	0	0	0	0	0	0	0	0	0
100600229	3/31/2010	8:50:19 AM	Yes	39.08435	-74.86985	0	0	1	0	0	0	0	0	0	0	0	0
100600227	4/1/2010	8:47:02 AM	No	39.15683	-74.70051	0	0	0	1	0	0	0	0	0	0	0	0
100601815	4/2/2010	5:50:11 PM	No	39.16537	-74.72015	0	0	0	1	0	0	0	0	0	0	0	0
100600244	4/3/2010	8:54:23 AM	No	39.21486	-74.69411	0	0	0	1	0	0	0	0	0	0	0	0
100600211	4/4/2010	8:40:37 AM	No	39.21669	-74.69918	0	0	0	1	0	0	0	0	0	0	0	0
100601337	4/5/2010	2:43:40 PM	No	39.24036	-74.66607	0	0	0	1	0	0	0	0	0	0	0	0
100601521	4/6/2010	3:35:29 PM	yes	39.24765	-74.73399	0	0	0	1	0	0	0	0	0	0	0	0

Figure 56: Months (Jan, Feb,Mar, Apri, May, Jun, Jul, Aug, Sept, Oct, Nov, Dec)

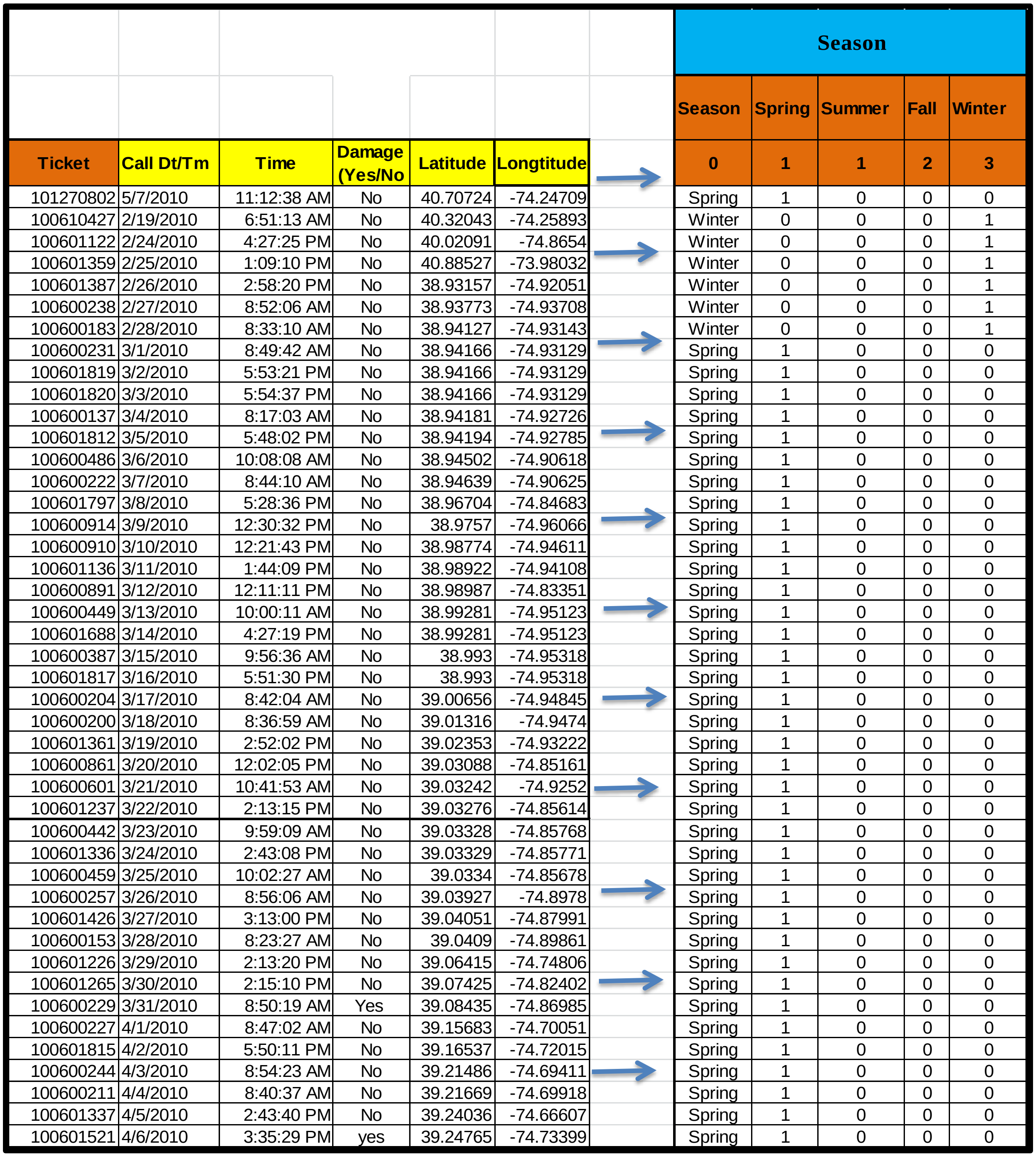
4- Year (2010, 2011, 2012), the Year was selected as attribute and categorized into 3 categories: so the function will look where is the ticket falling in and place the ticket in the Year which happened in. See figure 57.



						Year		
						2010	2011	2012
Ticket	Call Dt/Tm	Time	Damage (Yes/No	Latitude	Longtitude	1	2	3
101270802	5/7/2010	11:12:38 AM	No	40.70724	-74.24709	1		
100610427	2/19/2010	6:51:13 AM	No	40.32043	-74.25893	1		
100601122	2/24/2010	4:27:25 PM	No	40.02091	-74.8654	1		
100601359	2/25/2010	1:09:10 PM	No	40.88527	-73.98032	1		
100601387	2/26/2010	2:58:20 PM	No	38.93157	-74.92051	1		
100600238	2/27/2010	8:52:06 AM	No	38.93773	-74.93708	1		
100600183	2/28/2010	8:33:10 AM	No	38.94127	-74.93143	1		
100600231	3/1/2010	8:49:42 AM	No	38.94166	-74.93129	1		
100601819	3/2/2010	5:53:21 PM	No	38.94166	-74.93129	1		
100601820	3/3/2010	5:54:37 PM	No	38.94166	-74.93129	1		
100600137	3/4/2010	8:17:03 AM	No	38.94181	-74.92726	1		
100601812	3/5/2010	5:48:02 PM	No	38.94194	-74.92785	1		
100600486	3/6/2010	10:08:08 AM	No	38.94502	-74.90618	1		
100600222	3/7/2010	8:44:10 AM	No	38.94639	-74.90625	1		
100601797	3/8/2010	5:28:36 PM	No	38.96704	-74.84683	1		
100600914	3/9/2010	12:30:32 PM	No	38.9757	-74.96066	1		
100600910	3/10/2010	12:21:43 PM	No	38.98774	-74.94611	1		
100601136	3/11/2010	1:44:09 PM	No	38.98922	-74.94108	1		
100600891	3/12/2010	12:11:11 PM	No	38.98987	-74.83351	1		
100600449	3/13/2010	10:00:11 AM	No	38.99281	-74.95123	1		
100601688	3/14/2010	4:27:19 PM	No	38.99281	-74.95123	1		
100600387	3/15/2010	9:56:36 AM	No	38.993	-74.95318	1		
100601817	3/16/2010	5:51:30 PM	No	38.993	-74.95318	1		
100600204	3/17/2010	8:42:04 AM	No	39.00656	-74.94845	1		
100600200	3/18/2010	8:36:59 AM	No	39.01316	-74.9474	1		
100601361	3/19/2010	2:52:02 PM	No	39.02353	-74.93222	1		
100600861	3/20/2010	12:02:05 PM	No	39.03088	-74.85161	1		
100600601	3/21/2010	10:41:53 AM	No	39.03242	-74.9252	1		
100601237	3/22/2010	2:13:15 PM	No	39.03276	-74.85614	1		
100600442	3/23/2010	9:59:09 AM	No	39.03328	-74.85768	1		
100601336	3/24/2010	2:43:08 PM	No	39.03329	-74.85771	1		
100600459	3/25/2010	10:02:27 AM	No	39.0334	-74.85678	1		
100600257	3/26/2010	8:56:06 AM	No	39.03927	-74.8978	1		
100601426	3/27/2010	3:13:00 PM	No	39.04051	-74.87991	1		
100600153	3/28/2010	8:23:27 AM	No	39.0409	-74.89861	1		
100601226	3/29/2010	2:13:20 PM	No	39.06415	-74.74806	1		
100601265	3/30/2010	2:15:10 PM	No	39.07425	-74.82402	1		
100600229	3/31/2010	8:50:19 AM	Yes	39.08435	-74.86985	1		
100600227	4/1/2010	8:47:02 AM	No	39.15683	-74.70051	1		
100601815	4/2/2010	5:50:11 PM	No	39.16537	-74.72015	1		
100600244	4/3/2010	8:54:23 AM	No	39.21486	-74.69411	1		
100600211	4/4/2010	8:40:37 AM	No	39.21669	-74.69918	1		
100601337	4/5/2010	2:43:40 PM	No	39.24036	-74.66607	1		
100601521	4/6/2010	3:35:29 PM	yes	39.24765	-74.73399	1		

Figure 57 YEARS (2010,2011, 2012)

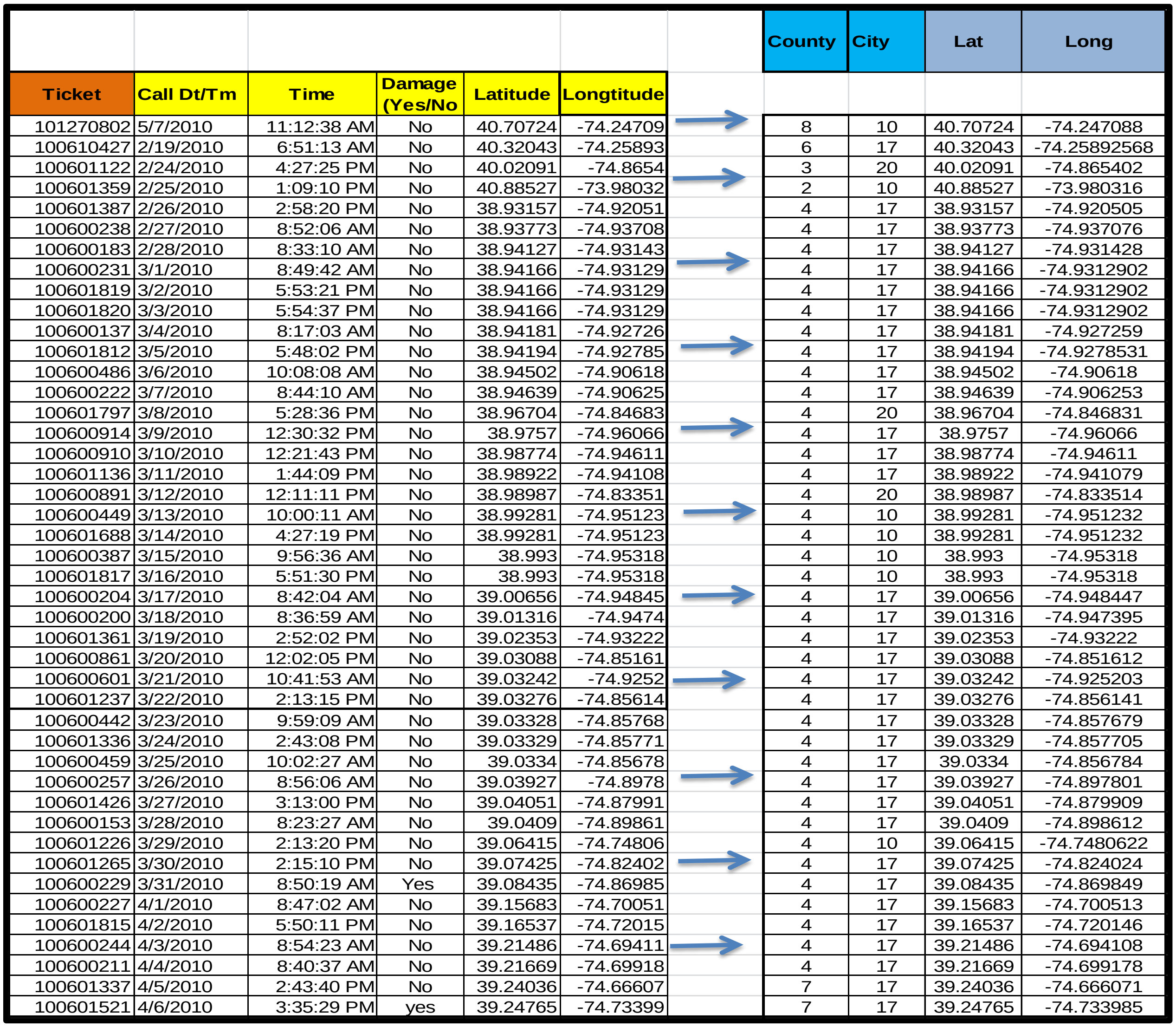
5- Season (Spring, Summer, Fall, Winter), the seasons was selected as attribute and categorized into 4 categories: so the function will look where is the ticket falling in and place the ticket in the right season which happened in. See figure 58.



						Season				
						Season	Spring	Summer	Fall	Winter
Ticket	Call Dt/Tm	Time	Damage (Yes/No	Latitude	Longtitude	0	1	1	2	3
101270802	5/7/2010	11:12:38 AM	No	40.70724	-74.24709	Spring	1	0	0	0
100610427	2/19/2010	6:51:13 AM	No	40.32043	-74.25893	Winter	0	0	0	1
100601122	2/24/2010	4:27:25 PM	No	40.02091	-74.8654	Winter	0	0	0	1
100601359	2/25/2010	1:09:10 PM	No	40.88527	-73.98032	Winter	0	0	0	1
100601387	2/26/2010	2:58:20 PM	No	38.93157	-74.92051	Winter	0	0	0	1
100600238	2/27/2010	8:52:06 AM	No	38.93773	-74.93708	Winter	0	0	0	1
100600183	2/28/2010	8:33:10 AM	No	38.94127	-74.93143	Winter	0	0	0	1
100600231	3/1/2010	8:49:42 AM	No	38.94166	-74.93129	Spring	1	0	0	0
100601819	3/2/2010	5:53:21 PM	No	38.94166	-74.93129	Spring	1	0	0	0
100601820	3/3/2010	5:54:37 PM	No	38.94166	-74.93129	Spring	1	0	0	0
100600137	3/4/2010	8:17:03 AM	No	38.94181	-74.92726	Spring	1	0	0	0
100601812	3/5/2010	5:48:02 PM	No	38.94194	-74.92785	Spring	1	0	0	0
100600486	3/6/2010	10:08:08 AM	No	38.94502	-74.90618	Spring	1	0	0	0
100600222	3/7/2010	8:44:10 AM	No	38.94639	-74.90625	Spring	1	0	0	0
100601797	3/8/2010	5:28:36 PM	No	38.96704	-74.84683	Spring	1	0	0	0
100600914	3/9/2010	12:30:32 PM	No	38.9757	-74.96066	Spring	1	0	0	0
100600910	3/10/2010	12:21:43 PM	No	38.98774	-74.94611	Spring	1	0	0	0
100601136	3/11/2010	1:44:09 PM	No	38.98922	-74.94108	Spring	1	0	0	0
100600891	3/12/2010	12:11:11 PM	No	38.98987	-74.83351	Spring	1	0	0	0
100600449	3/13/2010	10:00:11 AM	No	38.99281	-74.95123	Spring	1	0	0	0
100601688	3/14/2010	4:27:19 PM	No	38.99281	-74.95123	Spring	1	0	0	0
100600387	3/15/2010	9:56:36 AM	No	38.993	-74.95318	Spring	1	0	0	0
100601817	3/16/2010	5:51:30 PM	No	38.993	-74.95318	Spring	1	0	0	0
100600204	3/17/2010	8:42:04 AM	No	39.00656	-74.94845	Spring	1	0	0	0
100600200	3/18/2010	8:36:59 AM	No	39.01316	-74.9474	Spring	1	0	0	0
100601361	3/19/2010	2:52:02 PM	No	39.02353	-74.93222	Spring	1	0	0	0
100600861	3/20/2010	12:02:05 PM	No	39.03088	-74.85161	Spring	1	0	0	0
100600601	3/21/2010	10:41:53 AM	No	39.03242	-74.9252	Spring	1	0	0	0
100601237	3/22/2010	2:13:15 PM	No	39.03276	-74.85614	Spring	1	0	0	0
100600442	3/23/2010	9:59:09 AM	No	39.03328	-74.85768	Spring	1	0	0	0
100601336	3/24/2010	2:43:08 PM	No	39.03329	-74.85771	Spring	1	0	0	0
100600459	3/25/2010	10:02:27 AM	No	39.0334	-74.85678	Spring	1	0	0	0
100600257	3/26/2010	8:56:06 AM	No	39.03927	-74.8978	Spring	1	0	0	0
100601426	3/27/2010	3:13:00 PM	No	39.04051	-74.87991	Spring	1	0	0	0
100600153	3/28/2010	8:23:27 AM	No	39.0409	-74.89861	Spring	1	0	0	0
100601226	3/29/2010	2:13:20 PM	No	39.06415	-74.74806	Spring	1	0	0	0
100601265	3/30/2010	2:15:10 PM	No	39.07425	-74.82402	Spring	1	0	0	0
100600229	3/31/2010	8:50:19 AM	Yes	39.08435	-74.86985	Spring	1	0	0	0
100600227	4/1/2010	8:47:02 AM	No	39.15683	-74.70051	Spring	1	0	0	0
100601815	4/2/2010	5:50:11 PM	No	39.16537	-74.72015	Spring	1	0	0	0
100600244	4/3/2010	8:54:23 AM	No	39.21486	-74.69411	Spring	1	0	0	0
100600211	4/4/2010	8:40:37 AM	No	39.21669	-74.69918	Spring	1	0	0	0
100601337	4/5/2010	2:43:40 PM	No	39.24036	-74.66607	Spring	1	0	0	0
100601521	4/6/2010	3:35:29 PM	yes	39.24765	-74.73399	Spring	1	0	0	0

Figure 58 Season (Spring, Summer, Fall, Winter)

6- Location (County, City, Latitude , Longitude), the Location was selected as attribute and categorized into 4 categories: so the function will look where is the ticket falling in and place the ticket in the right location within the right Lat,& Long which happened in. See figure 59.



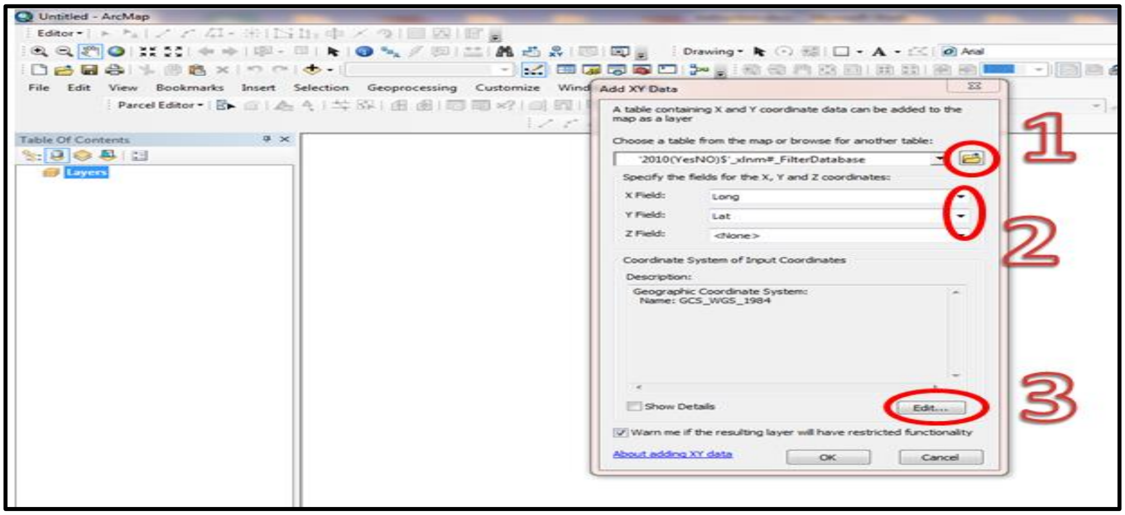
						County	City	Lat	Long
Ticket	Call Dt/Tm	Time	Damage (Yes/No	Latitude	Longtitude				
101270802	5/7/2010	11:12:38 AM	No	40.70724	-74.24709	8	10	40.70724	-74.247088
100610427	2/19/2010	6:51:13 AM	No	40.32043	-74.25893	6	17	40.32043	-74.25892568
100601122	2/24/2010	4:27:25 PM	No	40.02091	-74.8654	3	20	40.02091	-74.865402
100601359	2/25/2010	1:09:10 PM	No	40.88527	-73.98032	2	10	40.88527	-73.980316
100601387	2/26/2010	2:58:20 PM	No	38.93157	-74.92051	4	17	38.93157	-74.920505
100600238	2/27/2010	8:52:06 AM	No	38.93773	-74.93708	4	17	38.93773	-74.937076
100600183	2/28/2010	8:33:10 AM	No	38.94127	-74.93143	4	17	38.94127	-74.931428
100600231	3/1/2010	8:49:42 AM	No	38.94166	-74.93129	4	17	38.94166	-74.9312902
100601819	3/2/2010	5:53:21 PM	No	38.94166	-74.93129	4	17	38.94166	-74.9312902
100601820	3/3/2010	5:54:37 PM	No	38.94166	-74.93129	4	17	38.94166	-74.9312902
100600137	3/4/2010	8:17:03 AM	No	38.94181	-74.92726	4	17	38.94181	-74.927259
100601812	3/5/2010	5:48:02 PM	No	38.94194	-74.92785	4	17	38.94194	-74.9278531
100600486	3/6/2010	10:08:08 AM	No	38.94502	-74.90618	4	17	38.94502	-74.90618
100600222	3/7/2010	8:44:10 AM	No	38.94639	-74.90625	4	17	38.94639	-74.906253
100601797	3/8/2010	5:28:36 PM	No	38.96704	-74.84683	4	20	38.96704	-74.846831
100600914	3/9/2010	12:30:32 PM	No	38.9757	-74.96066	4	17	38.9757	-74.96066
100600910	3/10/2010	12:21:43 PM	No	38.98774	-74.94611	4	17	38.98774	-74.94611
100601136	3/11/2010	1:44:09 PM	No	38.98922	-74.94108	4	17	38.98922	-74.941079
100600891	3/12/2010	12:11:11 PM	No	38.98987	-74.83351	4	20	38.98987	-74.833514
100600449	3/13/2010	10:00:11 AM	No	38.99281	-74.95123	4	10	38.99281	-74.951232
100601688	3/14/2010	4:27:19 PM	No	38.99281	-74.95123	4	10	38.99281	-74.951232
100600387	3/15/2010	9:56:36 AM	No	38.993	-74.95318	4	10	38.993	-74.95318
100601817	3/16/2010	5:51:30 PM	No	38.993	-74.95318	4	10	38.993	-74.95318
100600204	3/17/2010	8:42:04 AM	No	39.00656	-74.94845	4	17	39.00656	-74.948447
100600200	3/18/2010	8:36:59 AM	No	39.01316	-74.9474	4	17	39.01316	-74.947395
100601361	3/19/2010	2:52:02 PM	No	39.02353	-74.93222	4	17	39.02353	-74.93222
100600861	3/20/2010	12:02:05 PM	No	39.03088	-74.85161	4	17	39.03088	-74.851612
100600601	3/21/2010	10:41:53 AM	No	39.03242	-74.9252	4	17	39.03242	-74.925203
100601237	3/22/2010	2:13:15 PM	No	39.03276	-74.85614	4	17	39.03276	-74.856141
100600442	3/23/2010	9:59:09 AM	No	39.03328	-74.85768	4	17	39.03328	-74.857679
100601336	3/24/2010	2:43:08 PM	No	39.03329	-74.85771	4	17	39.03329	-74.857705
100600459	3/25/2010	10:02:27 AM	No	39.0334	-74.85678	4	17	39.0334	-74.856784
100600257	3/26/2010	8:56:06 AM	No	39.03927	-74.8978	4	17	39.03927	-74.897801
100601426	3/27/2010	3:13:00 PM	No	39.04051	-74.87991	4	17	39.04051	-74.879909
100600153	3/28/2010	8:23:27 AM	No	39.0409	-74.89861	4	17	39.0409	-74.898612
100601226	3/29/2010	2:13:20 PM	No	39.06415	-74.74806	4	10	39.06415	-74.7480622
100601265	3/30/2010	2:15:10 PM	No	39.07425	-74.82402	4	17	39.07425	-74.824024
100600229	3/31/2010	8:50:19 AM	Yes	39.08435	-74.86985	4	17	39.08435	-74.869849
100600227	4/1/2010	8:47:02 AM	No	39.15683	-74.70051	4	17	39.15683	-74.700513
100601815	4/2/2010	5:50:11 PM	No	39.16537	-74.72015	4	17	39.16537	-74.720146
100600244	4/3/2010	8:54:23 AM	No	39.21486	-74.69411	4	17	39.21486	-74.694108
100600211	4/4/2010	8:40:37 AM	No	39.21669	-74.69918	4	17	39.21669	-74.699178
100601337	4/5/2010	2:43:40 PM	No	39.24036	-74.66607	7	17	39.24036	-74.666071
100601521	4/6/2010	3:35:29 PM	yes	39.24765	-74.73399	7	17	39.24765	-74.733985

Figure 59 : Location (County, City, Lat, Long)

7- Damages within 10 miles diameter, 20 miles diameter, 30 miles diameter (Diameter 10, Diameter 20, Diameter 30), the number of damages within different diameters for the same incident was selected as attribute and categorized into 3 categories. Joint Spatial method was used to calculate these values.

Joint spatial was introduced to transfer attributes from one layer to another based on their spatial relationship. Joint Spatial is a process used in this study to transfer data from one feature layer's attribute to combined it with another layer's attributes. In addition, by developing this attribute, we should be able to see how surrounding damages within different density and concentration could influence the damage of future underground gas pipeline. The following steps show how D10,D20, and D30 developed.

Steps 1. First step is to create XY events from excel file, then browse for the excel file, set latitude and longitude fields and set coordinate system to WGS 84, then OK. See figure 60



Add XY Data

A table containing X and Y coordinate data can be added to the map as a layer

Choose a table from the map or browse for another table:
'2010(YesNO)$'_xlnm#_FilterDatabase

Specify the fields for the X, Y and Z coordinates:

X Field: Long

Y Field: Lat

Z Field: <None>

Coordinate System of Input Coordinates

Description:
Geographic Coordinate System:
Name: GCS_WGS_1984

□ Show Details   Edit...

☑ Warn me if the resulting layer will have restricted functionality

About adding XY data   OK   Cancel

Figure 60 Creating X,Y fileds in ARC GIS

Step 2. Plot the data on the map within provided Lat, and Long. There are some data with bad coordinate: see figure 61. This process is done as a part of a data cleaning: which is very important to get accurate and precise results. The locations which have bad coordinates needed to be removed from the data.

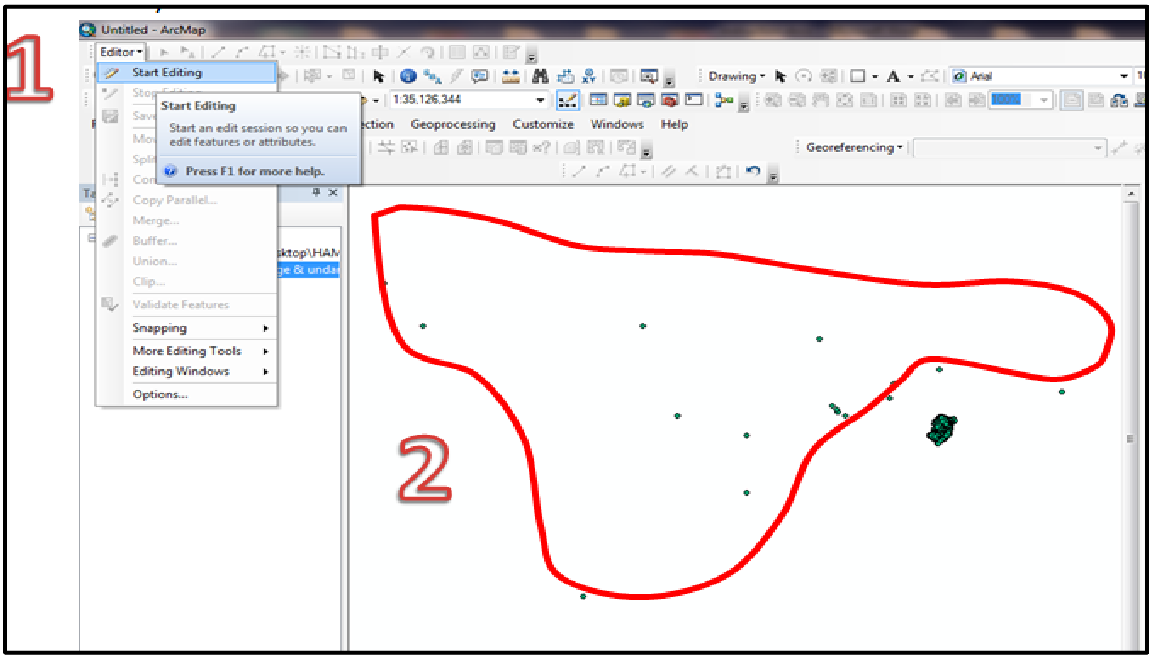


Figure 61Locating Bad Coordinates within the plotted Locations

Step 3. Create buffer areas of specific radius (10miles, 20 miles, 30 miles): go to search tab (CTRL+F) and type “Buffer” Select Buffer (Analysis) tool and then we set the required parameters (Input Features, Output Features, Linear units set to miles and type10). Do the same thing for 20 and 30 miles separately. See figure 62.

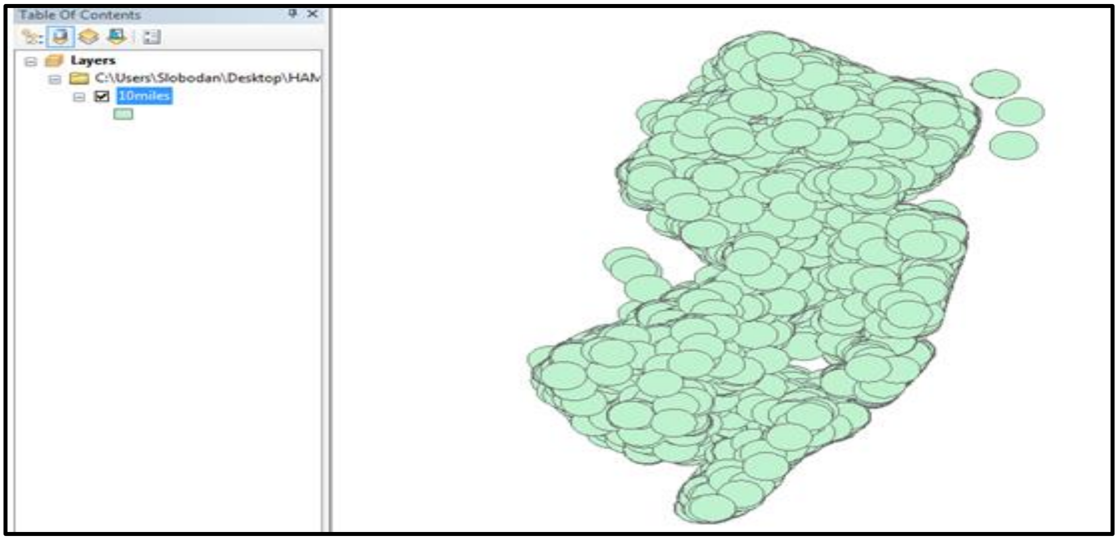
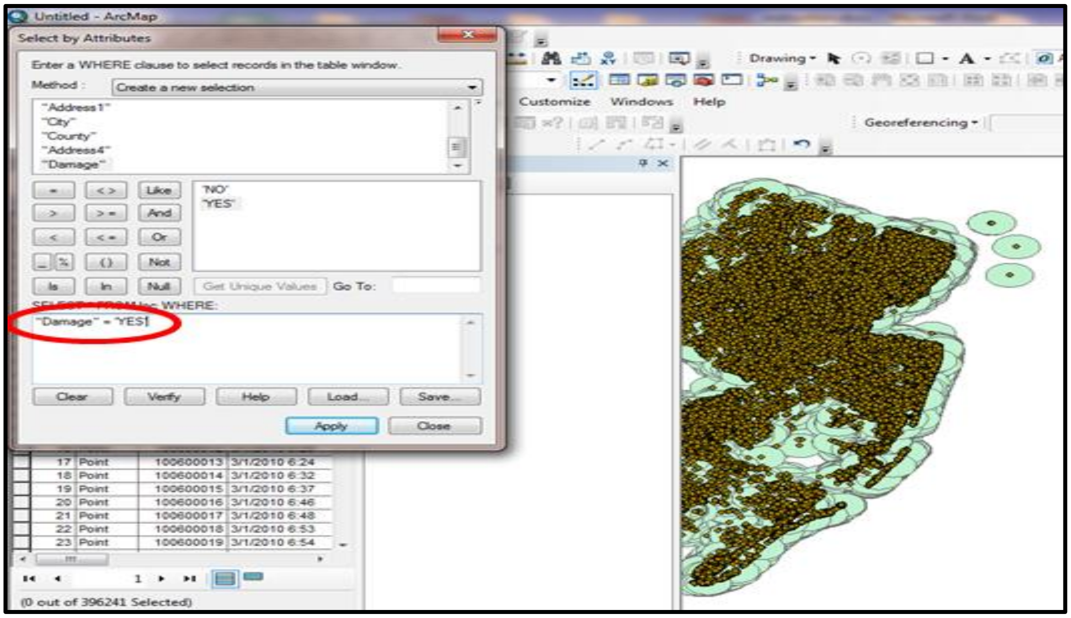


Figure 62: Buffer Circle 10 Miles Diameter

Step 4. Is this step we set the parameters needed to perform the task and select attributes, such as damaged (Yes), city, ticket, county, and merge between two layers. A new layer that contain Join_Count field showing the number of locations inside 10 miles diameter area was created and joint spatial was conducted. See figure 63, 64.



Select by Attributes

Enter a WHERE clause to select records in the table window.

Method : Create a new selection

"Address1"
"City"
"County"
"Address4"
"Damage"

'NO'
'YES'

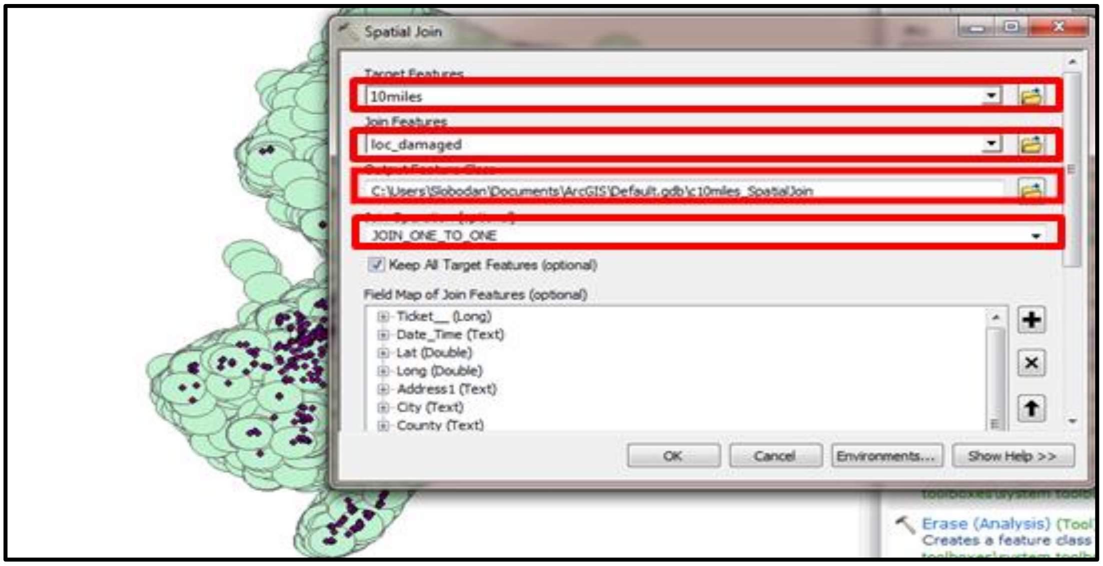
SELECT * FROM loc WHERE:

"Damage" = 'YES'

17	Point	100600013	3/1/2010 6:24
18	Point	100600014	3/1/2010 6:32
19	Point	100600015	3/1/2010 6:37
20	Point	100600016	3/1/2010 6:46
21	Point	100600017	3/1/2010 6:48
22	Point	100600018	3/1/2010 6:53
23	Point	100600019	3/1/2010 6:54

(0 out of 396241 Selected)

Figure 63 Joint Spatial Buffers within 10 miles Diameter


Spatial Join

Target Features: 10miles

Join Features: loc_damaged

Output Feature Class: C:\Users\Slobodan\Documents\ArcGIS\Default.gdb\c10miles_SpatialJoin

Join Operation (optional): JOIN_ONE_TO_ONE

☑ Keep All Target Features (optional)

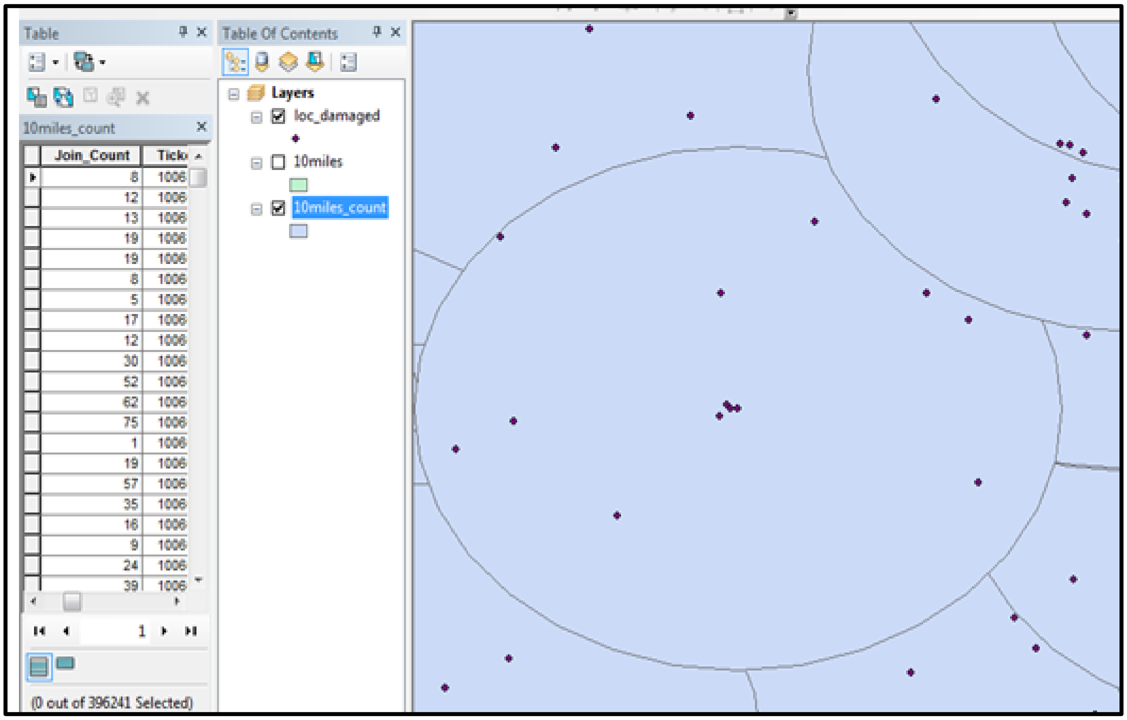
Field Map of Join Features (optional)

- Ticket__ (Long)
- Date_Time (Text)
- Lat (Double)
- Long (Double)
- Address1 (Text)
- City (Text)
- County (Text)

Figure 64 forming Joint Spatial

Step 5. Is this step we double check the selected parameters and cross reference some of the damage numbers within certain ticket number. Finally, the results look like ellipses because of projection: but actually these are circles (Figure 65).

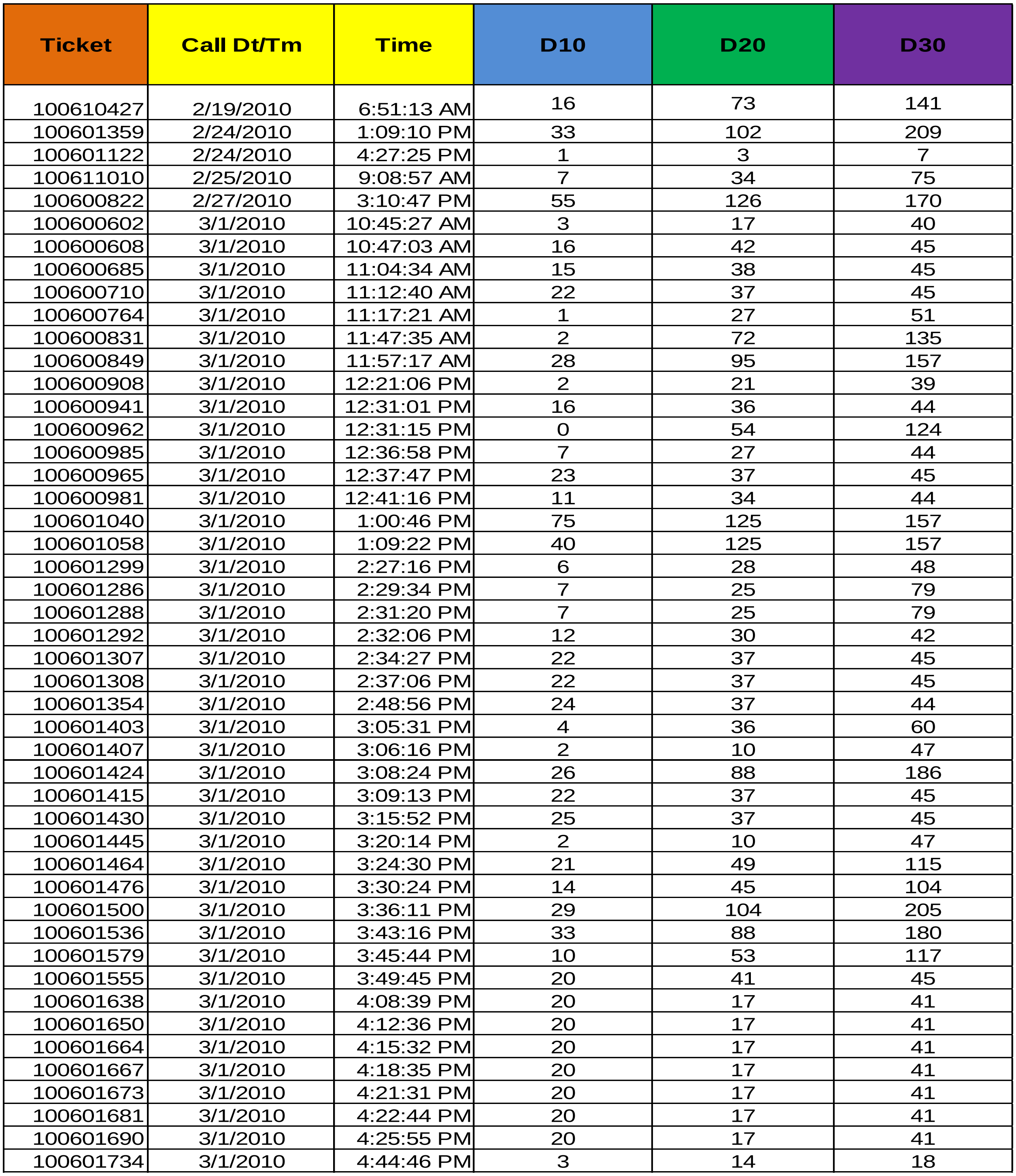
The results are ready to be exported to excel. All five steps are repeated to determine the damages within 20 miles, 30 miles. See figure 66 for final outcome.



Join_Count	Ticke
8	1006
12	1006
13	1006
19	1006
19	1006
8	1006
5	1006
17	1006
12	1006
30	1006
52	1006
62	1006
75	1006
1	1006
19	1006
57	1006
35	1006
16	1006
9	1006
24	1006
39	1006

(0 out of 396241 Selected)

Figure 65 Magnified10 miles Circle.



Ticket	Call Dt/Tm	Time	D10	D20	D30
100610427	2/19/2010	6:51:13 AM	16	73	141
100601359	2/24/2010	1:09:10 PM	33	102	209
100601122	2/24/2010	4:27:25 PM	1	3	7
100611010	2/25/2010	9:08:57 AM	7	34	75
100600822	2/27/2010	3:10:47 PM	55	126	170
100600602	3/1/2010	10:45:27 AM	3	17	40
100600608	3/1/2010	10:47:03 AM	16	42	45
100600685	3/1/2010	11:04:34 AM	15	38	45
100600710	3/1/2010	11:12:40 AM	22	37	45
100600764	3/1/2010	11:17:21 AM	1	27	51
100600831	3/1/2010	11:47:35 AM	2	72	135
100600849	3/1/2010	11:57:17 AM	28	95	157
100600908	3/1/2010	12:21:06 PM	2	21	39
100600941	3/1/2010	12:31:01 PM	16	36	44
100600962	3/1/2010	12:31:15 PM	0	54	124
100600985	3/1/2010	12:36:58 PM	7	27	44
100600965	3/1/2010	12:37:47 PM	23	37	45
100600981	3/1/2010	12:41:16 PM	11	34	44
100601040	3/1/2010	1:00:46 PM	75	125	157
100601058	3/1/2010	1:09:22 PM	40	125	157
100601299	3/1/2010	2:27:16 PM	6	28	48
100601286	3/1/2010	2:29:34 PM	7	25	79
100601288	3/1/2010	2:31:20 PM	7	25	79
100601292	3/1/2010	2:32:06 PM	12	30	42
100601307	3/1/2010	2:34:27 PM	22	37	45
100601308	3/1/2010	2:37:06 PM	22	37	45
100601354	3/1/2010	2:48:56 PM	24	37	44
100601403	3/1/2010	3:05:31 PM	4	36	60
100601407	3/1/2010	3:06:16 PM	2	10	47
100601424	3/1/2010	3:08:24 PM	26	88	186
100601415	3/1/2010	3:09:13 PM	22	37	45
100601430	3/1/2010	3:15:52 PM	25	37	45
100601445	3/1/2010	3:20:14 PM	2	10	47
100601464	3/1/2010	3:24:30 PM	21	49	115
100601476	3/1/2010	3:30:24 PM	14	45	104
100601500	3/1/2010	3:36:11 PM	29	104	205
100601536	3/1/2010	3:43:16 PM	33	88	180
100601579	3/1/2010	3:45:44 PM	10	53	117
100601555	3/1/2010	3:49:45 PM	20	41	45
100601638	3/1/2010	4:08:39 PM	20	17	41
100601650	3/1/2010	4:12:36 PM	20	17	41
100601664	3/1/2010	4:15:32 PM	20	17	41
100601667	3/1/2010	4:18:35 PM	20	17	41
100601673	3/1/2010	4:21:31 PM	20	17	41
100601681	3/1/2010	4:22:44 PM	20	17	41
100601690	3/1/2010	4:25:55 PM	20	17	41
100601734	3/1/2010	4:44:46 PM	3	14	18

Figure 66 Shows all damages within 10,20, 30 mile Diameter

1.30 Choosing Machine Learning Algorithms

Logistic Regression analysis is a form of predictive modeling technique which investigates the relationship between a dependent (Target) and independent (Predictor) variable. In this study damage condition is the dependent variable and independent variable (s) are data attributes, Time, Days, weeks, Months, Years, Cities, Counties, D10, D20, D30. Logistic Regression technique is used for forecasting, time series modeling and finding the causal effect relationship between the variables. For example, relationship between damage and undamaged UG gas pipes will be determined through the attributes of the predictive model. More specifically, the regression analysis will go through each point of attribute (predictor) and try to fit most of the points through patterns. In addition, regression analysis is an important tool for modeling and analyzing data. In our study, regression analysis fit a curve / line to the data points, in such a manner that the differences between the distances of data points from the curve or line is minimized.

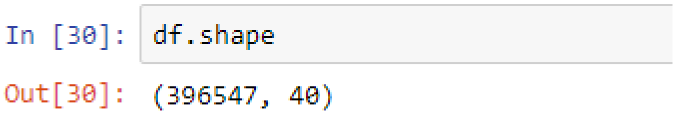
Why we employed Regression Analysis in our study Model? First, Regression Analysis indicate the significant relationships between dependent variable (Target (YES/NO)) and (Predictors (data Attributes)) independent variable. Second, Regression Analyses indicate the strength of impact of multiple independent variables (data attributes) on a dependent variable (Damage/ Not).k-nearest neighbors algorithm (K-NN) is a simple algorithm that store all available cases (Predictors) and classifies new cases based on a similarity measure (e.g., distance functions).

KNN analysis was chosen because it works as pattern recognition though specific distance for each neighborhood, then it classifies the data based ion the distances between the generated neighborhoods.

Random forest builds multiple decision trees for the data attributes Time, Week days, Months, and other attributes and merges them together to get a more accurate and stable prediction. More specifically, it creates a forest of data attributes and makes random selection. In our study model we chose Random Forest Model, because it can be used for both regression and classification tasks and that it's easy to view the relative importance it assigns to the input Attribute. A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyper plane. In addition, given labeled training data (supervised learning), the algorithm outputs an optimal hyper plane which categorizes all the damaged, and undamaged UG Gas pipe. Advantages of S.V Machine is work really well with clear margin of separation: for example, we have damaged, and undamaged UG Gas pipe data.

1.31 Predictive Model Methodology

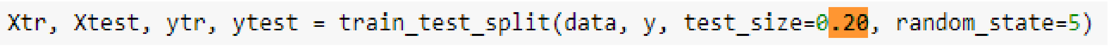
Now as we understand the machine-learning problem want to solve for: predicting gas pipe damage, and dominant risk factors for future UG gas pipe operations. The next step is to build a model which is to employ data science methodologies like Logistic regression, Random Forest, KNN, Bayesian. Looking at the historical data we have, we want to produce a model that estimates a particular variable specific. Which is (YES/NO) damages or undamaged. The following steps explain the preliminary steps to input data into Python (Anaconda). The total number of records used in the model were 396,547 (including undamaged & damages): see figure 67.



```
In [30]: df.shape
Out[30]: (396547, 40)
```

Figure 67: Shows total number of data (Anaconada)

The total inputted data attributes into the model were equal 40 attributes including (Time, am, pm, days, Mon, Tue, Wed, Thurs, Fri, Sat, Sun, Month, Jan, Feb, March, May, Jun, Jul, Aug, Sept, Oct, Nov, Dec, Year 2010-2014, Season, Winter, Summer, Autumn, Spring, damages within 10 miles, 20 miles, and 30 miles. Then it was developed in the model into 752 attributes, cities were put in the columns and values (1, or 0) was assigned based on the ticket, damaged or not. (Zero) value replaced the undamaged tickets, and (1) value replaced the damaged tickets. This step was performed to transfer the data into numerical which make it useful by the algorithm to process it, instead of having text data. Furthermore, data cleaning was performed in Excel, and Python (Anaconda) to make sure the data was not having any missing values, repeated values, or corrupted numbers. Next, the data split was 80% Training, and 20% Testing see figure 68.



```
Xtr, Xtest, ytr, ytest = train_test_split(data, y, test_size=0.20, random_state=5)
```

Figure 68: Shows Data Split between Training, and Testing

Which means; Data in Training = 0.8 *396,547 = 317,237

Data in Testing = 0.2* 396,547= 79,309

Micro analysis was selected from (Micro, Macro, and Binary). The following chart explain the methodology of the Model see figure 69. The process starts by preparing the data, then inputting data into Anaconda, then cleaning the data again, then splitting the data, then running the algorithms in 80 % training. Then we run remaining 20% testing data into testing and test the model. The following step is to select a model based on the testing metrics such as confusion matrix, precession, recall. then run the confusion matrix, precession, recall.

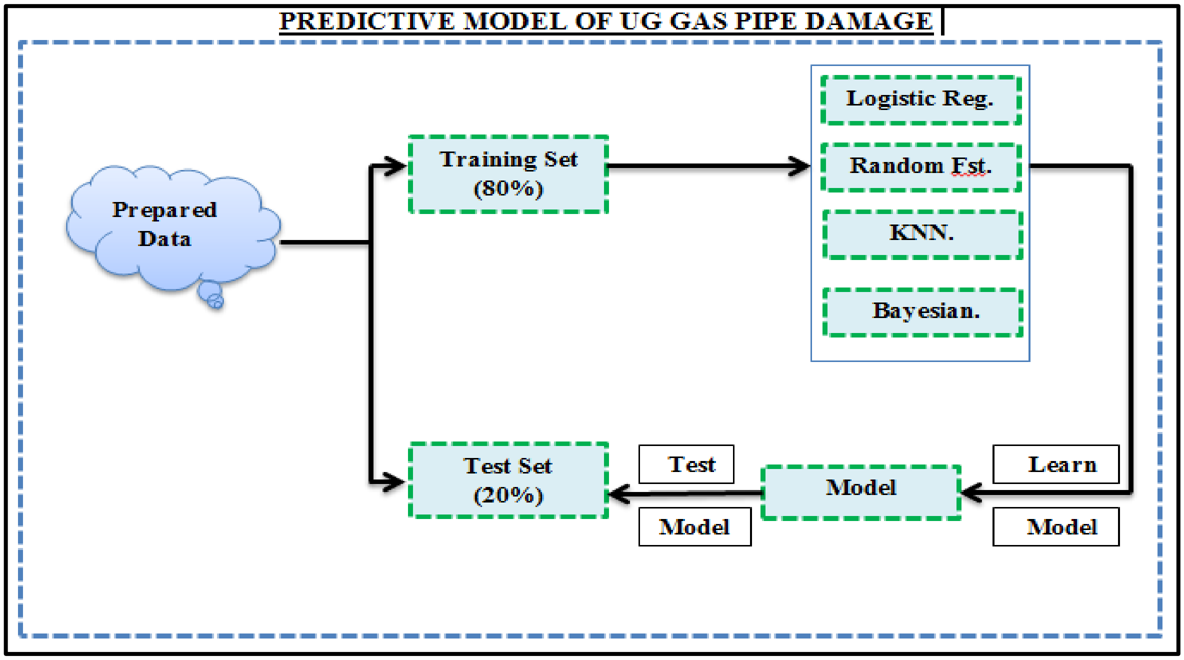
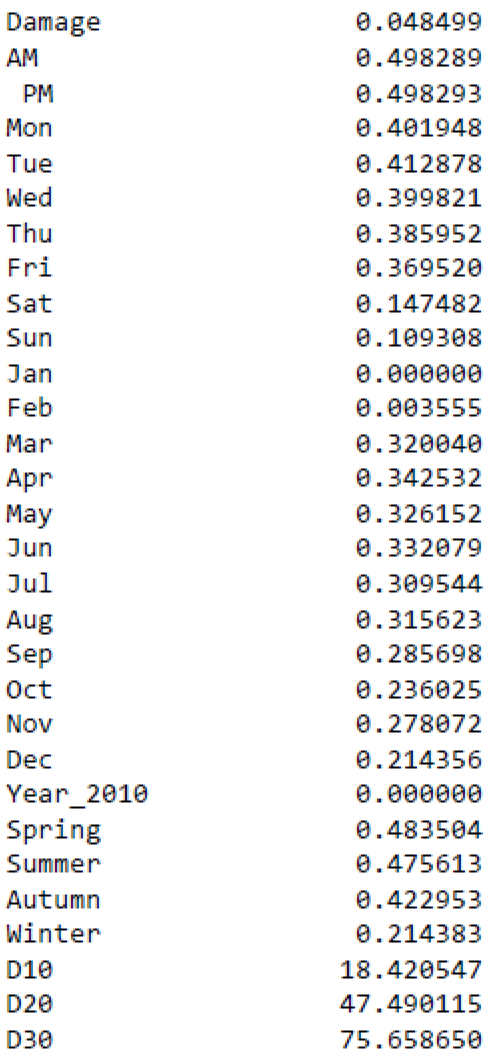


Figure 69: Predictive Model for Underground Gas Pipe Damages.

1.32 Processing the Predictive Model (Training The Model)

The process of training a predictive model involves providing PM algorithm (that is, the learning algorithm) with training data to learn from, and develop patterns . The term PM model in our study refers to the model artifact that is created by the training process. The provided training data to the algorithm contains the correct answers, which is known as a target or target attribute. The learning algorithm finds patterns in the training data that map the inputted data attributes to the target (the answer that we want to predict), and it outputs an PM model that captures these patterns. Then, the model will be used to get predictions on new data for which the Target Answer is not known to the model. we do not know the target. In our study we provided the algorithm with 40 attributes called predictors, and the Target which is (YES = 1) or (No = 0), damage or undamaged.

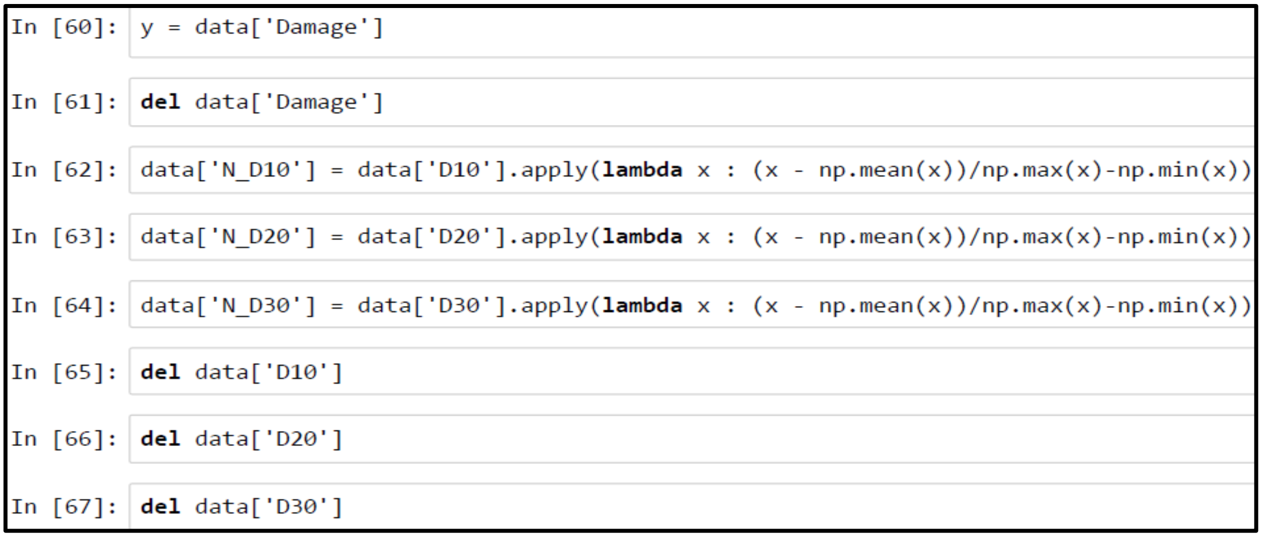
This step is viewing the attributes type, and checking the attributes' normal distribution, for the predictors to give good results, all of the data attributes needed to be normally distributed. Thus, standard deviation test was performed on some of attribute of the data. The results were some of the attributes were not normally distributed, Anaconda was used for that test see figure 72.



```
Damage                0.048499
AM                    0.498289
 PM                   0.498293
Mon                   0.401948
Tue                   0.412878
Wed                   0.399821
Thu                   0.385952
Fri                   0.369520
Sat                   0.147482
Sun                   0.109308
Jan                   0.000000
Feb                   0.003555
Mar                   0.320040
Apr                   0.342532
May                   0.326152
Jun                   0.332079
Jul                   0.309544
Aug                   0.315623
Sep                   0.285698
Oct                   0.236025
Nov                   0.278072
Dec                   0.214356
Year_2010             0.000000
Spring                0.483504
Summer                0.475613
Autumn                0.422953
Winter                0.214383
D10                  18.420547
D20                  47.490115
D30                  75.658650
```

Figure 70 figure shows the standard deviation of all attributes

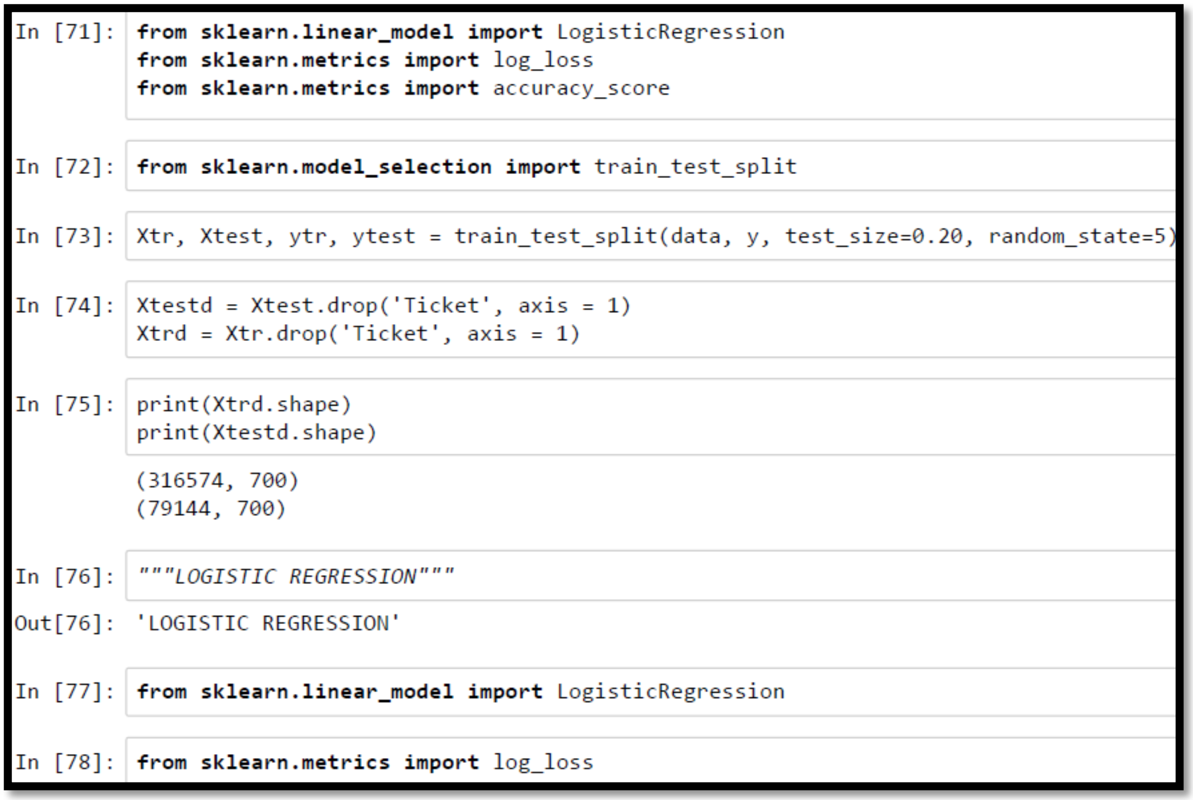
As can be seen in figure 70, damages within 10 miles diameter, 20 miles diameter, and 30 miles diameter have the values 18.420547, 47.490115, and 75.658650. These values will have negative impact on the modal. Thus, all these three values were normalized for all data set. Standard deviation equation was used to normalize these values, see figures 71.



```
In [60]: y = data['Damage']

In [61]: del data['Damage']

In [62]: data['N_D10'] = data['D10'].apply(lambda x : (x - np.mean(x))/np.max(x)-np.min(x))

In [63]: data['N_D20'] = data['D20'].apply(lambda x : (x - np.mean(x))/np.max(x)-np.min(x))

In [64]: data['N_D30'] = data['D30'].apply(lambda x : (x - np.mean(x))/np.max(x)-np.min(x))

In [65]: del data['D10']

In [66]: del data['D20']

In [67]: del data['D30']
```

Figure 71 standard Deviation for D10, D20, D30

All Libraries, Methodologies, and needed tools were imported through Python Anaconda Library Numpy was imported to the model because Numpy will speed up the data workflow, and interface with other packages in the Python system, like scikit-learn, that use Numpy. In addition, Numpy has a much more natural and convenient integration of mathematical operations. Pandas was called in to the Model, because Pandas is needed to perform some of the functions. Seaborn was called in to the model, is needed as a tool that does statistical data visualization. Seaborn is a Python visualization library based on matplotlib. It's imported because it gives a high-level interface for drawing attractive statistical graphics. Which is needed for some steps for the model. Matplotlib library was imported to provide data visualization, and needed calculations. Sklearn was imported as well because it include many tools / features such as classifications, clustering and regression. In addition, Sklearn cross validation library was imported to cross validate some of the data attributes. Logistic Regression Algorithm, the followings are specific codes were used in Logistic Regression Algorithm see figure 72. The algorithm start by importing all needed Libraries, such as Sklearn, Metrics, Numpy, and Matplotlib. Then importing all data CSV file into the model. Next, dropping and defining X, Y axes into the systems. Then, applying Ytr, Ytest, Xtrain, and Xtest to the model. Then, run the model. Next, exporting the outcome from the model as training outcome with target Yes, and No.

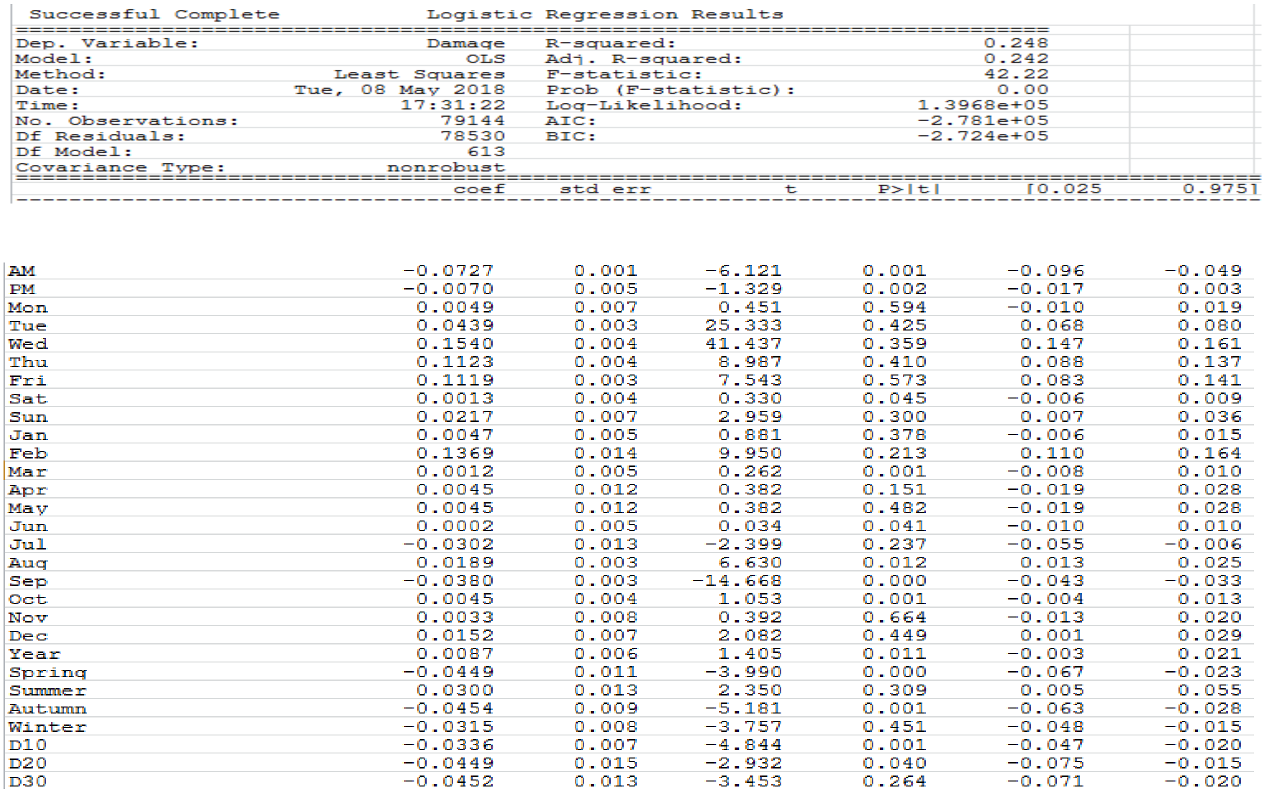


```
In [71]: from sklearn.linear_model import LogisticRegression
         from sklearn.metrics import log_loss
         from sklearn.metrics import accuracy_score

In [72]: from sklearn.model_selection import train_test_split

In [73]: Xtr, Xtest, ytr, ytest = train_test_split(data, y, test_size=0.20, random_state=5)

In [74]: Xtestd = Xtest.drop('Ticket', axis = 1)
         Xtrd = Xtr.drop('Ticket', axis = 1)

In [75]: print(Xtrd.shape)
         print(Xtestd.shape)

         (316574, 700)
         (79144, 700)

In [76]: """LOGISTIC REGRESSION"""

Out[76]: 'LOGISTIC REGRESSION'

In [77]: from sklearn.linear_model import LogisticRegression

In [78]: from sklearn.metrics import log_loss
```

Figure 72: Logistic Regression Algorithm

In more details on how the data processed in the predictive model. The data was divided to X, Y , where is X is equal to all attributes, and Y is equal to Damage / Target Variables. The data split was 80 % of the data into training, 20% into testing. Moreover, 70% of the data is in X-train, Y-train. Then, the classifier was chosen to be either Logistic regression, KNN, Random Forest, Or Bayesian. Then the equation, CLF.fit(x-train, Y-train) finds the patterns in all the attributes with respect to Y- train, and save the patterns. Results, the algorithm, found the pattern in the training data. Next, the model uses the pattern found from the training data which maps all the attributes to target value which is damage (YES/NO).

Using the pattern found above combined with another function called "PREDICT". Equation Data-Predict = CCF. Predict (X – test), Data – Predict = = (Y – Test) Actual. Finally, we score the model or test the model by comparing Data-Predict Vs Y- Test. Model could not converge with more than 29 variables. Furthermore, Statistical significance (P Value) was used to determine dominant damage attributes. The model shows not all selected attributes are effective in predicting the future UG gas pipe damage see figure 73.



```
 Successful Complete            Logistic Regression Results
==============================================================================
Dep. Variable:                 Damage   R-squared:                       0.248
Model:                            OLS   Adj. R-squared:                  0.242
Method:                 Least Squares   F-statistic:                     42.22
Date:                Tue, 08 May 2018   Prob (F-statistic):               0.00
Time:                        17:31:22   Log-Likelihood:             1.3968e+05
No. Observations:               79144   AIC:                        -2.781e+05
Df Residuals:                   78530   BIC:                        -2.724e+05
Df Model:                         613
Covariance Type:            nonrobust
=========================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------

AM                            -0.0727      0.001     -6.121      0.001     -0.096      -0.049
PM                            -0.0070      0.005     -1.329      0.002     -0.017       0.003
Mon                            0.0049      0.007      0.451      0.594     -0.010       0.019
Tue                            0.0439      0.003     25.333      0.425      0.068       0.080
Wed                            0.1540      0.004     41.437      0.359      0.147       0.161
Thu                            0.1123      0.004      8.987      0.410      0.088       0.137
Fri                            0.1119      0.003      7.543      0.573      0.083       0.141
Sat                            0.0013      0.004      0.330      0.045     -0.006       0.009
Sun                            0.0217      0.007      2.959      0.300      0.007       0.036
Jan                            0.0047      0.005      0.881      0.378     -0.006       0.015
Feb                            0.1369      0.014      9.950      0.213      0.110       0.164
Mar                            0.0012      0.005      0.262      0.001     -0.008       0.010
Apr                            0.0045      0.012      0.382      0.151     -0.019       0.028
May                            0.0045      0.012      0.382      0.482     -0.019       0.028
Jun                            0.0002      0.005      0.034      0.041     -0.010       0.010
Jul                           -0.0302      0.013     -2.399      0.237     -0.055      -0.006
Aug                            0.0189      0.003      6.630      0.012      0.013       0.025
Sep                           -0.0380      0.003    -14.668      0.000     -0.043      -0.033
Oct                            0.0045      0.004      1.053      0.001     -0.004       0.013
Nov                            0.0033      0.008      0.392      0.664     -0.013       0.020
Dec                            0.0152      0.007      2.082      0.449      0.001       0.029
Year                           0.0087      0.006      1.405      0.011     -0.003       0.021
Spring                        -0.0449      0.011     -3.990      0.000     -0.067      -0.023
Summer                         0.0300      0.013      2.350      0.309      0.005       0.055
Autumn                        -0.0454      0.009     -5.181      0.001     -0.063      -0.028
Winter                        -0.0315      0.008     -3.757      0.451     -0.048      -0.015
D10                           -0.0336      0.007     -4.844      0.001     -0.047      -0.020
D20                           -0.0449      0.015     -2.932      0.040     -0.075      -0.015
D30                           -0.0452      0.013     -3.453      0.264     -0.071      -0.020
```

Figure 73: Shows the outcome from Logistic Regression Model

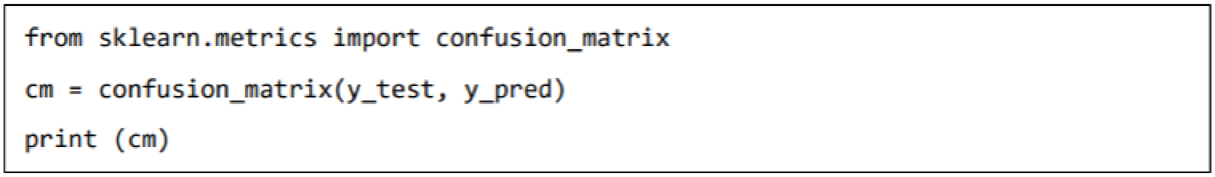
1.33 Model Testing, and Validation

20 % of the data was used in testing the predictive model, and the metrics were used to evaluate the model were Confusion Matrix, Recall, and Precision. These three Metrics were conducted on all of the outcomes of the algorithms including Logistic Regression, Random Forest, KNN, and Bayesian. Results, Logistic Regression Model gave the best result according to Confusion Matrix, Recall, precision. Sample was used for illustration purposes, see figure below. More details will be added later. In addition, comparison

between confusion matrix for KNN, Bayesian, Logistic regression, and Random Forest will be added later.

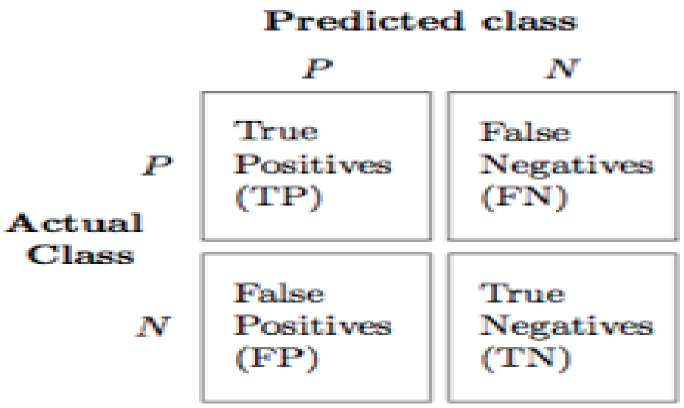
1.33.1 Confusion Matrix

A confusion matrix is a table that is we used on our study used to describe the performance of a classification models we employed in our study (Bayesian, Logistic Regression, KNN, and random Forest) on a set of test UG gas pipe data for which the true values. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing. Therefore, these the definitions regarding confusion matrix, true positives (TP): These are cases in which we predicted yes (they have the UG gas pipe damage), and they do have the damage. true negatives (TN): We predicted no, and they don't have the gas pipe damage. false positives (FP): We predicted yes, but they don't actually have the damage. False negatives (FN): We predicted no, but they actually do have the damage see figure 74, and 75.



```
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, y_pred)
print (cm)
```

Figure 74 : Show confusion Matrix Code



		Predicted class	
		P	*N*
Actual Class	*P*	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Figure 75: Show Confusion Matrix

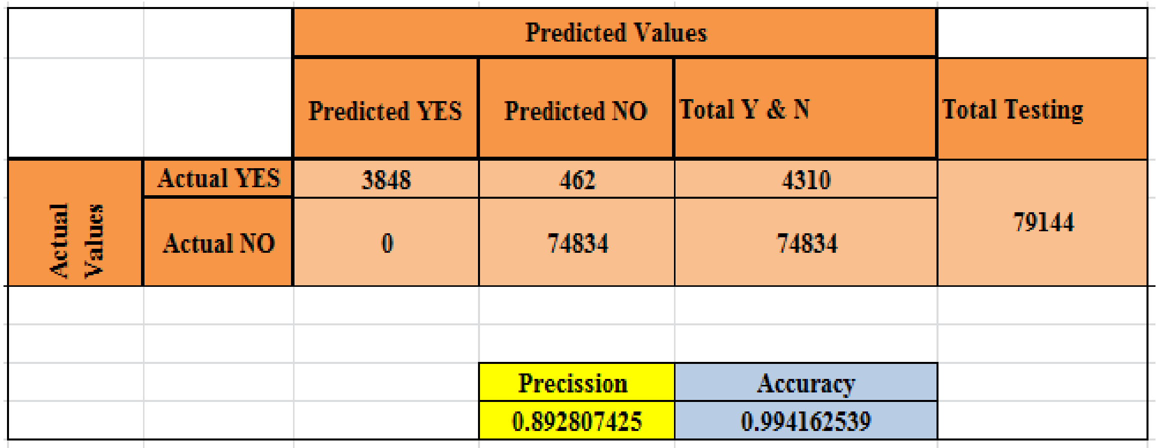
1.33.2 Accuracy

We used accuracy measure in our study so it tell us how often is the classifier correct?

Equation = (TP+TN)/total. See figure 76.

1.33.3 Precession

We used accuracy measure in our study so it tell us when it predicts yes, how often is it correct? Equation =TP/predicted yes. See figure 76

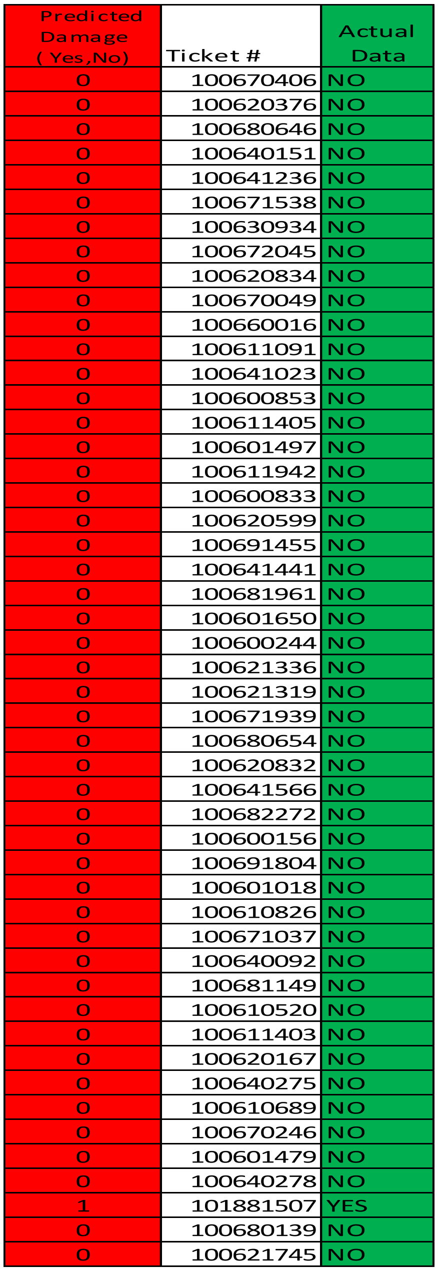


		Predicted Values			
		Predicted YES	Predicted NO	Total Y & N	Total Testing
Actual Values	Actual YES	3848	462	4310	79144
	Actual NO	0	74834	74834	
			Precission	Accuracy	
			0.892807425	0.994162539	

Figure 76:Show confusion matrix results for testing data set

The outcome from the tested model was exported into excel sheets to compare the predicted target with the actual target see table 1 below. Ticket number was used as unique value which is shared between the data in the actual form and the predicted results / target

Table 1: Show compared exported test data in excel



Predicted Damage (Yes,No)	Ticket #	Actual Data
0	100670406	NO
0	100620376	NO
0	100680646	NO
0	100640151	NO
0	100641236	NO
0	100671538	NO
0	100630934	NO
0	100672045	NO
0	100620834	NO
0	100670049	NO
0	100660016	NO
0	100611091	NO
0	100641023	NO
0	100600853	NO
0	100611405	NO
0	100601497	NO
0	100611942	NO
0	100600833	NO
0	100620599	NO
0	100691455	NO
0	100641441	NO
0	100681961	NO
0	100601650	NO
0	100600244	NO
0	100621336	NO
0	100621319	NO
0	100671939	NO
0	100680654	NO
0	100620832	NO
0	100641566	NO
0	100682272	NO
0	100600156	NO
0	100691804	NO
0	100601018	NO
0	100610826	NO
0	100671037	NO
0	100640092	NO
0	100681149	NO
0	100610520	NO
0	100611403	NO
0	100620167	NO
0	100640275	NO
0	100610689	NO
0	100670246	NO
0	100601479	NO
0	100640278	NO
1	101881507	YES
0	100680139	NO
0	100621745	NO

1.34 Discussion of Results, and Future work.

The developed predictive model encompasses a variety of statistical techniques from modeling, machine learning, data mining and others that analyze UG gas pipeline historical facts to make predictions about future events. In our study, predictive models exploit patterns found in UG gas pipe historical data and selected attributes to identify the Target variable (YES/NO) (Damage/Undamaged) . The Produced Predictive Model capture relationships among many risk factors to allow assessment of risk or potential associated with a particular set of attributes, guiding decision making for agencies dealing with UG gas pipe digging process.

Even though, there are some limitation in this study, such as missing data, and not having some valuable data attributes such as pipe size, diameter, pipe materials type …etc. The developed model from the derived data attributes was able to predict more than 80 % of the future UG gas pipe damages.

In other hand, it’s advised as part for future work to collect more data attributes related the UG gas pipeline, such as pipe size, pipe materials, reason for damage, temperature of the gas pipe, and age of the gas pipe. Analysis can be further conducted after collecting these attributes to assess the impact on the predictive model, and accuracy of future prediction. In addition, for future study, the collected attributes should be collected for both damaged, and undamaged data so the analysis can be performed.

CHAPTER SIX: BAYSIAN NETWORK

1.35 Introduction.

The proposed Bayesian network is a graphical technology for describing cause and effects relationship. Bayesian network consists of nodes, arcs, and condition probabilistic table (Yuan et al., 2015). The construction of Bayesian Network started with mapping both Bow-tie and Fault tree models to assess the preliminary influence factors involved in the underground gas (UG) line damage process. Then, the next step was to identify Bayesian nodes, Bayesian network structure, and Bayesian preliminary combined model (Figure 77). Different from bow-tie method, the Bayesian network is an inference probabilistic method, which can overcome the static limitation of bow-tie method due to its updating mechanism. It can also implement forward and backward linear predictions as well as diagnosis analysis (Bhandari et al., 2015). In addition, calculating the probability of the nodes is illustrated as combination of probability analysis, merge probability basic events, and probability per-sub notes.

The Bayesian network (BN) does not have a clear structure due to non-existing nodes. As a result, it does not demonstrate the evolution process of UG gas pipe damage. Therefore, more specific and detailed BN model needed to be constructed to study the risk involved in the UG gas pipe damage. In this research, we apply a Bayesian approach to learning Bayesian networks, containing decision-graphs generalizations of decision trees, that can encode arbitrary equality constraints to represent the conditional probability distributions in the nodes.

The proposed Bayesian network structure consists of UG gas pipe network risk model, made up of nodes that represent variables. Moreover, relationship between variables nodes can represent direct causal dependencies based on process understanding, statistical, or other types of associations. A conditional probability table (CPT) is used to describe the probability of each value of the child node, conditioned on every possible combination of values of its parent nodes. These describe the strength of the causal relationships between variables. If a variable has no parents, it is described by a marginal probability distribution. The posterior probability distribution for a variable is calculated for new observations. Bayesian network exploits the distributional simplifications of a network structure by calculating how probable events change given subsequent observations or external interventions (Korb and Nicholson, 2004). The data was collected and prepared to serve the purpose of this research. The collected data was for both damaged, and undamaged of UG gas pipe. The data was converted from raw data to excel tables by running the data through multiple excel functions. Then the data was cleaned from errors, missing.

Sequence of UG gas pipe was developed by using Bow-tie which includes two parts. The left of bow-tie is a FT that describes the latent causes for an initial UG gas pipe damage event. The right of bow-tie is an ET which describes the sequential failure of damage preventive barrier. It also presents the evolution process from initial event to final latent consequence. To overcome the complexity of the UG gas damage network, Fault Tree was used to develop the model. Fault tree analysis was used to calculate reliability of the complex UG pipe damages model. It is used to provide a logical and diagrammatic approach for evaluating the possibility of an UG gas pipe damage resulting from sequences and combinations of failure events. By using the fault tree UG gas pipe damage

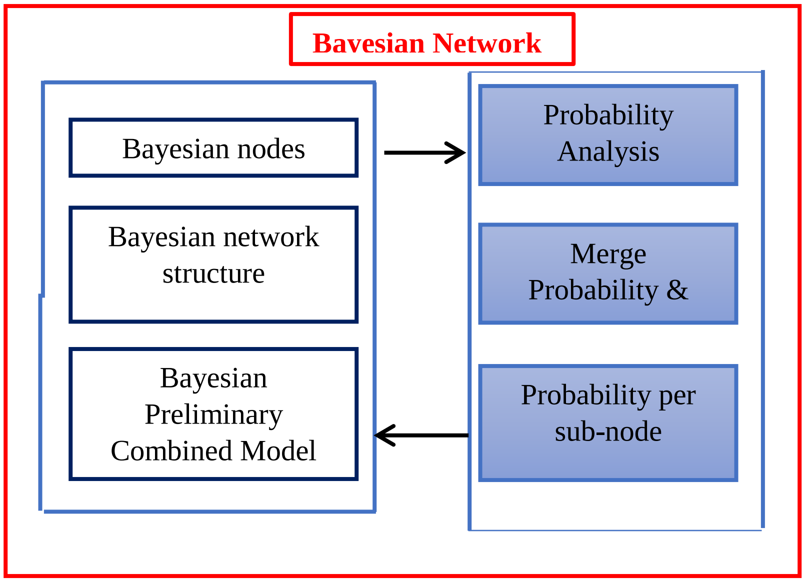


Figure 77: Show Bayesian Model

model, we were able to explain the relationship among malfunction of UG gas pipe components and observed system. Preliminary risk model was developed by using Bayesian network which in future work will determine the dominant risk factors involved (Figure 77). Furthermore, 775,000 out of 2 million data records was geo-coded and plotted by using Arc GIS which enabled us to perform preliminary spatial, and hotspot analysis. More specifically, the cluster analysis performed by Rapid miner showed that certain attributes has more cluster of damage than others which is clear indication of possible risk factor. The performed Hotspots analysis showed the cities which has more probability of UG gas damage. These cities can be used in future research to extract the probability to be used in Bayesian node. Perform cluster Analysis by using rapid miner to determine to examine what attributes contribute to the risk factors. In addition, performing the Hotspot, and cluster analysis will give a clear indication to the latent risk factors involved in UG gas pipe damages. After extracting the risk factors will build the network using Bow tie, and Fault tree method. Then will build Bayesian model and calculate the probability of the Bayesian Nodes. Agena risk software will be used to analysis the Bayesian network and produce the results.

1.36 Bayesian Structure.

1.36.1 Risk evolution process of underground gas pipe damage modeling with Bow-tie

In this research Bow-tie is used as an approach to integrate a fault tree to represent causes, threat (UG pipe damage) and consequences of multiple attributes. As stated before, traditional ‘bow-tie’ approach is not able to produce detailed risk model because each risk event is independent. Therefore, in order to deal with the complexity of the data, the Fault tree logic is employed. The goal is to derive probabilities (likelihood) of basic events in fault tree and to estimate nodes probabilities (likelihood) of output event consequences. Furthermore, Bow tie study model also explores how interdependencies among various risk factors might influence the results of the analyses. It also demonstrates different possible scenarios of UG gas pipe damages.

Methodology starts with defining the system and collecting the data. Collecting the underground gas pipe damaged data is the process of gathering one call center collected data and measuring information on targeted attributes in an established framework. This process enables us to answer relevant questions and evaluate risk outcomes. More specifically, in Bow-tie, the UG damage scenario is the link between damage and all its possible causes can be represented in the form of a fault tree. In the same time, the relationship between pipe damage and its possible multiple consequences can be represented by means of an event tree. Fault trees can be integrated in the form of a bow-tie diagram which then can be used to analyze underground gas pipe damages later: as their causes and consequences remain linked together. Moreover, this framework of Bow-tie provides the research with a simplified classification process where the usually varied information available in one call center damage reports can be consistently stored and summarized according to a set of fixed common criteria.

1.36.2 Developing the dependencies, and relationship of Bow-tie model

One of the main processes in Bow-tie risk model development is developing the dependencies. As shown in Figure 78, the process start with gas leakage. Excavator or the customer notices the leak and notifies the company. The company then notify one call center to manage the process: which is call before you dig code 811. Factors causing the gas leakage may vary widely. The backhoe may hit the gas pipe and cause the gas leak; dislocating the gas pipe, and marking up the wrong location also can cause UG gas pipe damage. Thus, the first introduced threat is excavator hit wrongly locating the UG gas pipe. This barrier was chosen as it does relate to case number 2 in Figure 78. The Tech consultant performs all required steps in terms of initiating the tickets with one call center, and then going through the steps with UG operator by having those checking utilities in the area of the incident. However, because of the mis-locating street number which mistakenly located the UG gas pipe on 820 West St instead of 1820 West St.

Consequently, the excavator to hit the UG gas pipe, and cause the UG gas incident? In addition, there are many other scenarios where is the UG gas pipe can be at risk. The second factor is the reliability of One Call center that manages information distribution and communication flow process. Communication plays vital role in making sure information flows to the right direction and to the right party. Once call center was chosen as a barrier because it does play a vital role in preventing UG gas damages. The reason is that if we compare the UG gas pipe damages between years 2009-14 in terms of specific categories such as contractor, we find that total UG gas damages in year 2009 by a contractor were 1350 incidents compared to 656 in year 2014. Which means one call center played a significant role on mitigating the risk of UG gas pipe incidents.

As shown in Figure 78, the process flows through UG operator response barrier, third party locating barrier, and finally the preventive barrier of risk of hitting excavator. Moreover, the UG operators receives the request from one call center and then transfer the request to locating party or in house staff to go out and locate the UG gas pipe. Finally, if nothing works from the preventive barriers, the gas pipe damage happens and the risk happens.

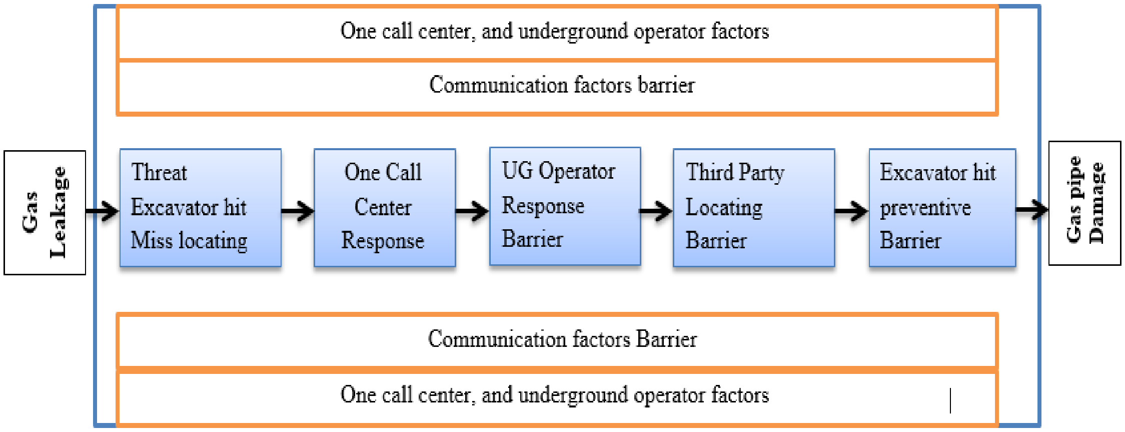

One call center, and underground operator factors
Communication factors barrier
Gas Leakage
Threat Excavator hit Miss locating
One Call Center Response
UG Operator Response Barrier
Third Party Locating Barrier
Excavator hit preventive Barrier
Gas pipe Damage
Communication factors Barrier
One call center, and underground operator factors

Figure 78 Bow-tie model for UG gas pipe damage

1.36.3 Risk evolution process of underground gas pipe damage modeling with Fault tree.

While the risk factors may include multiple factors (e.g., gas leakage, interference from the third party, material defects, malfunction, and natural hazard), references show that most of the pipelines fail in a mode of leak. This is because the gas transmission pipelines are mainly installed underground. Therefore, UG operators of the pipeline have to ensure safety and reliability of them. In this paper, a fault tree of the UG gas pipe damage model was constructed Figure 79 to evaluate risk event and Gas pipe damage was defined as the top event of the fault tree. Three kinds of damages modes, such as mislocating, miscommunication, and excavator defect, were considered as sub-top events. This fault tree was comprised of 33 basic events.

The Fault tree was used in this research as extension to Bow-tie method. By using Fault tree method, it was possible to develop a more detailed risk network structure, risk cause, and risk consequence network. Next, the developed risk model from FTA will be used to identify nodes, basic events.

1.36.4 Developing the dependencies, and relationship of Fault tree model

The developed risk model of UG gas pipe of a fault tree creates a visual record of the logical relationships between gas pipe damage events and causes. FTA is a useful tool to understand the results of UG gas pipe incidents analysis and pinpoint weaknesses in the design and identify risk events. Moreover, the developed FTA flow chart will help prioritize risk events. Based on comprehensive analysis of many UG gas pipe networks damages, we propose 14 nodes under three main influence factors as shown in Figure 79:

- **Locating defect:** Mismarking the UG gas pipe is one of the common causes of UG gas pipe damage. The UG gas pipe will break down if the gas pipe the exact location of the gas pipe is incorrectly located or marked (Figure 79). For example, in case number 2 in Figure 79, the Tech consultant performed all required steps in terms of initiating the tickets with one call center. However, because of the miss locating street number which mistakenly located the UG gas pipe on 820 West St instead of 1820 West St. This resulted in the excavator to hit the UG gas pipe, and caused the UG gas pipe damage.

Miscommunication: transferring the right information to the right party such as underground operator, excavator, and one call center on the right time is critical for protecting UG gas pipe from damages. For example, in case 2 of gas line damage, a backhoe was digging a trench behind a building; then the backhoe operator damaged a ¾-inch steel natural gas service line (Figure 79). This resulted in two leaks in the natural gas service line, which was operated at 35 psig. The reason is miscommunication between the contractor and locating party as the contractor told investigators that blue paint was used to mark both service lines because that was the only paint that they had. However, representative later could not find any blue or other line markings on the ground at the accident scene.

Excavator defect: One of the causes of UG gas pipeline damages is excavator mistakenly hit the UG gas pipe. This can be caused by many factors such as marking out the wrong location of the UG gas pipe, the excavator uses the wrong depth of the gas pipe (Figure 39).

One of the steps is to identify the effect of miscommunication on the failure of UG gas pipe. This research contains a complex network with undefined risk events, which caused by combinations of other risk events: rather than a low-level of damage with simple causes. Next step is to identify the UG pipe damage effect in a box at the top-center of the diagram area.

1.36.4.1 Detailed explanation of developing Fault Tree (FTA) network

Excavator Defect: one of the causes of UG gas pipe damage is excavator defect which in turn is influenced by excavator type (Figure 79). The excavator type contains five categories which are commercial, utility maintenance, general excavator, homeowner, and private contractor. All these types have different influence over the risk probability of the UG gas pipe been damaged. Another factor in this group is excavator zip code. The excavators were grouped by zip codes and risk probability will be derived based on the zip code.

Miscommunication defect: this contains two categories: Locating request, and UG Facility Operator (Figure 79). Locating request means that when the digging was required to do maintenance, the one call center is supposed to make request to locate the UG gas pipe. The excavator has no chance to know whether locating request has been made to the

One Call Center. In addition, there are two sub-categories under locating request: the number of outgoing calls per month; and the number of outgoing calls per day. Number of outgoing calls represents how many requests the one call center receive per day or month. As been derived from the analysis of the collected data, the volume of the daily and monthly request calls has direct impact on the UG gas pipe damage (Figure 79). After one call center receives request, the operator transfers the request to check what utilities in the area of digging. They will also check how many companies own theses utilities to inform them ongoing activities. UG facility operator plays a major rule on transferring the right information to the right utility company. If UG facility operator miss utilities in the area of the digging and does not inform them of the digging that may cause damage to the UG gas pipe. Under a UG facility operator, there are two sub-categories: number of damages per year and number of damages.

Locating Defect: locating the area of the digging is one of the main factors that plays a vital rule in UG gas pipe damage. When the locating contractor receives a request, the contractors go out to the field and put mark out, signs around the area of gas pipe. There are three main categories which could cause the damage of UG gas pipe damage: these are pipe size, damage zip code, and pipe material types.

When the collected damaged data was analyzed (Figure 79), there is direct correlation between the pipe size and the frequency of the damage. For example, pipe size ½" has more frequency of being damaged than pipe size 4". In addition, the pipe material of the gas pipe has impact on the probability of the risk involved in the damage; the plastic has more frequency of damage than steel.

As shown in figure 79, there are three types of issues: locating defect, miscommunication, and excavator defect. The majority of the damage events which may occur due the third party not locating the UG gas pipe when they receive the request. Miscommunication covers what could

go wrong in the information flow between the one call center, UG operator, and Locating

party. Excavator defect is another main contributing factor that is illustrated in Figure 79,to overcome the complexity of the model.

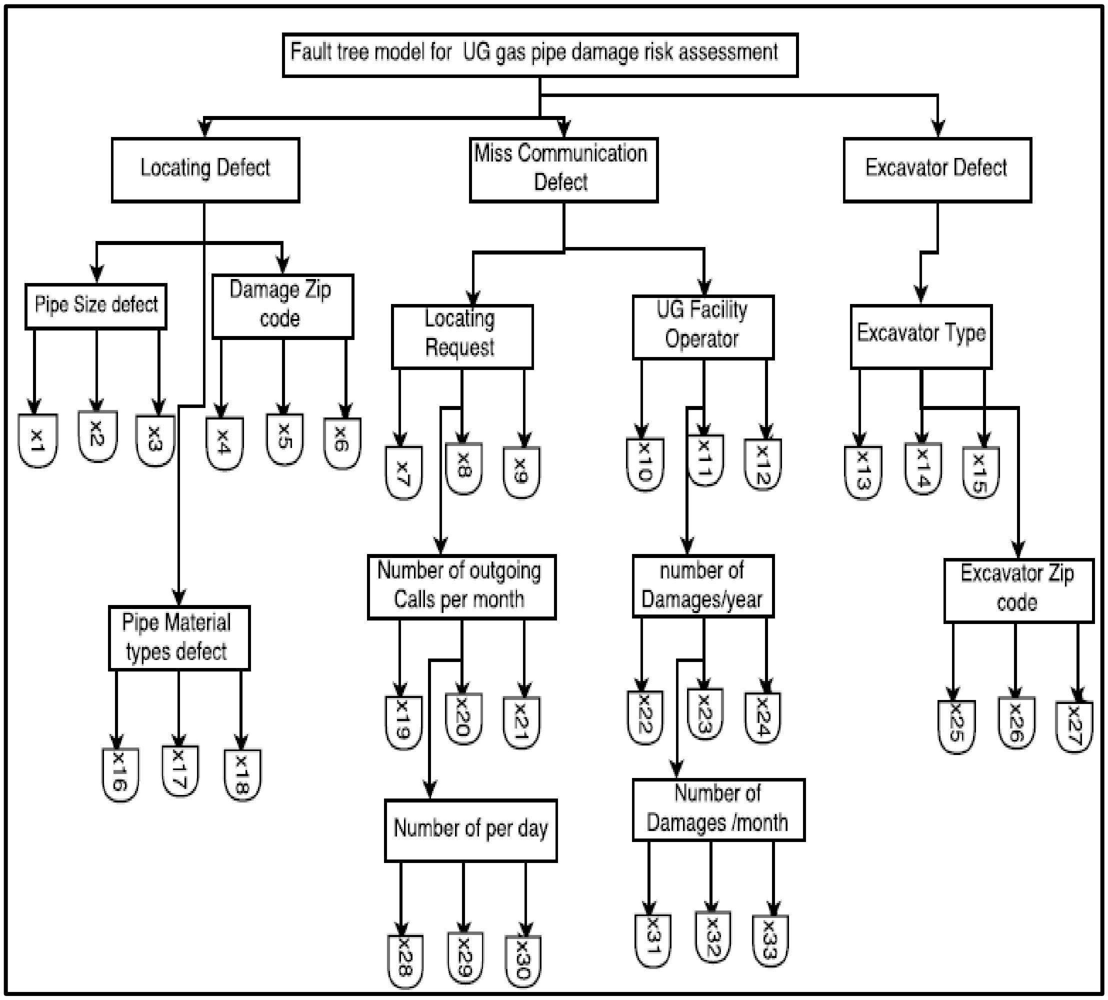


Figure 79: Fault Tree for Underground Gas Pipe Damage

1.37 Statistical Inferences (Bivariate, and Univaraite) Analysis.

Univaraite analysis is the simplest form of analyzing data. “Uni” means “one”; so in other words data has only one variable which is the time. It does not deal with UG gas pipe group of damages or relationships (unlike regression). Its major purpose is to describe; it takes data, summarizes that data and finds patterns in the data. We will describe patterns found in Univariate attribute include central tendency (mean, mode and median) and dispersion: range, variance, maximum, minimum, quartiles (including the interquartile range), and standard deviation. There are several options for describing data with Univaraite outcome. Some techniques, Frequency Distribution Tables, Bar Charts, Histograms, Frequency Polygons, Pie Charts, Histograms. Bivariate analysis, analysis with two UG gas pipe attributes that can change and are compared to find relationships. If one variable is influencing another variable, then we have bivariate data that has an independent (UG gas pipe data attributes) and a dependent variable (Target, Yes, NO). An independent variable is a condition or piece of data in an experiment that can be controlled or changed. Dependent variables (Predictor) are a condition or piece of UG pipe data in an experiment that is controlled or influenced by an outside factor: most often the independent variable (which is in our case damage).

1.37.1 Univaraite (Descriptive) Analysis (Time AM/PM)

The time was converted from (AM/PM) before, and afternoon to 24 hrs (24hrs) (1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24) (Figure 80). First the time was divided into (AM, PM), and then divided to 24 hrs. Univaraite and Bivariate analysis was performed. Findings, Timing attribute of the digging was determined to have significant effect on the UG gas pipe damage. Moreover, when all hours were analyzed as one attribute, P-Value was 0.001 which is less than the industry standard of 0.05 table 5. More specifically, 24hrs were analyzed separately; some hours have more damage percent than others. For, Instance, the damage percent within hour 8 represent 14.2% as compared to the rest of the 24 hrs, which is the highest percent of damage among all hours table 3. In addition, the damage percent within hour 10, 13, and 15 represent 11.6%, 11.4%, 11.3% consecutively as compared to the rest of the 24 hrs table 4. Therefore, the specific hours have significant impact on the UG gas pipe damage. In Summary, the received UG gas pipe digging requests by the agency during 10, 13, and 15 have more chance of been damaged than the other hours.

Table 2: Shows Descriptive Analysis for 24 hrs Attributes

Descriptive Statistics

	N	Range	Minimum	Maximum	Sum	Mean	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error
Time 24	396547	23	0	23	4786156	12.07	.005
Valid N (lis twise)	396547						

Table 3 Shows Discriptive Analysis (St. Deviation, Skewness, Kurtosis)

Descriptive Statistics

	Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Time 24	3.319	11.016	.270	.004	.085	.008

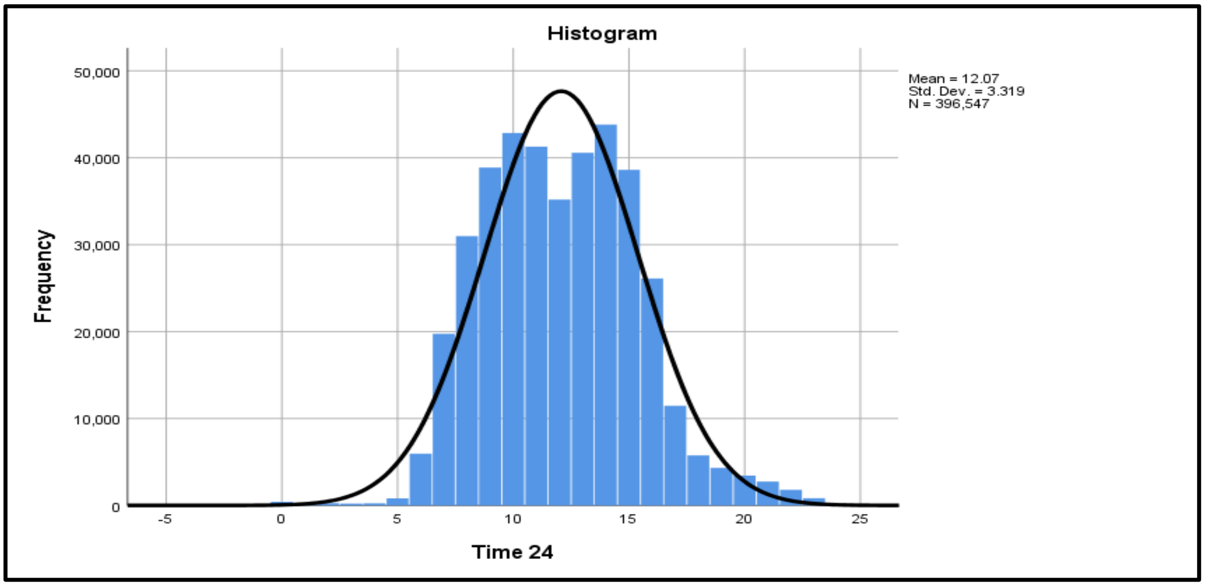


Figure 80 Shows Mean, Std. Deviation, And # of Frequencies

1.37.2 Bivariate Analysis (Time AM/PM) & 24hr

Table 4: Bivariate Analysis for the Time

Time 24 * DAMAGE Cross Tabulation					
			DAMG		Total
			0	1	
	8hr	Count	30835_a	150_b	30985
		Expected Count	30902.7	82.3	30985.0
		% within Time 24	99.5%	0.5%	100.0%
		% within DAMG	7.8%	14.2%	7.8%
		% of Total	7.8%	0.0%	7.8%
	9hr	Count	38770a	96a	38866
		Expected Count	38762.8	103.2	38866.0
		% within Time 24	99.8%	0.2%	100.0%
		% within DAMG	9.8%	9.1%	9.8%
		% of Total	9.8%	0.0%	9.8%
	10hr	Count	42722a	122a	42844
		Expected Count	42730.2	113.8	42844.0
		% within Time 24	99.7%	0.3%	100.0%
		% within DAMG	10.8%	11.6%	10.8%
		% of Total	10.8%	0.0%	10.8%

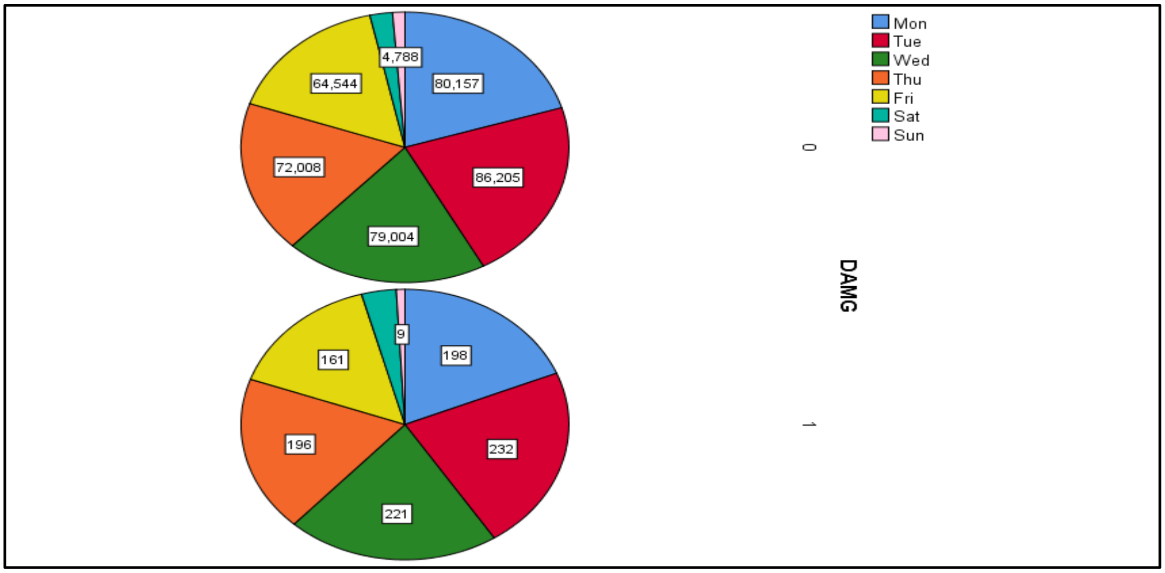
	11hr	Count	41169a	103a	41272
		Expected Count	41162.4	109.6	41272.0
		% within Time 24	99.8%	0.2%	100.0%
		% within DAMG	10.4%	9.8%	10.4%
		% of Total	10.4%	0.0%	10.4%
	12hr	Count	35144a	41b	35185
		Expected Count	35091.6	93.4	35185.0
		% within Time 24	99.9%	0.1%	100.0%
		% within DAMG	8.9%	3.9%	8.9%
		% of Total	8.9%	0.0%	8.9%
	13hr	Count	40458a	120a	40578
		Expected Count	40470.2	107.8	40578.0
		% within Time 24	99.7%	0.3%	100.0%
		% within DAMG	10.2%	11.4%	10.2%
		% of Total	10.2%	0.0%	10.2%
	14hr	Count	43697a	111a	43808
		Expected Count	43691.7	116.3	43808.0
		% within Time 24	99.7%	0.3%	100.0%
		% within DAMG	11.0%	10.5%	11.0%
		% of Total	11.0%	0.0%	11.0%
	15hr	Count	38491a	119a	38610
		Expected Count	38507.5	102.5	38610.0
		% within Time 24	99.7%	0.3%	100.0%
		% within DAMG	9.7%	11.3%	9.7%
		% of Total	9.7%	0.0%	9.7%
	16hr	Count	26040a	71a	26111
		Expected Count	26041.7	69.3	26111.0
		% within Time 24	99.7%	0.3%	100.0%
		% within DAMG	6.6%	6.7%	6.6%
		% of Total	6.6%	0.0%	6.6%
Total		Count	395494	1053	396547
		Expected Count	395494.0	1053.0	396547.0
		% within Time 24	99.7%	0.3%	100.0%
		% within DAMG	100.0%	100.0%	100.0%
		% of Total	99.7%	0.3%	100.0%

Table 5: Shows P-Value for the Time (AM/PM)

P- Value (Chi-Square Tests)			
	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	147.633[a]	23	.001

1.37.3 Univaraite (Descriptive) Analysis Mon, Tue, Wed, Thu, Fri, Sat, and Sun (Week Days)

The week as one attribute was determined to be significant figure 81, and table 6. However, as observed from the UG gas pipe damaged data distribute, the damage was happening across all week days. Thus, Univaraite, and Bivariate statistics analysis was performed on each day separate to see the impact for each individual day on the UG gas pipe damage. Moreover, as per conducted analysis, one days was having significant impact on UG gas pipe damage, Saturday, , P values were 0.045. Thus, by avoiding digging operation on Saturday, that will significantly decrease the damage of UG gas pipe damage. As can be seen in the detailed analysis above, P value for the rest of the week days was more than 0.05 which means it does not have significant impact table 7, & 8. In addition, odds ratio for Thursday, and Wednesday was 1.3. This means these two days have 1.29 chance of getting damaged as compared to rest of the week days. Results, any received digging requests during Thursday, and Wednesday will have 29% chance of been damage zthan



the other days.

Figure 81:Damage and Undamaged across week days

Table 6: Shows Descriptive Statistics for Week Days

Descriptive Statistics

	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Std. Error	Statistic	Std. Error
Mon	.162	1.480	.004	.189	.008
Tue	.170	1.366	.004	-.134	.008
Wed	.160	1.502	.004	.255	.008
Thu	.149	1.648	.004	.715	.008
Fri	.137	1.823	.004	1.324	.008
Sat	.022	6.478	.004	39.963	.008
Sun	.012	8.926	.004	77.679	.008

1.37.4 Bivariate Analysis Mon, Tue, Wed, Thu, Fri, Sat, and Sun (Week Days)

The Bivariate analysis was conducted for all days. In order to find out which day is more significant in causing/contributing to underground gas pipe damage, P-Value was calculated for each day Separate starting from Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday. for illustration purpose, Monday calculations are shown below in table 7, 8, and figure 82.

1.37.4.1 Bivariate Analysis for Monday.

Table 7: Shows Bivariate Analysis for Monday

Bivariate Analysis Crosstab					
			DAMG		
			0	1	Total
Mon	0	Count	315337a	855a	316192
		% within Mon	99.7%	0.3%	100.0%
		% within DAMG	79.7%	81.2%	79.7%
		% of Total	79.5%	0.2%	79.7%
		Standardized Residual	.0	.5	
	1	Count	80157a	198a	80355
		% within Mon	99.8%	0.2%	100.0%
		% within DAMG	20.3%	18.8%	20.3%
		% of Total	20.2%	0.0%	20.3%
		Standardized Residual	.1	-1.1	
Total		Count	395494	1053	396547
		% within Mon	99.7%	0.3%	100.0%
		% within DAMG	100.0%	100.0%	100.0%
		% of Total	99.7%	0.3%	100.0%
Each subscript letter denotes a subset of DAMG categories whose column proportions do not differ significantly from each other at the .05 level.					

Table 8: Shows P Value for Monday

P- Value (Chi-Square Tests)					
	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1.393[a]	1	.5942		

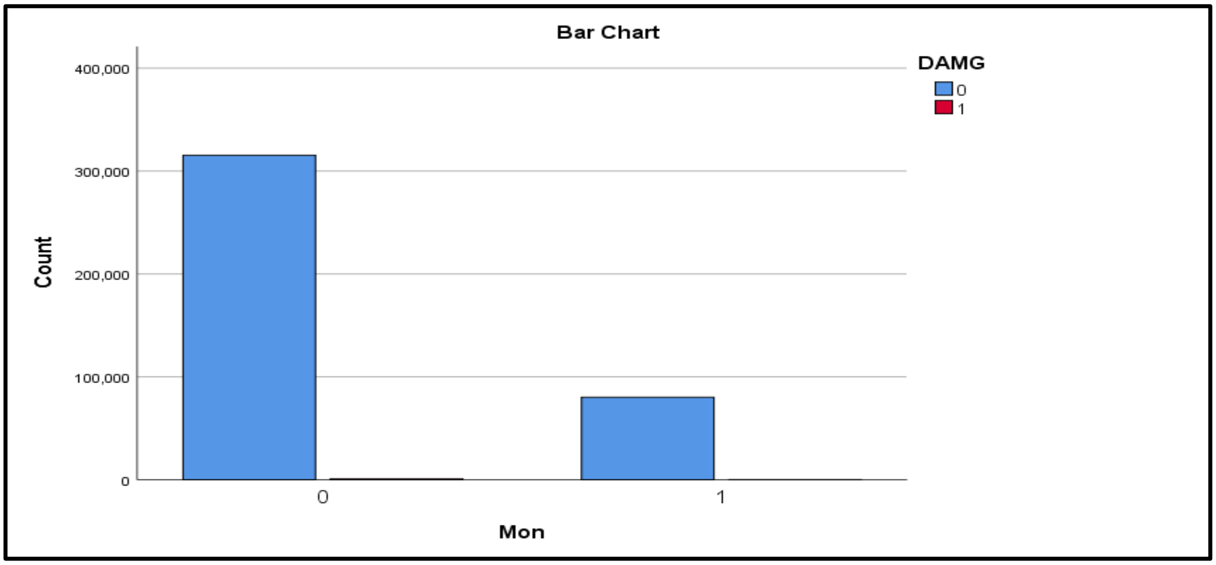


Figure 82 Show total Ticket Numer/Yera, Total undamages in Monday, & Total Damages in Monday

1.37.5 Univaraite (Descriptive) Analysis (Months)

Table 9: Show Descriptive Analysis for Months

Descriptive Statistics

	N	Range	Minimum	Maximum	Sum	Mean	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error
Jan	396547	0	0	0	0	.00	.000
Feb	396547	1	0	1	5	.00	.000
Mar	396547	1	0	1	45940	.12	.001
Apr	396547	1	0	1	53826	.14	.001
May	396547	1	0	1	47986	.12	.001
Jun	396547	1	0	1	50065	.13	.001
Jul	396547	1	0	1	42606	.11	.000
Aug	396547	1	0	1	44523	.11	.001
Sep	396547	1	0	1	35552	.09	.000
Oct	396547	1	0	1	23477	.06	.000
Nov	396547	1	0	1	33455	.08	.000
Dec	396547	1	0	1	19112	.05	.000
Valid N (list wise)	396547						

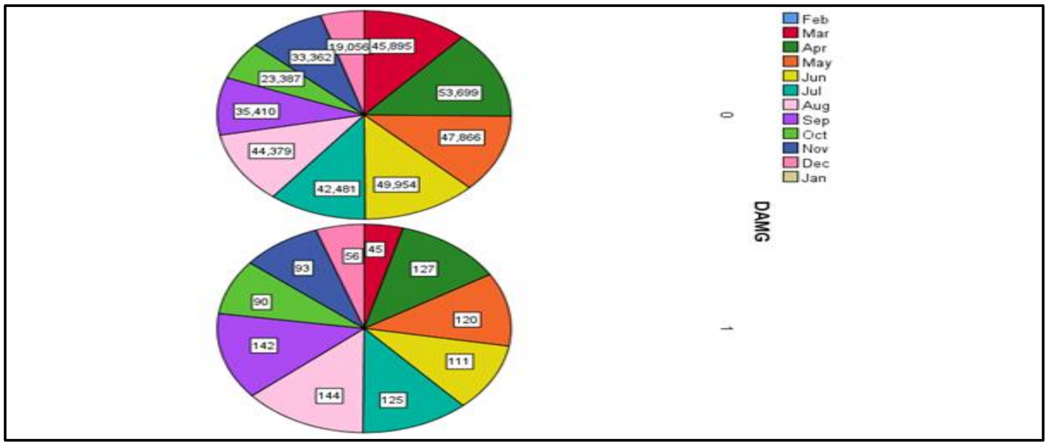


Figure 83: Show Damage, and Undamaged across Months

1.37.6 Bivariate Analysis (Months)

The Bivariate analysis was conducted separate for all months. In order to find out which month is more significant in causing/contributing to underground gas pipe damage, P-Value was calculated for each month separate, starting from Jan, Feb, Mar, April, May, Jun, July, Aug, Sept, Oct, Nov, and Dec. Bivariate analysis was conducted separate for all months together. In order to find out which month is more significant in causing/contributing to underground gas pipe damage, P-Value was calculated for each month separate, starting from Jan, Feb, Mar, April, May, Jun, July, Aug, Sept, Oct, Nov, and Dec. Results, the month of March, June, Aug, Sept, and Oct have P-value less than 0.05 table 18, and table 19 which means these months have significant impact on the UG gas pipe damage. In addition, by avoiding digging UG gas pipe during March, June, Aug, Sept, and Oct will have positive impact on the UG Gas pipe.

1.37.6.1 Bivariate Analysis for March.

Table 10: Shows Bivariate for March

Crosstab

			DAMG		
			0	1	Total
Mar	0	Count	349599_a	1008_b	350607
		% within Mar	99.7%	0.3%	100.0%
		% within DAMG	88.4%	95.7%	88.4%
		% of Total	88.2%	0.3%	88.4%
		Standardized Residual	-.1	2.5	
	1	Count	45895_a	45_b	45940
		% within Mar	99.9%	0.1%	100.0%
		% within DAMG	11.6%	4.3%	11.6%
		% of Total	11.6%	0.0%	11.6%
		Standardized Residual	.4	-7.0	
Total		Count	395494	1053	396547
		% within Mar	99.7%	0.3%	100.0%
		% within DAMG	100.0%	100.0%	100.0%
		% of Total	99.7%	0.3%	100.0%

Each subscript letter denotes a subset of DAMG categories whose column proportions do not differ significantly from each other at the .05 level.

Table 11: Show P Value for March

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	55.103[a]	1	.000		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 121.99.
b. Computed only for a 2x2 table

1.37.7 Univaraite (Descriptive) Analysis (Seasons)

Table 12: Show Descriptive Analysis for all Seasons

Descriptive

	Season	Statistic	Std. Error

Damage binary	Autumn			
		Variance	.004	
		Std. Deviation	.059	
		Minimum	0	
		Maximum	1	
		Range	1	
		Skewness	16.780	.008
		Kurtosis	279.585	.016
	Spring			
		Variance	.002	
		Std. Deviation	.044	
		Minimum	0	
		Maximum	1	
		Range	1	
		Interquartile Range	0	
		Skewness	22.428	.006
		Kurtosis	501.019	.013
	Summer			
		Variance	.003	
		Std. Deviation	.053	
		Minimum	0	
		Maximum	1	
		Range	1	
		Skewness	18.922	.007
		Kurtosis	356.053	.013
	Winter	Variance	.003	
		Std. Deviation	.054	
		Minimum	0	
		Maximum	1	
		Range	1	
		Interquartile Range	0	
		Skewness	18.396	.018
		Kurtosis	336.466	.035

1.37.7.1 Bivariate Analysis Seasons (Autumn, Spring, Summer, Winter)

This attribute was split into four seasons, Autumn, Spring, Summer, And Winter figure 88. As per the UG gas pipe data distribution, damages were happening across the four seasons. Statistics analysis was performed for each season separate. Results, P-Value for autumn, and spring were less than 0.05, and for summer, and winter was more than 0.05 table 39, and table 40. Which means digging during autumn, and spring could cause more damage to UG gas pipe than digging in summer, and winter. More detailed analysis included under Month Section above.

Table 13: Show Bivariate for all seasons

Season * DAMG Cross tabulation

Count

		DAMG		Total
		0	1	
Season	Autumn	92159_{a}	325_{b}	92484
	Spring	147460_{a}	292_{b}	147752
	Summer	136814_{a}	380_{a}	137194
	Winter	19061_{a}	56_{a}	19117
Total		395494	1053	396547

Each subscript letter denotes a subset of DAMG categories whose column proportions do not differ significantly from each other at the .05 level.

Table 14: P value for All Seasons

Chi-Square Tests for all Seasons Combined

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	52.701[a]	3	.000
Likelihood Ratio	52.801	3	.000
N of Valid Cases	396547		

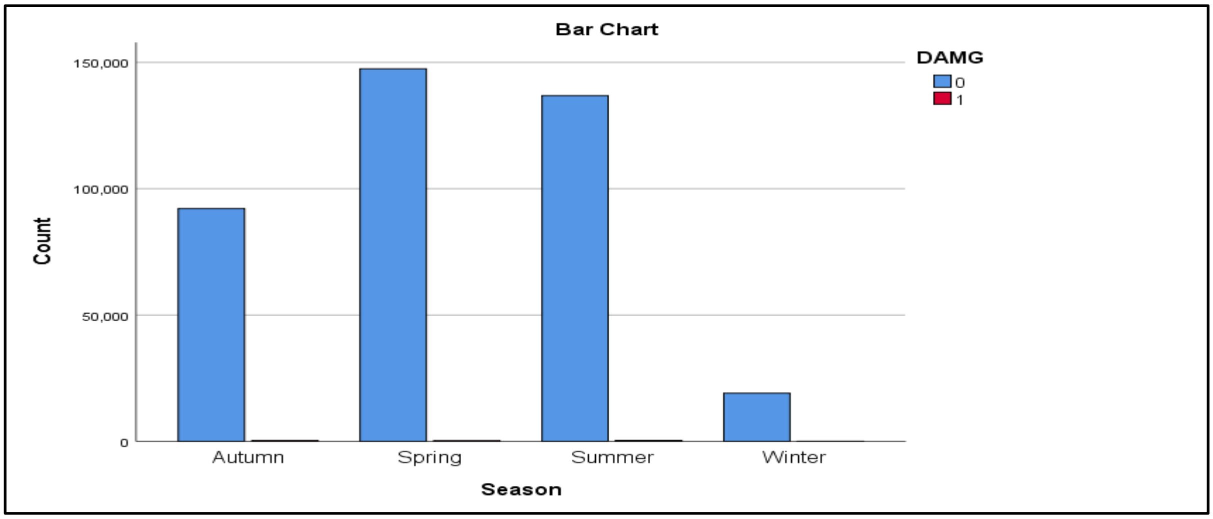


Figure 84 Total Undamaged & Damages per Season

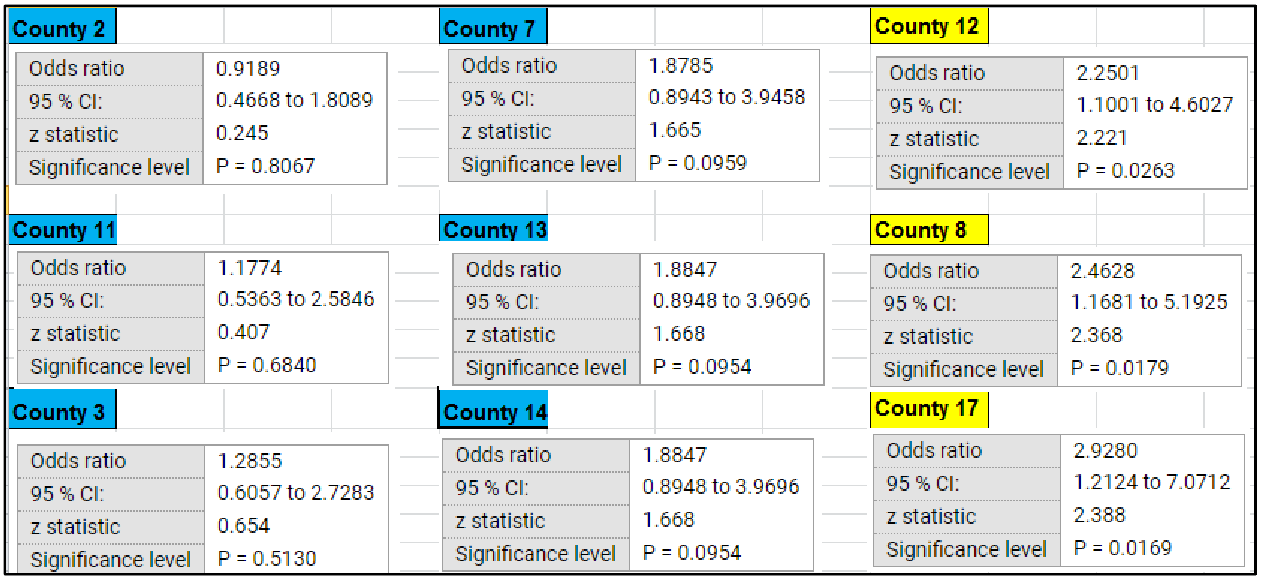
1.37.7.2 Spring P Value

Table 15: Show P Value for spring

Chi-Square Tests					
	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	41.013[a]	1	.000		

1.37.8 Univaraite (Descriptive) Analysis (County)

Statistics analysis was performed on all of the counties. For instance, County 5 has P-value equal to 0.001, and Odd ratio of 7.2. Which means county 5 has significant impact on the UG gas pipe damage, and any Gas pipe digging in county 5 has 7 times chance of been damage as compared to county 10 (reference county). All over all, 9 counties out of 21 counties total have P – value less than 0.05. As result, UG gas pipe digging in theses counties have significant impact on the gas pipe damage, and needed to be taken in consideration when digging request is received. In addition, by summarizing the odd ratios for all counties, and then classifying them in ascending order, we found the following; odds ratio in county # 5 equal to 7.3 figure 89, and 45. This means there is approximately 7 times more chance for the gas pipe to be damage in county 5 as compared to the referenced county. In ascending order, for counties 20, 1, 4, 6 and 8, the odd ratios were 3.3, 3.2, 3, 3, 2.9 , and 2.5 figure 90, and 45. This means any ticket requests received in these counties have around 3 times chance as compared to the referenced county.



County 2

Odds ratio	0.9189
95 % CI:	0.4668 to 1.8089
z statistic	0.245
Significance level	P = 0.8067

County 7

Odds ratio	1.8785
95 % CI:	0.8943 to 3.9458
z statistic	1.665
Significance level	P = 0.0959

County 12

Odds ratio	2.2501
95 % CI:	1.1001 to 4.6027
z statistic	2.221
Significance level	P = 0.0263

County 11

Odds ratio	1.1774
95 % CI:	0.5363 to 2.5846
z statistic	0.407
Significance level	P = 0.6840

County 13

Odds ratio	1.8847
95 % CI:	0.8948 to 3.9696
z statistic	1.668
Significance level	P = 0.0954

County 8

Odds ratio	2.4628
95 % CI:	1.1681 to 5.1925
z statistic	2.368
Significance level	P = 0.0179

County 3

Odds ratio	1.2855
95 % CI:	0.6057 to 2.7283
z statistic	0.654
Significance level	P = 0.5130

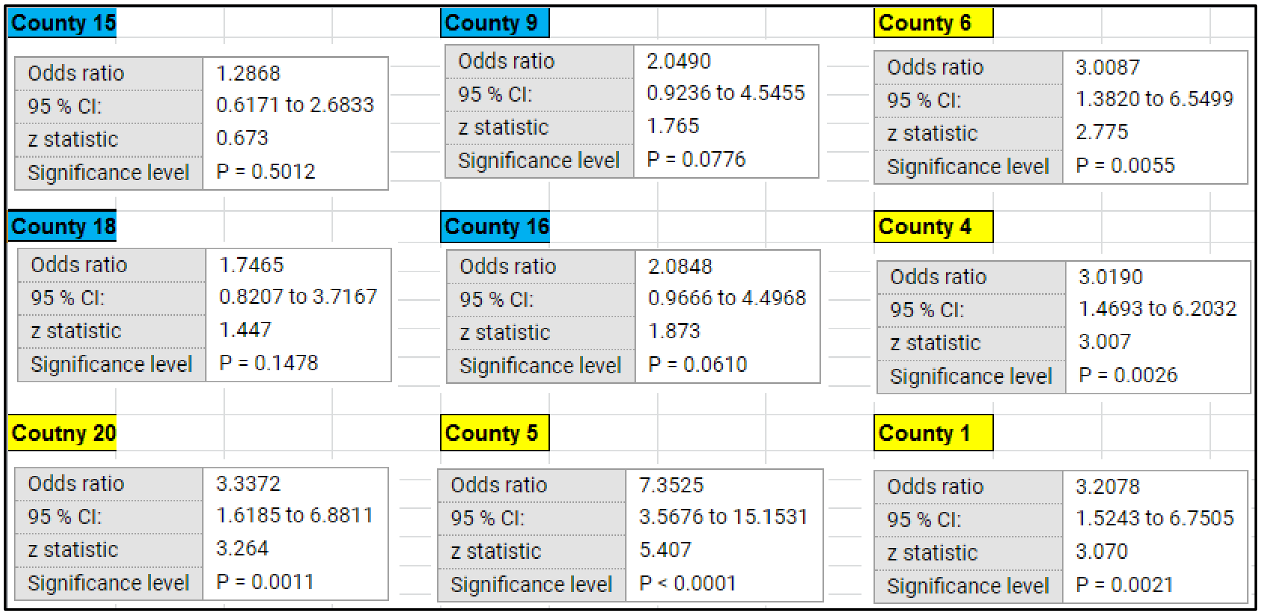
County 14

Odds ratio	1.8847
95 % CI:	0.8948 to 3.9696
z statistic	1.668
Significance level	P = 0.0954

County 17

Odds ratio	2.9280
95 % CI:	1.2124 to 7.0712
z statistic	2.388
Significance level	P = 0.0169

Figure 85:Shows counties with P Value & Odd Ratio



County 15

Odds ratio	1.2868
95 % CI:	0.6171 to 2.6833
z statistic	0.673
Significance level	P = 0.5012

County 9

Odds ratio	2.0490
95 % CI:	0.9236 to 4.5455
z statistic	1.765
Significance level	P = 0.0776

County 6

Odds ratio	3.0087
95 % CI:	1.3820 to 6.5499
z statistic	2.775
Significance level	P = 0.0055

County 18

Odds ratio	1.7465
95 % CI:	0.8207 to 3.7167
z statistic	1.447
Significance level	P = 0.1478

County 16

Odds ratio	2.0848
95 % CI:	0.9666 to 4.4968
z statistic	1.873
Significance level	P = 0.0610

County 4

Odds ratio	3.0190
95 % CI:	1.4693 to 6.2032
z statistic	3.007
Significance level	P = 0.0026

Coutny 20

Odds ratio	3.3372
95 % CI:	1.6185 to 6.8811
z statistic	3.264
Significance level	P = 0.0011

County 5

Odds ratio	7.3525
95 % CI:	3.5676 to 15.1531
z statistic	5.407
Significance level	P < 0.0001

County 1

Odds ratio	3.2078
95 % CI:	1.5243 to 6.7505
z statistic	3.070
Significance level	P = 0.0021

Figure 86: Shows sample of counties with P Value & Odd Ratio

Table 16: Shows The Risk Ratio, Odds Ratio, and P Value

County Code	Code/Value	Damage	Total Ticket / County	% Probability	Undamaged	Odds Ratio	P value
County 10	10	8	6273	0.001275307	6265		
County 2	2	52	35505	0.001464582	35453	1.148415716	0.7154
County 11	11	28	18652	0.001501179	18624	1.177112374	0.684
County 3	3	45	27459	0.001638807	27414	1.285029499	0.513
County 15	15	65	39622	0.001640503	39557	1.28635922	0.5012
County 18	18	43	19324	0.002225212	19281	1.744844494	0.1478
County 21	21	7	3033	0.002307946	3026	1.809718101	0.2514
County 7	7	55	22984	0.002392969	22929	1.876386834	0.0959
County 13	13	52	21659	0.00240085	21607	1.882566139	0.0954
County 14	14	52	21659	0.00240085	21607	1.882566139	0.0954
County 9	9	25	9580	0.002609603	9555	2.046255219	0.0776
County 16	16	35	13182	0.002655136	13147	2.081958352	0.061
County 12	12	123	42931	0.002865063	42808	2.246567166	0.0263
County 8	8	51	16268	0.003134989	16217	2.458223199	0.0179
County 17	17	13	3490	0.003724928	3477	2.920809456	0.0169
County 6	6	31	8100	0.00382716	8069	3.000972222	0.0055
County 4	4	102	26561	0.003840217	26459	3.011210045	0.0026
County 1	1	53	12992	0.004079433	12939	3.198785791	0.0021
County 20	20	90	21210	0.004243281	21120	3.327263083	0.0011
County 5	5	92	9891	0.009301385	9799	7.29344859	0.0001

1.37.9 Univaraite (Descriptive) Analysis (D10, D20, D30)

The input for this attribute was developed by geocoding the data, then plotting all data in ARC GIS, and after multiple processes including Joint Spatial, damages were calculated within 10 miles diameter. Processes were repeated for 20 milesand 30 miles diameter. Because all values are continuous swhich cannot be statistically analyzed, values were converted into ranges. Each range represents certain number of damages, then these ranges entered into SPSS for the statistics analysis. The tables below show only 50 samples of the total number (Table 5). The Ranges were classified by distributing the number of total damages into equal intervals of 4 ranges as following;

- Damages within 10 miles diameter
  - Divided into Ranges (Range 1 from 0-20 damages)
  - Divided into Ranges (Range 2 from 21-40 damages)
  - Divided into Ranges (Range 3 from 41-60 damages)
  - Divided into Ranges (Range 4 from 61-86damages)
- Damages within 20 miles diameter
  - Divided into Ranges (Range 1 from 0 - 49 damages)
  - Divided into Ranges (Range 2 from 50 - 98 damages)
  - Divided into Ranges (Range 3 from 99 - 147 damages)
  - Divided into Ranges (Range 4 from 148 – 196 damages)
- Damages within 30 miles diameter
  - Divided into Ranges (Range 1 from 0 - 73 damages)
  - Divided into Ranges (Range 2 from 74 - 147 damages)
  - Divided into Ranges (Range 3 from 148 - 222 damages)
  - Divided into Ranges (Range 4 from 223 – 295 damages)

Table 17: Classifying D10, D20, D30 into Four Ranges

D10	D10 (Ranges)	D20	D20 (Range)	D30	D30 (Range)
12	1	30	1	42	1
2	1	6	1	48	1
39	2	157	4	260	4
28	2	103	3	175	3
5	1	26	1	41	1
6	1	21	1	48	1
6	1	21	1	48	1
12	1	36	1	43	1
11	1	17	1	34	1
15	1	41	1	52	1
50	3	111	3	157	3
11	1	17	1	34	1
3	1	17	1	40	1
1	1	18	1	48	1
9	1	29	1	69	1
10	1	25	1	85	2
50	3	111	3	157	3
17	1	41	1	49	1
17	1	106	3	200	3
22	2	106	3	200	3
3	1	9	1	31	1
2	1	6	1	42	1
2	1	6	1	42	1
15	1	49	1	118	2
19	1	38	1	45	1
23	2	71	2	178	3
21	2	38	1	45	1
21	2	38	1	45	1
9	1	27	1	40	1
20	1	41	1	45	1
17	1	41	1	49	1
0	1	6	1	37	1
16	1	36	1	44	1
39	2	102	3	231	4
7	1	29	1	40	1
16	1	32	1	92	2
16	1	32	1	92	2
16	1	32	1	92	2
16	1	32	1	92	2
30	2	127	3	218	3
18	1	65	2	162	3
25	2	68	2	145	2
25	2	70	2	145	2
7	1	25	1	77	2
60	3	162	4	259	4
7	1	25	1	77	2
15	1	62	2	177	3
15	1	62	2	177	3
13	1	50	1	121	2
20	1	73	2	171	3

1.37.9.1 Bivariate Analysis (D10).

Table 18: Biveriate Analysis for D 10

DAMAGE within 10 mile Diameter Cross Tabulation

			DAMG		Total
			0	1	
Damage	Range 1	Count	225180	554	225734
		% with in low to high level of near area damage	99.8%	0.2%	100.0%
		% within DAMG	56.9%	52.6%	56.9%
		% of Total	56.8%	0.1%	56.9%
	Range 2	Count	116167	311	116478
		% within low to high level of near area damage	99.7%	0.3%	100.0%
		% within DAMG	29.4%	29.5%	29.4%
		% of Total	29.3%	0.1%	29.4%
	Range 3	Count	26929	89	27018
		% within low to high level of near area damage	99.7%	0.3%	100.0%
		% within DAMG	6.8%	8.5%	6.8%
		% of Total	6.8%	0.0%	6.8%
	Range 4	Count	27217	99	27316
		% within low to high level of near area damage	99.6%	0.4%	100.0%
		% within DAMG	6.9%	9.4%	6.9%
		% of Total	6.9%	0.0%	6.9%
Total		Count	395493	1053	396546
		% within low to high level of near area damage	99.7%	0.3%	100.0%
		% within DAMG	100.0%	100.0%	100.0%
		% of Total	99.7%	0.3%	100.0%

Table 19: P Value for D10

P Value / Chi-Square Tests (D 10)			
	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	17.303[a]	3	.001

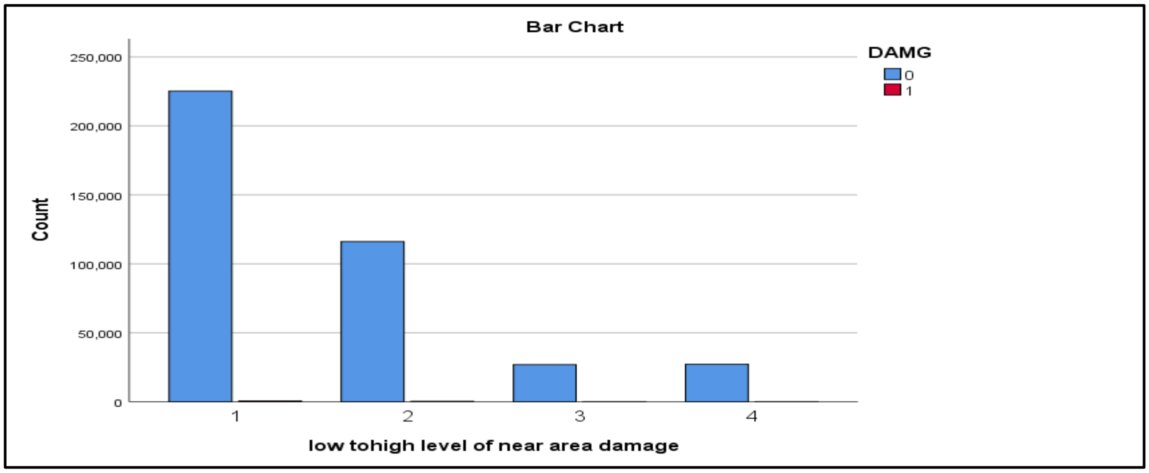


Figure 87 : Shows classification of Damages, and Undamaged per Range (D10)

1.37.9.2 Univaraite (Descriptive) Analysis (D20)

1.37.9.3 Bivariate Analysis (D20)

Table 20: Biveriate Analysis for D 20

Low to High D20 * DAMAGE Cross Tabulation					
			DAMG		
			0	1	Total
low to high D20	Range 1.00	Count	154390	375	154765
		% within low to high D20	99.8%	0.2%	100.0%
		% within DAMG	39.0%	35.6%	39.0%
		% of Total	38.9%	0.1%	39.0%
	Range 2.00	Count	119629	314	119943
		% within low to high D20	99.7%	0.3%	100.0%
		% within DAMG	30.2%	29.8%	30.2%
		% of Total	30.2%	0.1%	30.2%
	Range 3.00	Count	85125	256	85381
		% within low to high D20	99.7%	0.3%	100.0%

		% within DAMG	21.5%	24.3%	21.5%
		% of Total	21.5%	0.1%	21.5%
	Range 4.00	Count	36349	108	36457
		% within low to high D20	99.7%	0.3%	100.0%
		% within DAMG	9.2%	10.3%	9.2%
		% of Total	9.2%	0.0%	9.2%
Total		Count	395493	1053	396546
		% within low to high D20	99.7%	0.3%	100.0%
		% within DAMG	100.0%	100.0%	100.0%
		% of Total	99.7%	0.3%	100.0%

Table 21: Show P Value for D 20

Chi-Square/ P Value Tests			
	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	8.308a	3	.040
a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 96.81.			

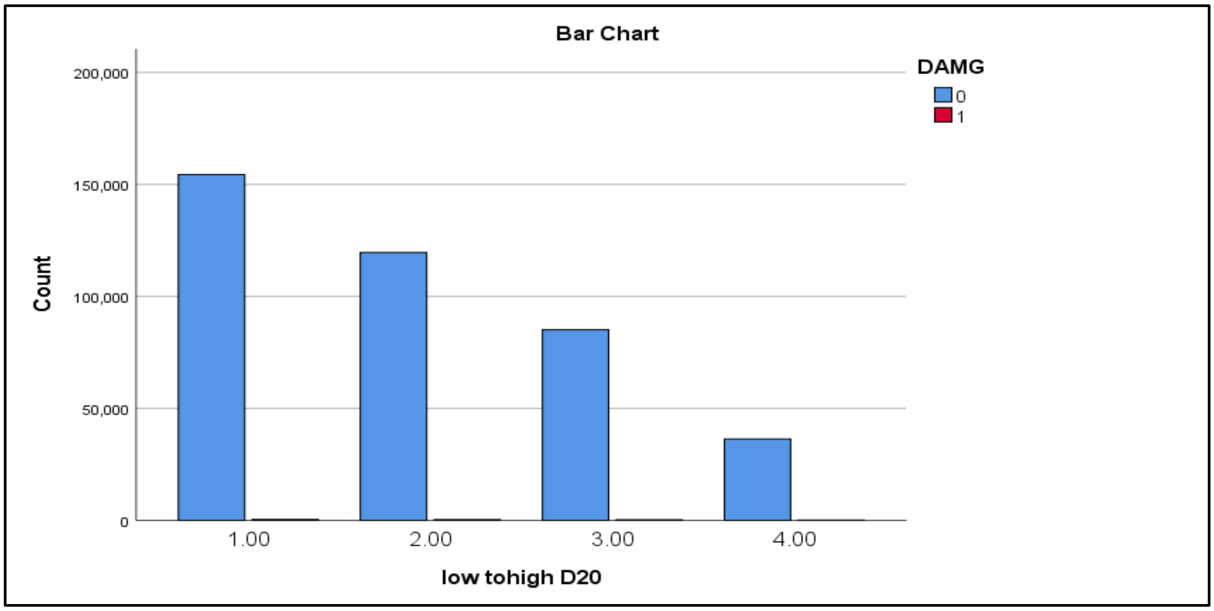


Figure 88 Show classification of Damages, and Undamaged per Range (D20)

1.37.9.4 Univaraite (Descriptive) Analysis (D30)

1.37.9.5 Bivariate Analysis (D30)

Table 22: Show Bivariate Analysis for D 30

Low to High D 30 * DAMG Cross Tabulation

			DAMG		Total
			0	1	
Low to High D 30	Range 1.00	Count	98560	266	98826
		% within Low to High D 30	99.7%	0.3%	100.0%
		% within DAMG	24.9%	25.3%	24.9%
		% of Total	24.9%	0.1%	24.9%
	Range 2.00	Count	126567	308	126875
		% within Low to High D 30	99.8%	0.2%	100.0%
		% within DAMG	32.0%	29.2%	32.0%
		% of Total	31.9%	0.1%	32.0%
	Range 3.00	Count	99066	280	99346
		% within Low to High D 30	99.7%	0.3%	100.0%
		% within DAMG	25.0%	26.6%	25.1%
		% of Total	25.0%	0.1%	25.1%
	Range 4.00	Count	71300	199	71499
		% within Low to High D 30	99.7%	0.3%	100.0%
		% within DAMG	18.0%	18.9%	18.0%
		% of Total	18.0%	0.1%	18.0%
Total		Count	395493	1053	396546
		% within Low to High D 30	99.7%	0.3%	100.0%
		% within DAMG	100.0%	100.0%	100.0%
		% of Total	99.7%	0.3%	100.0%

Table 23: Show P value for D 30

Chi-Square/ P Value Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	3.974[a]	3	.264

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 189.86.

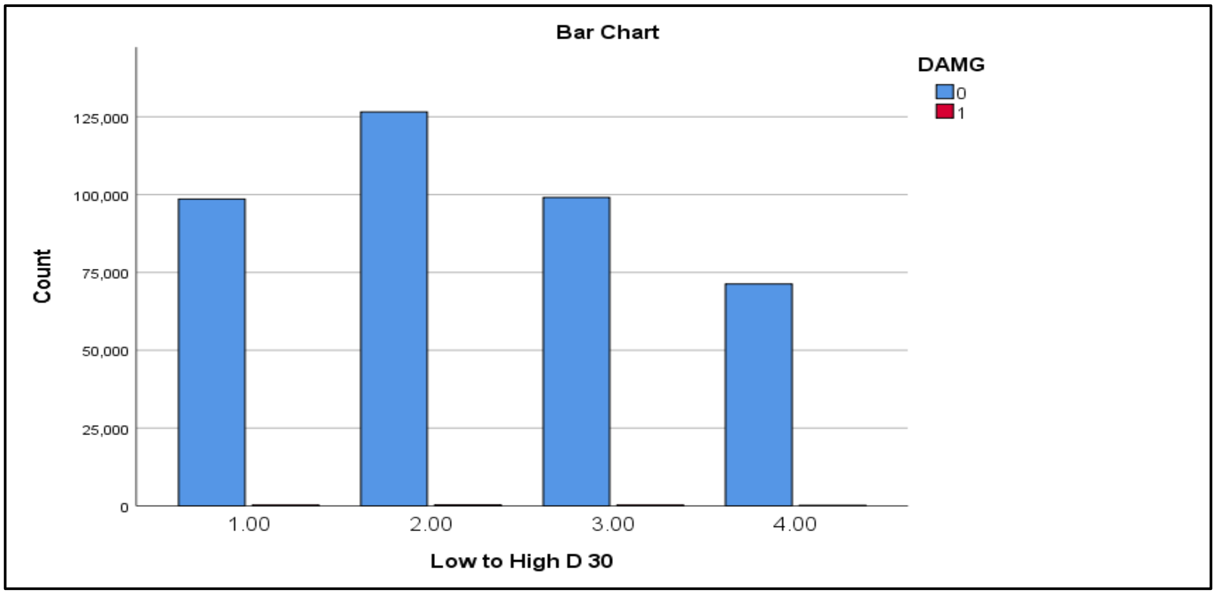


Figure 89 Figure : Classification of Damages, and Undamaged per Range (D30)

1.38 Determine the conditional probability for the factors.

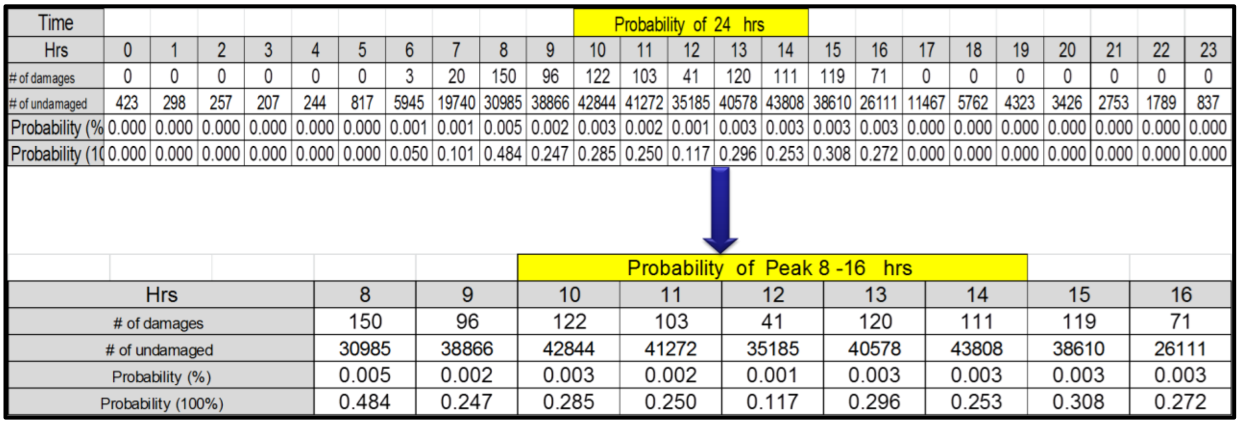
When calculating the probability of the underground gas pipe damage, it is important to compute the probability for each node in the Bayesian Network. There are many ways probability of the nodes can be calculated: for example, by Survey or asking expertise. For example, is this research will use real collected data and will run survey to collect expert judgment for missing attributes?

In order to perform the probability calculation of underground gas pipe damages, it is necessary to collect the conditional probability distribution for each node in the Bayesian network. The collected data is sufficient to elaborate the evolution of natural gas pipeline network accidents; it will be used to build conditional probability table (CPT) of BN with historical data by parameter learning.

Table 1 shows the BN nodes of natural gas pipeline network damages and their classifications. The causal relationships between each node is determined based on

Comprehensive case studies of many typical natural gas pipeline network accidents and further evaluation by evaluating the collected data and analyzing different factors affecting the underground pipeline damage process. The Bayesian network of natural gas pipeline network accident is established.

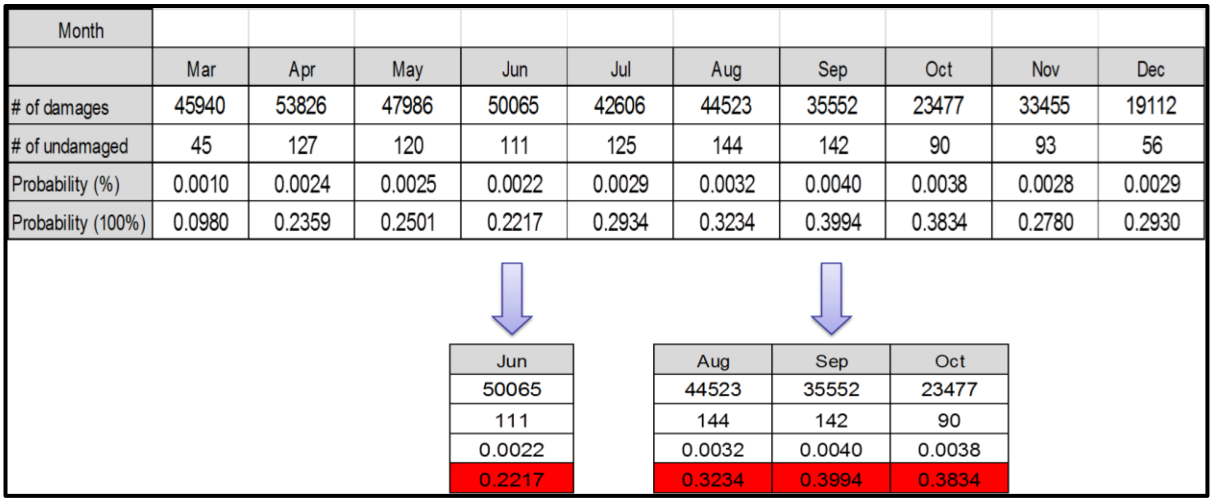
There are two types of nodes in Bayesian network. The first type does not have parent nodes such as “Gas Pipe Damage Cause”, and “Gas Pipe Damage Area In Urban Zones”. In addition, processes are perform using the received data to derive the probability of each of the parent nodes from the received data. Moreover, the network is divided into branches. Multiple approaches have been suggested to study and calculate the probability of the Bayesian Network. The data collected to derive the probability for each individual node is for years 2010, 2011, 2012, 2013, and 2014. More specifically, characteristics and data attributes will be selected to process the data. The first five years of collected data will be used in the current model, including probability for the nodes and Bayesian network path itself. In contrast, data from 2014 will be used later to test the developed model of Bayesian network for underground gas pipe damages. Analyzing the data is the process of systematically applying statistical and/or logical techniques to describe and illustrate, condense and recap, and evaluate data. According to Shamoo and Resnik (2003) various analytic procedures “provide a way of drawing inductive inferences from data and distinguishing the signal (the phenomenon of interest) from the noise (statistical fluctuations) present in the data. In this research a set of criteria was developed for processing the data, preparing the data, condensing the data, cross referencing the data. Thus, the goal of data analyses with the risk model was to discover useful information and support decision-making. Data analysis can be performed using diverse techniques under a variety of names, in different business, science, and social science domains. The probability is the measure of how likely the damage of UG gas pipe is to occur out of the number of possible received calls by one call center agency. Calculating probabilities of damage for UG gas pipes can seem complicated at first. However, once the data cleaned and prepared with certain attributes it become a matter of applying the calculation formula. In other words, the probability is the likelihood of UG gas pipe damage happening divided by the number of possible received calls by one call center. First, the data was organized by damaged and undamaged UG gas pipe data. Then the number of damaged data was calculated per hour starting from Hour (0) to Hour (23) figure 90. The probability was calculated by dividing the number of damaged UG gas pipe records per hour by the total number of the undamaged records of the UG gas pipe. Same steps were performed to calculate the probability for week days, months, years, Diameters (10, 20, 30) figure 91, 92.



Time										Probability of 24 hrs													
Hrs	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
# of damages	0	0	0	0	0	0	3	20	150	96	122	103	41	120	111	119	71	0	0	0	0	0	0	0
# of undamaged	423	298	257	207	244	817	5945	19740	30985	38866	42844	41272	35185	40578	43808	38610	26111	11467	5762	4323	3426	2753	1789	837
Probability (%	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.001	0.005	0.002	0.003	0.002	0.001	0.003	0.003	0.003	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Probability (1(	0.000	0.000	0.000	0.000	0.000	0.000	0.050	0.101	0.484	0.247	0.285	0.250	0.117	0.296	0.253	0.308	0.272	0.000	0.000	0.000	0.000	0.000	0.000	0.000

			Probability of Peak 8 -16 hrs					
Hrs	8	9	10	11	12	13	14	15	16
# of damages	150	96	122	103	41	120	111	119	71
# of undamaged	30985	38866	42844	41272	35185	40578	43808	38610	26111
Probability (%)	0.005	0.002	0.003	0.002	0.001	0.003	0.003	0.003	0.003
Probability (100%)	0.484	0.247	0.285	0.250	0.117	0.296	0.253	0.308	0.272

Figure 90: Calculate the probability of UG gas pipe damage for 24 hrs

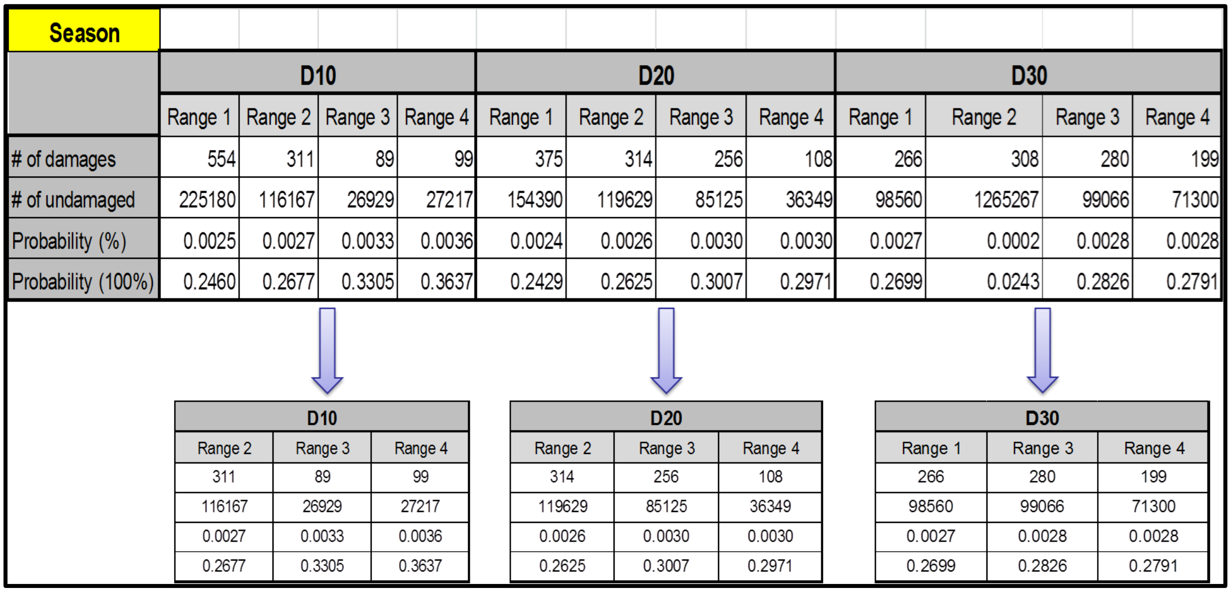


Month										
	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
# of damages	45940	53826	47986	50065	42606	44523	35552	23477	33455	19112
# of undamaged	45	127	120	111	125	144	142	90	93	56
Probability (%)	0.0010	0.0024	0.0025	0.0022	0.0029	0.0032	0.0040	0.0038	0.0028	0.0029
Probability (100%)	0.0980	0.2359	0.2501	0.2217	0.2934	0.3234	0.3994	0.3834	0.2780	0.2930

Jun
50065
111
0.0022
0.2217

Aug	Sep	Oct
44523	35552	23477
144	142	90
0.0032	0.0040	0.0038
0.3234	0.3994	0.3834

Figure 91: Calculate the probability of UG gas pipe damage for months



Season												
	D10				D20				D30			
	Range 1	Range 2	Range 3	Range 4	Range 1	Range 2	Range 3	Range 4	Range 1	Range 2	Range 3	Range 4
# of damages	554	311	89	99	375	314	256	108	266	308	280	199
# of undamaged	225180	116167	26929	27217	154390	119629	85125	36349	98560	1265267	99066	71300
Probability (%)	0.0025	0.0027	0.0033	0.0036	0.0024	0.0026	0.0030	0.0030	0.0027	0.0002	0.0028	0.0028
Probability (100%)	0.2460	0.2677	0.3305	0.3637	0.2429	0.2625	0.3007	0.2971	0.2699	0.0243	0.2826	0.2791

D10		
Range 2	Range 3	Range 4
311	89	99
116167	26929	27217
0.0027	0.0033	0.0036
0.2677	0.3305	0.3637

D20		
Range 2	Range 3	Range 4
314	256	108
119629	85125	36349
0.0026	0.0030	0.0030
0.2625	0.3007	0.2971

D30		
Range 1	Range 3	Range 4
266	280	199
98560	99066	71300
0.0027	0.0028	0.0028
0.2699	0.2826	0.2791

Figure 92:Calculate the probability of UG gas pipe damage for months

1.39 Building Bayesian Network / Model

The Bayesian Network (BN) composed of several nodes and directed edges. It reflects on the target analyses and represents cause and effect relationships of different nodes respectively. This network will include a probabilistic inference technology for reasoning under uncertainty by taking advantage of Probabilities Table of Bayesian Network nodes. Bayesian network was firstly presented by Pearl in 1985 (Pearl, 1985) and then has proven to be an effective cause-effect analysis tool. Bayesian Network represents uncertain knowledge in probabilistic systems. BN has been applied to a variety of safety assessment and risk analysis problems (Khakzad et al., 2011; Hossain and Muromachi, 2012; Francis et al., 2014; Tan et al., 2014; Kabir et al., 2015; Wu et al., 2016). Furthermore, based on data processing and comprehensive analyses of the selected attributes of underground gas line network, we propose 11 basic nodes (Figure 41). In addition, the relationship will be developed on which the evolution of natural gas pipeline network damages can be described explicitly. The investigation of UGPLD damages collected data guarantees the universality of Bayesian network. The description of each BN node for representing UGPLD damages is as follows:

a) Time (24 hrs)

b) Week days (Mon, Tue, Wed, Thurs, Fir, Sat, and Sun)

c) Months (Jan – Dec)

d) Year (2011- 2014)

e) Seasons (Winter, Summer, Spring, Autumn)

f) Damage within Diameter (D10, D20, D30)

Chart consists of 10 factors, were selected random to build the flow chart. The probability of the underground gas pipe being damaged under these 10 factors is 1.2% Scenario 1. The underground gas pipe has 1.2 % chance of been damage under these 10 factors figure 93. Chart 10 factors were randomly selected random to build the flow chart. The probability of the underground gas pipe been damage under these 10 factors is 0.025% scenario 2. The underground gas pipe has 0.025% chance of been damage under these 10 factors (Figure 94). The probability of the underground gas pipe been damage under these

10 factors is 0.15% scenario 3. The underground gas pipe has 0.15% chance of been damage under these 10 factors (Figure95).

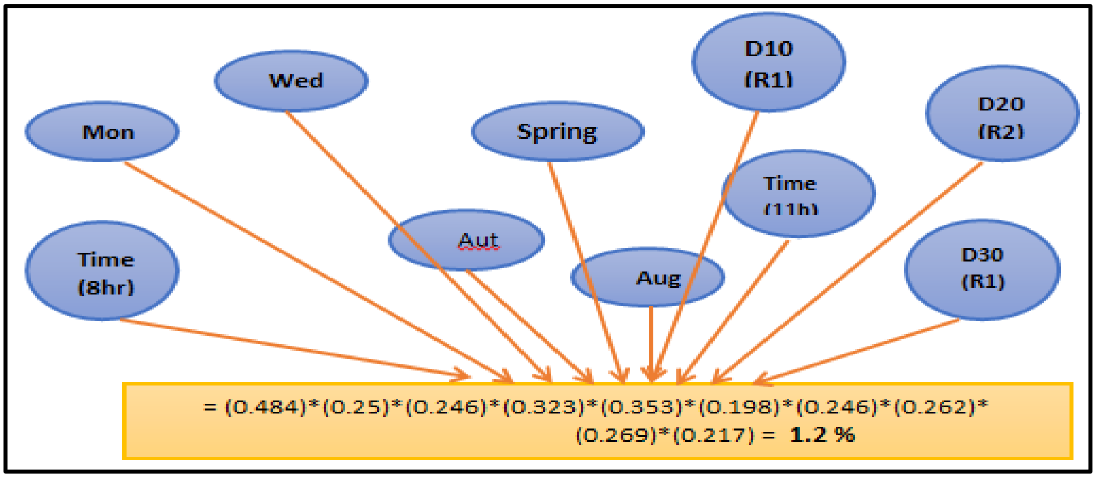


Figure 93 : Scenario 1 of developing Bayesian Model to predict the probability

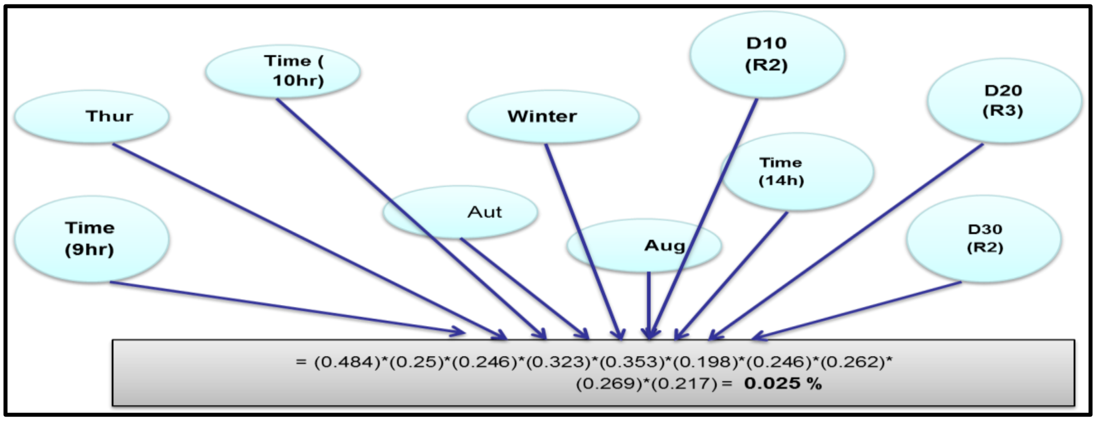


Figure 94:Scenario 2 of developing Bayesian Model to predict the probability

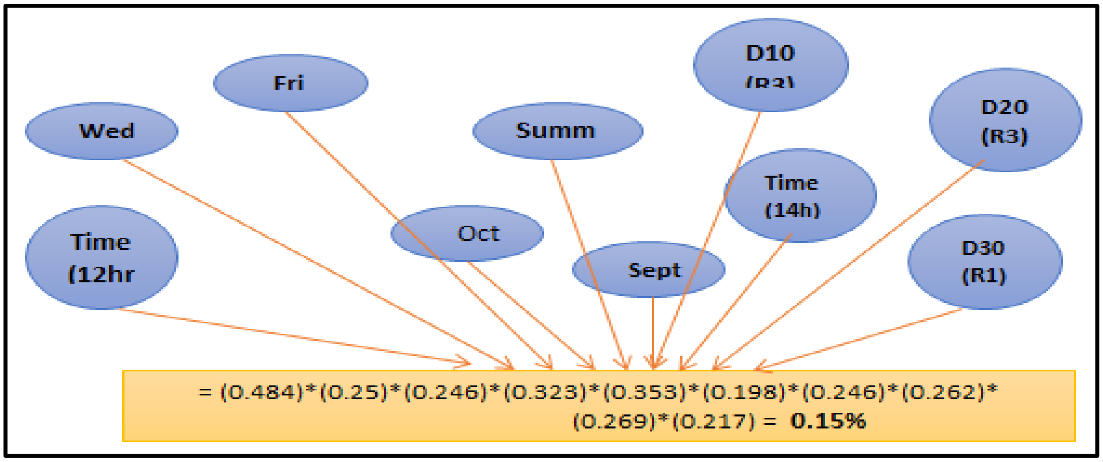


Figure 95: Scenario 3 of developing Bayesian Model to predict the probability

1.40 Results, and Discussion.

Time: First the time was divided into (AM, PM), and then divided to 24 hrs. Univariate and Bivariate analysis was performed. Findings, Timing attribute of the digging was determined to have significant effect on the UG gas pipe damage. Moreover, when all hours were analyzed as one attribute: P-Value was 0.001 which is less than the industry standard of 0.05. More specifically, 24hrs were analyzed separately; some hours have more damage percent than others. For instance, the damage percent within hour 8 represent 14.2% table 1 as compared to the rest of the 24 hrs, which is the highest percent of damage among all hours. In addition, the damage percent within hour 10, 13, and 15 represent 11.6%, 11.4%, 11.3% table 1 consecutively as compared to the rest of the 24 hrs. Therefore, the specific hours have significant impact on the UG gas pipe damage. In summary, the received UG gas pipe digging requests by the agency during 10, 13, and 15 have more chance of been damaged than the other hours. Thus, by avoiding digging when it's possible during these hours, UG gas pipe damage risk will be decreased.

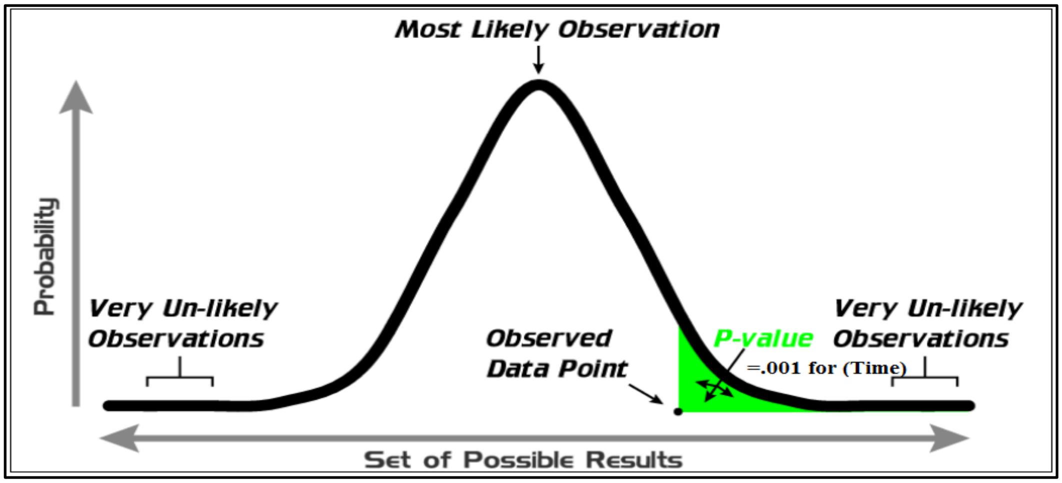


Figure 96: P Value for the attribute of (Time)

Week: First, the week as one attribute was determined to be significant. However, as observed from the UG gas pipe damaged data distribute, the damage was happening across all week days. Thus, Univaraite, and Bivariate statistics analysis was performed on each day separate to see the impact for each individual day on the UG gas pipe damage. Moreover, our analysis suggests that by avoiding digging operation on Saturday, that will significantly (p=0.045) decrease the damage of UG gas pipe damage. P value for the rest of the week days was more than 0.05 which means it does not have significant impact. In addition, odds ratio for Thursday, and Wednesday was 1.3. This means these two days have 1.3 chance of getting damaged as compared to rest of the week days. Any received digging requests during Thursday, and Wednesday will have 30% chance of been damage than the other days.

Months: The Bivariate analysis was conducted separate for all months. In order to find out which month is more significant in causing/contributing to underground gas pipe damage, P-Value was calculated for each month separate, starting from Jan, Feb, Mar, April, May, Jun, July, Aug, Sept, Oct, Nov, and Dec. Results, the month of March, June, Aug, Sept, and Oct have P-value less than 0.05 which means these months have significant impact on the UG gas pipe damage. In addition, by avoiding digging UG gas pipe during March, June, Aug, Sept, and Oct We may avoid UG pipe damage.

Seasons: this attribute was split into four seasons: Autumn, Spring, Summer, and Winter. As per the UG gas pipe data distribution, damages were happening across the four seasons. Statistics analysis was performed for each season separate. P-Value for autumn, and spring were less than 0.05, and for summer, and winter was more than 0.05. Therefore, digging during autumn, and spring could cause more damage to UG gas pipe than digging in summer, and winter. More detailed analysis included under Month Section above.

Counties, there are 21 counties included in the research. Statistical analysis was performed on all of the counties. For instance, County 5 has P-value equal to 0.001, and Odd ratio of 7.2. Which means county 5 has significant impact on the UG gas pipe damage, and any Gas pipe digging in county 5 has 7 times chance of been damage as compared to county 10 (reference county). All over all, 9 counties out of 21 counties total have P – value less than 0.05. As result, UG gas pipe digging in theses counties have significant impact on the gas pipe damage. This needed to be taken in consideration when digging request is received. In addition, by summarizing the odd ratios for all counties, and then classifying them in ascending order, we found the following; odds ratio in county # 5 equal to 7.3. This means there is approximately 7 times more chance for the gas pipe to be damage in county 5 as compared to the referenced county. In ascending order, for counties 20, 1, 4, 6 and 8, the odd ratios were 3.3, 3.2, 3, 3, 2.9, and 2.5 figure 97. This means any ticket requests received in these counties have around 3 times chance as compared to the referenced county.

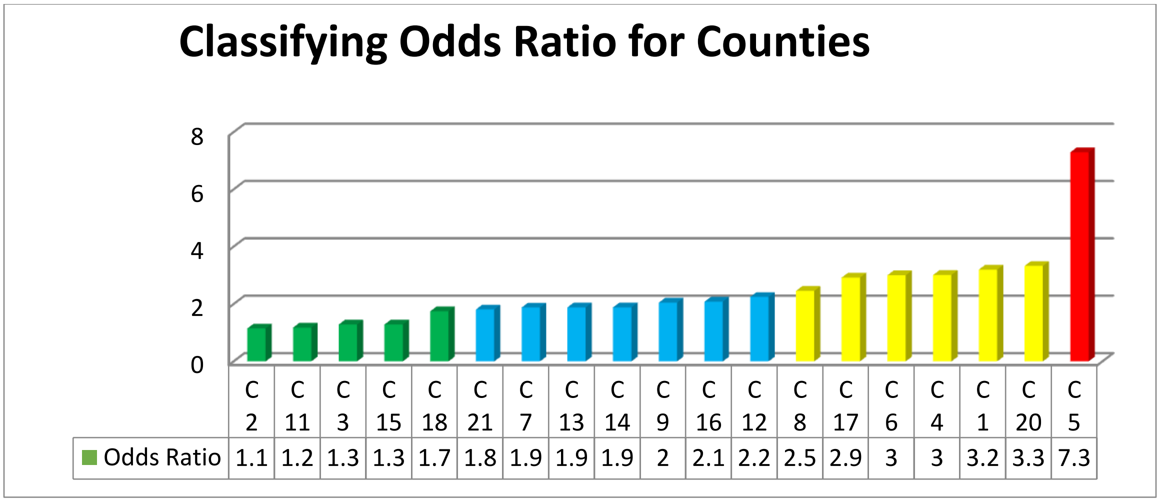


Figure 97:Odd Ratio ranking among the Counties

D10, D20, D30, attributes were split into four ranges divided equally in order to be able to perform the statistics analysis. Results, damages within 10 miles diameter have the highest impact as compared to D20, and D30. More specifically, P value for D10 was equal to 0.001 which is less than 0.005. In addition, D20 has P value of 0.04; D30 has P Value of 0.264. This means when the agency receive digging requests, they should look at the number of damages within 10 miles diameter, and 20 miles diameter to determine the chance of the UG gas pipe been damage. However, this study suggests that damages within 30 miles have no impact on the UG gas pipe damage. In addition, by analyzing the effect of ranges per diameter D10, D20, and D30 on the UG gas pipe damage, the percentage within damage was higher on Range 1, than Range 2. And percentage of damage was higher on Range 2 than Range 3. And percentage of damage in Range 3 was higher than Range 4.

CHAPTER SEVEN: CONCLUSION AND RECOMMENDATION

1.41 Summary and Conclusions

The significant number of damages of underground gas pipelines and their consequences have motivated many researchers to study UG pipeline damage. A comprehensive study of these research efforts revealed the lack of a comprehensive predictive model for estimating damages. Some studies focused on corrosion or third-party failures and could not assess the effects related to information flow on damages. Most studies that consider multiple damages focuses on qualitative investigation or develop physical models that are very expensive to implement. Qualitative models rely on survey which makes it difficult (if not impossible) to obtain due to the location of most UG pipelines. In addition to that, is the high cost of needed operators, the shortage of complete historical data on gas pipe has been a challenge for all researchers. This research addresses the lack of effective predicative models by developing a comprehensive risk model: for the processes involved in the information flow process starting from receiving digging request, till the completion of the excavation of gas pipelines. The model developed in this research provides an overall image throughout the digging processes of gas pipelines. This model is able to predict the probability of the underground gas pipe damage. Damage probability is calculated through Bayesian model based the derived important factors by machine learning model. A Bayesian risk model was developed to minimize pipeline damage rates through assessing and ranking the risk of various sections of natural gas pipelines. It would explore the interaction among significant factors. Inputs required for an effective evaluation of the risk encountered in exchange of information between different parties involved. In addition, the past data were used to develop a risk model to study future risk associated with the excavation requests and risk factors. This study also provides a research base by using Logistic Regression to develop risk model to investigate the interactive effects of various factors causing underground gas line damage. The regression also help predicting the probability of future damage occurrence. In addition, regression analysis help identifyingthe important factors can be used by digging agencies when they receive digging requests to minimize the possibility of gas pipe been damage. Basically the responsible agency can cross reference the request attribute with the developed risk factors; then they can make educated decision on where to be caution and pay more attention to the digging processes in the potential areas.

Risk involved in the UG gas pipe is a critical aid in the decision-making process of underground gas pipe system. The predictive model will be useful in assisting the operators of such facilities in the maintenance and inspection planning. The model can rank the selected tools based on their probabilities of happening. This research develops framework for the development of risk assessment models completely based on the historical damage of UG gas pipe data. The methodology is applied on the infrastructure of oil and gas pipelines. However, it can be expanded to be used in other infrastructure types. The main value of such a predicative models is that they reduce the cost of damage prediction with or without abundance of a valuable data. These models assess the probability of risk of different infrastructure types and to plan accordingly for the life cycle of such infrastructures.

1.42 Research contributions

- Provide the important risk factors and inputs required for an effective evaluation and assessment of the risk encountered in information exchange among different parties involved during the repair of underground gas pipelines. This can be used by responsible agencies to mitigate the future damages of UG gas pipelines.
- Provide a framework for developing risk predictive models for different infrastructure types of underground gas pipelines using historical data.
- An integrated risk assessment model to evaluate the risk level of a pipeline based on calculated probability.
- A probabilistic Bayesian based model to predict the probability of damage in underground gas pipelines.

1.43 Research Limitations

- The probability of damage prediction model does not consider the interdependency of basic events (the selection was random).
- The developed model still needs the users to enter the general attributes for each scenario type: based on the received call request by agency for the digging.
- The Bayesian does not consider the effect of safety barriers that can reduce the probability of gas pipeline damages.
- In the absence of required data on the identified attributes, the model could only be developed for UG gas pipes. It can be used for other infrastructure systems if data were to be collected from infrastructure agencies.

- The information flow process only proposes a processes of study based on the received data from the agency which is incomplete data.

1.44 Future Work and Recommendations

- Develop a dynamic 'age of damage' prediction model and consider it in the rehabilitation planning model. It should take into account the infrastructure age, type, and location.
- Consider the attributes of offshore pipelines, or water pipelines for applying the machine learning developed model by this study, and evaluate the results.
- Collect more data in UG gas pipelines order to consider the interdependency of the damage sources among each other, especially the materials type, pipe size, sources of damage. All that included could lead to more precision, and accuracy in the prediction process.
- Develop a Bayesian model to consider the attributes that were not available through the historical data and develop a network . The development of the probability of failure assessment model was limited to the availability of historical data.
- Evaluate the possible failure types in developing maintenance scenarios. This model might be able to predict the types of defects that can cause the failure of a pipeline and thus make it possible to plan for the maintenance accordingly.
- Develop a consequence of failure prediction model on non-common attributes of consequences, for example the amount of damage to the environment based on a damage

scenario. The prediction of the environmental effects of pipe damage needs more attributes, including the depth of the pipe, the pipe size, request time, and city.

References

1- Li, Shuai, Hubo Cai, and Vineet R. Kamat. "Uncertainty-aware geospatial system for mapping and visualizing underground utilities." Automation in Construction 53 (2015): 105-119.

2- Li, Shuai, et al. "Estimating features of underground utilities: Hybrid GPR/GPS approach." Journal of Computing in Civil Engineering 30.1 (2014): 04014108.

3- Al-Nuaimy, W., Huang, Y., Nakhkash, M., Fang, M. T. C., Nguyen, V. T.,and Eriksen, A. (2000). “Automatic detection of buried utilities and solid objects with GPR using neural networks and pattern recognition.” J. Appl. Geophys., 43(2), 157–165.

4- Jeong, H. S., Abraham, D. M., and Lew, J. J. (2004). “Evaluation of an emerging market in subsurface utility engineering.” J. Constr. Eng. Manage., 10.1061/(ASCE)0733-9364(2004)130:2(225), 225–234.

5- Li, Shuai, Hubo Cai, and Vineet R. Kamat. "Uncertainty-aware geospatial system for mapping and visualizing underground utilities." Automation in Construction 53 (2015): 105-119.

6- Jaw, Siow Wei, and Mazlan Hashim. "Locational accuracy of underground utility mapping using ground penetrating radar." Tunnelling and Underground Space Technology 35 (2013): 20-29.

7- Talmaki, Sanat, Vineet R. Kamat, and Hubo Cai. "Geometric modeling of geospatial data for visualization-assisted excavation." Advanced Engineering Informatics 27.2 (2013): 283-298.

8- Talmaki, Sanat, and Vineet R. Kamat. "Real-time hybrid virtuality for prevention of excavation related utility strikes." Journal of Computing in Civil Engineering 28.3 (2012): 04014001.

9- PHMSA. (2012a). “Significant pipeline incidents through 2010 by cause.” ⟨http://primis.phmsa.dot.gov/comm/reports/safety/SigPSIDet_2010 _2010_US.html? nocache=8248#all⟩ (Jan. 11, 2012).

10- Feng, C., et al. "Vision-based articulated machine pose estimation for excavation monitoring and guidance." ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction. Vol. 32. Vilnius Gediminas Technical University, Department of Construction Economics & Property, 2015.

11- Azar, Ehsan Rezazadeh, Chen Feng, and Vineet R. Kamat. "Feasibility of in-plane articulation monitoring of excavator arm using planar marker tracking." Journal of Information Technology in Construction (ITcon) 20.15 (2015): 213-229.

12- Bernold L.E. (2002). Spatial integration in construction, Journal of Construction Engineering and Management, Vol. 128, No.5, 400–408.

13- Bernold, Leonhard E. "Control schemes for tele-robotic pipe installation." Automation in Construction 16.4 (2007): 518-524.

14- Stentz A., Bares J., Singh S. and Rowe P. (1999). A robotic excavator for autonomous truck loading, Autonomous Robots, Vol. 7, No. 2, 175-186.

15- Chiang M.H. and Huang C.C. (2004). Experimental implementation of complex path tracking control for large robotic hydraulic excavators, The International Journal of Advanced Manufacturing Technology, Vol. 23, No. 1-2, 126-132.

16- Bai, L., Liang, J., Sui, C., & Dang, C. (2013). Fast global k-means clustering based on local geometrical information. Information Sciences, 245(0), 168-180.

doi:http://dx.doi.org.proxy.libraries.rutgers.edu/10.1016/j.ins.2013.05.023

17- Yamamoto H., Moteki M., Shao H., Ootuki T., Kanazawa H. and Tanaka Y. (2009). Basic technology toward autonomous hydraulic excavator, in 26th International Symposium on Automation and Robotics in Construction (ISARC 2009) in Austin, USA, 288-295.

18- Talmaki, Sanat, Vineet R. Kamat, and Hubo Cai. "Geometric modeling of geospatial data for visualization-assisted excavation." Advanced Engineering Informatics 27.2 (2013): 283-298.

19- Baksh, Al-Amin, et al. "Network based approach for predictive accident modelling." Safety science 80 (2015): 274-287.

20- Cagno, E., Caron, F., Mancini, M., Ruggeri, F., 2000. Using AHP in determining the prior distributions on gas pipeline failures in a robust Bayesian approach. Reliab. Eng. Syst. Saf. 67, 275e284.

21- Yuhua, Dong, and Yu Datao. "Estimation of failure probability of oil and gas transmission pipelines by fuzzy fault tree analysis." Journal of loss prevention in the process industries 18.2 (2005): 83-88.

22- Markowski, Adam S., and M. Sam Mannan. "Fuzzy logic for piping risk assessment (pfLOPA)." Journal of loss prevention in the process industries 22.6 (2009): 921-927.

23- Han, Z. Y., and W. G. Weng. "Comparison study on qualitative and quantitative risk assessment methods for urban natural gas pipeline network." Journal of Hazardous Materials 189.1 (2011): 509-518.

24- Guo, Yanbao, et al. "Comprehensive risk evaluation of long-distance oil and gas transportation pipelines using a fuzzy Petri net model." Journal of Natural Gas Science and Engineering 33 (2016): 18-29.

25- Li, Xinhong, Guoming Chen, and Hongwei Zhu. "Quantitative risk analysis on leakage failure of submarine oil and gas pipelines using Bayesian network." Process Safety and Environmental Protection 103 (2016): 163-173.

26- Kabir, Golam, Rehan Sadiq, and Solomon Tesfamariam. "A fuzzy Bayesian belief network for safety assessment of oil and gas pipelines." Structure and Infrastructure Engineering 12.8 (2016): 874-889.

27- DIRT Report www.commongroundalliance.com • www.cga-dirt.com

28- Chapter 2. UNDERGROUND FACILITIES: ONE-CALL Damage Preventions by BU.

29- PHMSA Inspector Training and Qualifications http://www.phmsa.dot.gov/pipeline/tq

30- PHMSA Pipeline Safety Regulation http://www.phmsa.dot.gov/pipeline/tq/regs

31- PHMSA Inspector Enforcement Guidance http://www.phmsa.dot.gov/foia/e-reading-room

32- PHMSA Website:http://www.phmsa.dot.gov

33- One Call Center in The State of New jersey http://www.nj1-call.org/

34- CGA | DIRT Overview. (2005). Retrieved December 8, 2014, from http://www.commongroundalliance.com/Template.cfm?Section=DIRT_Overview&Template=/TaggedPage/TaggedPageDisplay.cfm&TPLID=39&ContentID=2206

35- Common Ground Alliance. (2014). DIRT Annual Report for 2013. Retrieved from http://www.cga-dirt.com/annual/2013/DIRT_Report_for_2013_Final_20141008_REDUCED.pdf

36- Eynard, D. (Professor) (2012). Pattern Analysis and Machine Learning Intelligence. Lecture conducted from Milan.

37- Gentile, K. (Technical Services Manager) (2013, December 17). Public Awareness. New Jersey Pipeline Safety Seminar. Lecture conducted by Pipeline and Hazardous Materials Safety Administration, Edison, NJ.

38- Gong, J. (Assistant Professor) (2014, October 30). CE546 Advanced Construction Management II: Clustering (Part II). Lecture conducted from Piscataway, NJ.

39- Karypis, G., Han, E.H., & Kumar, V. (August 1991). CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling. IEEE Computer, 32(8):68-75

40- Roda, A. M. (2012). 2009-2011 New Jersey Gas Line Damage Report. (No. CAIT-RU433986). Piscataway, NJ: Center for Advanced Infrastructure & Transportation.

41- Safe Digging. (2014). Retrieved December 8, 2014, from http://www.call811.com/how-811-works/safe-digging.aspx

42- Sanchez, R. (2014). New York explosion exposes nation's aging and dangerous gas mains. Retrieved from http://www.cnn.com/2014/03/15/us/aging-gas-infrastructure/

43- Tan, P., Steinbach, M., & Kumar, V. (2006). Introduction to Data Mining (First ed.). Boston: Pearson Addison Wesley.

44- West, P. (Pipeline Safety Specialist) (2013). PHMSA: Organization and Regulatory Overview. Lecture conducted by Pipeline and Hazardous Materials Safety Administration.

45- Alsabti, K., Ranka, S., & Singh, V. (1997). An efficient k-means clustering algorithm. Electical Engineering and Computer Science. Retrieved from http://surface.syr.edu/eecs/43/

46- Anselin, Luc. "Local Indicators of Spatial Association—LISA," Geographical Analysis 27(2): 93–115, 1995.

47- Anselin, L., Bera, A. K., Florax, R., & Yoon, M. J. (1996). Simple diagnostic tests for spatial dependence. Regional Science and Urban Economics, 26(1), 77–104. doi:10.1016/0166-0462(95)02111-6

48- Commissioners, B. O. F., & Prevention, D. (n.d.). Underground Facilities : One-Call Damage Prevention System, 1–11.

49- Genolini, C., & Falissard, B. (n.d.). A package to cluster longitudal data. Computer Methods and Programs in Biomedicine, 2011

50- Genovese, C. R., Lazar, N. A., & Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuroimage, 15(4), 870–878. doi:10.1006/nimg.2001.1037\rS1053811901910377 [pii]

51- Getis, A., & Ord, J. K. (1992). The Analysis of Spatial Association. Geographical Analysis, 24(3), 189–206. doi:10.1111/j.1538-4632.1992.tb00261.x

52- Likas, A., Vlassis, N., & J. Verbeek, J. (2003). The global k-means clustering algorithm. Pattern Recognition, 36(2), 451–461. doi:10.1016/S0031-3203(02)00060-2

53- Ord, J. K., & Getis, A. (2010). Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. Geographical Analysis, 27(4), 286–306. doi:10.1111/j.1538-4632.1995.tb00912.

54- Wagstaff, K., Cardie, C., Rogers, S., & Schroedl, S. (2001). Constrained K-means Clustering with Background Knowledge. International Conference on Machine Learning, 577–584. doi:10.1109/TPAMI.2002.1017616

55- Weisstein, E. W. (n.d.). Least Squares Fitting. Wolfram Research, Inc. Retrieved from http://mathworld.wolfram.com/LeastSquaresFitting.html

56- Zhao, H., Shang, Z., Tang, Y. Y., & Fang, B. (2013). Multi-focus image fusion based on the neighbor distance. Pattern Recognition, 46(3), 1002–1011. doi:10.1016/j.patcog.2012.09.012

57- How Grouping Analysis works—ArcGIS Pro | ArcGIS for Desktop. (n.d.). Retrieved March 2, 2016, from http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/how-grouping-analysis-works.htm

58- How Cluster and Outlier Analysis (Anselin Local Moran's I) works—ArcGIS Pro ArcGIS for Desktop. (n.d.). Retrieved March 2, 2016, from http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/h-how-cluster-and-outlier-analysis-anselin-local-m.htm

59- How Hot Spot Analysis (Getis-Ord Gi*) works—Help | ArcGIS for Desktop. (n.d.). Retrieved March 2, 2016, from http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-statistics-toolbox/h-how-hot-spot-analysis-getis-ord-gi-spatial-stati.htm

60- New Jersey One Call - Home. (n.d.). Retrieved March 2, 2016, from http://www.nj1-call.org/

61- ArcGIS - FEMA MOTF Hurricane Sandy Impact Analysis. (n.d.). Retrieved March 2, 2016, from https://www.arcgis.com/home/item.html?id=307dd522499d4a44a33d7296a5da5ea0

62- New Jersey Geographic Information Network. (n.d.). Retrieved March 2, 2016, from https://njgin.state.nj.us/NJ_NJGINExplorer/index.jsp

63- Modeling spatial relationships—Help | ArcGIS for Desktop. (n.d.). Retrieved April 9, 2016, from http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-statistics-toolbox/modeling-spatial-relationships.htm#GUID-729B3B01-6911-41E9-AA99-8A4CF74EEE27

64- Everitt, Brian S., Landau, Sabine, Leese, Morven, Stahl, D. (2011). Cluster Analysis, 65–124. http://doi.org/10.1007/BF00154794

65- global perspective and focus on technology innovations and market impact, GTI http://www.gastechnology.org/Pages/default.aspx

66- Zhao, Yafan, and Mingda Song. "Failure analysis of a natural gas pipeline." Engineering Failure Analysis 63 (2016): 61-71.

67- Mohsin, R., Z. A. Majid, and F. L. Tan. "Numerical analysis of wall shear patterns on the external wall of an API 5L X42 natural gas pipe." Engineering Failure Analysis 48 (2015): 30-40.

68- American Gas Association (AGA) 2005, "Natural Gas Industry Safety Programs",http://www.aga.org/Kc/aboutnaturalgas/additional/NGSafetyPrograms.htm

69- American Gas Association (AGA) 2010, "What causes natural gas pipeline accidents?"http://www.aga.org/Kc/aboutnaturalgas/consumerinfo/CausesofNGPipelineAccidents.htm

70- Belson, K., DePalma, A (2007), "Asbestos and Aging Pipes Remain Buried Hazards", NYTimes, July 19, 2007,

71- http://www.nytimes.com/2007/07/19/nyregion/19steam.html

Call811 (2010), "Frequently Asked Questions", http://www.call811.com/default.aspx

72- Positioning Buried Utilities using an integrated GNSS approach https://www.researchgate.net/publication/257527875

73- Makowski, A. S., and S. M. Mannan. "Fuzzy logic for piping risk assessment." Journal of Loss Prevention in the Process Industries 22.6 (2009): 921-927.

74- Jamshidi, Ali, et al. "Developing a new fuzzy inference system for pipeline risk assessment." Journal of loss prevention in the process industries 26.1 (2013): 197-208.

75- Wu, Jiansong, et al. "Probabilistic analysis of natural gas pipeline network accident based on Bayesian network." Journal of Loss Prevention in the Process Industries 46 (2017): 126-136.

76- Yang, Ya Ping, Yong Mei Hao, and Zhi Xiang Xing. "Multi-factor and polymorphism failure probability calculation for natural gas pipelines." Applied Mechanics and Materials. Vol. 401. Trans Tech Publications, 2013.

77- Sørup, Hjalte Jomo Danielsen, et al. "Sustainable flood risk management–What is sustainable?." Proceedings of 9th International Conference on Planning and Technologies for Sustainable Urban Water Management. 2016.

78- Hochschulkolleg, Chinesisch-Deutsches. "Integrated Analyses and Assessment of Operational Risk: An Influence Diagrams Approach Based on Topological Data Model. 2014."

79- Zhang, Nevin L., and David Poole. "A simple approach to Bayesian network computations." Proc. of the Tenth Canadian Conference on Artificial Intelligence. 1994.

80- Berardi, Luigi, et al. "Development of pipe deterioration models for water distribution systems using EPR." Journal of Hydroinformatics 10.2 (2008): 113-126.

81- Torres-Arredondo, Miguel Angel, et al. "Damage detection and classification in pipework using acousto-ultrasonics and non-linear data-driven modelling." Journal of Civil Structural Health Monitoring 3.4 (2013): 297-306.

82- Ying, Yujie. A data-driven framework for ultrasonic structural health monitoring of pipes. Diss. Carnegie Mellon University, 2012.

83- Janes, Kevin A., and Michael B. Yaffe. "Data-driven modelling of signal-transduction networks." Nature reviews. Molecular cell biology 7.11 (2006): 820.

84- Marsh, Julie A., John F. Pane, and Laura S. Hamilton. "Making sense of data-driven decision making in education." (2006).

85- Kelangath, Subin, et al. "Risk analysis of damaged ships–a data-driven Bayesian approach." Ships and Offshore Structures 7.3 (2012): 333-347.

86- Yin, Shen, et al. "A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process." Journal of Process Control 22.9 (2012): 1567-1581.

87- Delessert, Christian, et al. "Spatial and temporal analysis of the local response to wounding." Plant molecular biology 55.2 (2004): 165-181.

88- Biswal, Bharat, et al. "Functional connectivity in the motor cortex of resting human brain using echo-planar mri." Magnetic resonance in medicine 34.4 (1995): 537-541.

89- Bishop, Christopher M. Pattern recognition and machine learning. springer, 2006.

90- Ringnér, Markus. "What is principal component analysis?." Nature biotechnology 26.3 (2008): 303-304.

91- Tipping, Michael E., and Christopher M. Bishop. "Probabilistic principal component analysis." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61.3 (1999): 611-622.

92- Delessert, Christian, et al. "Spatial and temporal analysis of the local response to wounding." Plant molecular biology 55.2 (2004): 165-181.

93- Rozenfeld, Ophir, Rafael Sacks, and Yehiel Rosenfeld. "'CHASTE': construction hazard assessment with spatial and temporal exposure." Construction Management and Economics 27.7 (2009): 625-638.

94- Feyer, A.-M., and Williamson, A. M. 2006. "Chapter 56: Human factors in accident modelling." ILO encyclopaedia of occupational healthand safety, International Labour Organization, Geneva.

95- Winter, Robert, and Bernhard Strauch. "A method for demand-driven information requirements analysis in data warehousing projects." System Sciences, 2003. Proceedings of the 36th Annual Hawaii International Conference on. IEEE, 2003.

96- Nielsen, Elizabeth, DeAnn Hyder, and Chao Deng. "A data-driven approach to risk-based source data verification." Therapeutic Innovation & Regulatory Science 48.2 (2014): 173-180.

97- Markowski, Adam S., M. Sam Mannan, and Agata Bigoszewska. "Fuzzy logic for process safety analysis." Journal of loss prevention in the process industries 22.6 (2009): 695-702.

98- Duijm, Nijs Jan. "Safety-barrier diagrams as a safety management tool." Reliability Engineering & System Safety 94.2 (2009): 332-341.

99- De Dianous, Valérie, and Cécile Fiévez. "ARAMIS project: A more explicit demonstration of risk control through the use of bow–tie diagrams and the evaluation of safety barrier performance." Journal of Hazardous Materials 130.3 (2006): 220-233.

100- Shahriar, Anjuman, Rehan Sadiq, and Solomon Tesfamariam. "Risk analysis for oil & gas pipelines: A sustainability assessment approach using fuzzy based bow-tie analysis." Journal of Loss Prevention in the Process Industries 25.3 (2012): 505-523.

101- Reason, J. T., J. Carthey, and M. R. De Leval. "Diagnosing "vulnerable system syndrome": an essential prerequisite to effective risk management." Quality and safety in health care 10.suppl 2 (2001): ii21-ii25.

102- Khakzad, Nima, Faisal Khan, and Paul Amyotte. "Dynamic risk analysis using bow-tie approach." Reliability Engineering & System Safety 104 (2012): 36-44.

103- 103-Markowski, Adam S., M. Sam Mannan, and Agata Bigoszewska. "Fuzzy logic for process safety analysis." Journal of loss prevention in the process industries 22.6 (2009):

104- 104-Jacinto, Celeste, and Cristina Silva. "A semi-quantitative assessment of occupational risks using bow-tie representation." Safety Science 48.8 (2010): 973-979.

105- Hyun, Ki-Chang, et al. "Risk analysis using fault-tree analysis (FTA) and analytic hierarchy process (AHP) applicable to shield TBM tunnels." Tunnelling and Underground Space Technology 49 (2015): 121-129.

106- Dawotola, Alex W., P. H. A. J. M. Van Gelder, and J. K. Vrijling. "Risk Assessment of Petroleum Pipelines using a com-bined Analytical Hierarchy Process-Fault Tree Analysis (AHP-FTA)." Proceedings of the 7th international probabilistic workshop, Delft. 2009.

107- Chang, C. L. The Research of Risk Assessment Model for Information Security of e-Business by Fault Tree Analysis. Diss. Master Thesis, Huafan University, 2007.

108- Geum, Y., H. Seol, S. Lee, and Y. Park. 2009. “Application of Fault Tree Analysis to the Service Process: Service Tree Analysis Approach.” Journal of Service Management 20 (4): 433–454.

109- Lindhe, A., L. Rosén, T. Norberg, and O. Bergstedt. 2009. “Fault Tree Analysis for Integrated and Probabilistic Risk Analysis of Drinking Water Systems.” Water Research 43 (6): 1641–1653. http://www.sciencedirect.com/science/article/B6V73-4V94WX0-9/2/d545bb7892f81ecf53d77c2754a41012

110- Ortmeier, F., and G. Schellhorn. 2007. “Formal Fault Tree Analysis – Practical Experiences.” Electronic Notes in Theoretical Computer Science 185:

139–151. http://www.sciencedirect.com/science/article/B75H1-4P40BN4-C/2/f1f3349832e9c565a1404b975f03df92

111- Park, A. and S. J. Lee. 2009. “Fault Tree Analysis on Handwashing for Hygiene Management.” Food Control 20 (3): 223–229. http://www.sciencedirect.com/science/article/B6T6S-4SGKB9R-1/2/e2d664678e98f25d3f1befbc184be9ae

112- Shua, M.-H., C.-H. Cheng, and J.-R. Chang. 2006. “Using Intuitionistic Fuzzy Sets for Fault-tree Analysis on Printed Circuit Board Assembly.” Microelectronics Reliability 46 (12): 2139–2148. http://www.sciencedirect.com/science/article/pii/ S0026271406000217

113- Yuhua, D., and Y. Datao. 2005. “Estimation of Failure Probability of Oil and Gas Transmission Pipelines by Fuzzy Fault Tree Analysis.” Loss Prevention in the Process Industries 18: 83–88. http://www.sciencedirect.com/science/article/pii/S0950423005000148

114- Flage, Roger, et al. "Probability and Possibility-Based Representations of Uncertainty in Fault Tree Analysis." Risk analysis 33.1 (2013): 121-133.

115- Ruijters, Enno, and Mariëlle Stoelinga. "Fault tree analysis: A survey of the state-of-the-art in modeling, analysis and tools." Computer science review 15 (2015): 29-62.

116- O. Coudert, J.C. Madre, MetaPrime: An interactive fault-tree analyzer, IEEE Trans. Rel. 43 (1994) 121–127. http://dx.doi.org/10.1109/24.285125.

117- O. Coudert, J.C. Madre, Fault tree analysis: 1020 Prime implicants and beyond, in: Proc. Reliability and Maintainability Symposium, RAMS, 1993, pp. 240–245. http://dx.doi.org/10. 1109/RAMS.1993.296849

118- Doytchev, Doytchin E., and Gerd Szwillus. "Combining task analysis and fault tree analysis for accident and incident analysis: a case study from Bulgaria." Accident Analysis & Prevention 41.6 (2009): 1172-1179.

119- K. Stecher, Evaluation of large fault-trees with repeated events using an efficient bottom-up algorithm, IEEE Trans. Rel. 35 (1986) 51–58. http://dx.doi.org/10.1109/TR. 1986.4335344.

120- Y. Dutuit, A.B. Rauzy, Efficient algorithms to assess component and gate importance in fault tree analysis, Reliab. Eng. Syst. Safety 72 (2) (2001) 213–222. http://dx.doi. org/10.1016/S0951-8320 (01)00004-7.

121- H. Boudali, P. Crouzen, M. Stoelinga, Dynamic fault tree analysis using input/output interactive Markov chains, in: Proc. 37th Int. Conf. Dependable Systems and Networks,DSN, IEEE, 2007, pp. 708–717. http://dx.doi.org/10.1109/DSN.2007.37.

122- Ericson, Clifton A. "Fault tree analysis." System Safety Conference, Orlando, Florida. 1999.

123- Chen, L., Wang, G. L., & Meng, H. R. (1995). Fault tree analysis of an oilwell pump. Acta Petroleum Sinica, 16(3), 145–151.

124- Liao, K. X., Yao, A. L., & Zhang, H. X. (2001). Fault tree analysis of pipelines. Oil and Gas Transportation, 20(1), 27–30.

125- Lin, C. T., & Wang, M. J. J. (1997). Hybrid fault tree analysis using fuzzy sets. Reliability Engineering and System Safety, 58, 205–213.

126- Liu, Q. Y., & Chen, H. (1999). Failure analysis of casing. Journal of Southwest Petroleum Institute, 21(4), 75–77.

127- Khakzad, Nima, Faisal Khan, and Paul Amyotte. "Safety analysis in process facilities: Comparison of fault tree and Bayesian network approaches." Reliability Engineering & System Safety 96.8 (2011): 925-932.

128- Andrews, John D., and Thomas Robert Moss. Reliability and risk assessment. Wiley-Blackwell, 2002.

129- Addo, Peter Martey, Dominique Guegan, and Bertrand Hassani. "Credit Risk Analysis Using Machine and Deep Learning Models." Risks 6.2 (2018): 38.

130- Choi, Joon Yul, et al. "Multi-categorical deep learning neural network to classify retinal images: A pilot study employing small database." PloS one 12.11 (2017): e0187336.

131- Fritz, Bradley A., et al. "Using machine learning techniques to develop forecasting algorithms for postoperative complications: protocol for a retrospective study." BMJ open8.4 (2018): e020124.

132- Azahara “Make data count; predict the future with machine learning” (2016)

133- JONATHAN VANIAN “4 Things Everyone Should Fear About Artificial Intelligence and the Future” (2018)

134- Huerta, Ramon, Fernando Corbacho, and Charles Elkan. "Nonlinear support vector machines can systematically identify stocks with high and low future returns." Algorithmic Finance 2.1 (2013): 45-58.

135- Isa, Dino, Rajprasad Rajkumar, and Ko-Choong Woo. "Pipeline Defect Detection Using Support Vector Machine." 6th WSEAS International conference on circuits, systems, electronics, control and signal processing, Cairo, Egypy. 2007.

136- Yu, Wei, et al. "Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes." BMC medical informatics and decision making 10.1 (2010): 16.

137- Isa, Dino, and Rajprasad Rajkumar. "Pipeline defect prediction using support vector machines." Applied Artificial Intelligence 23.8 (2009): 758-771.

138- Imandoust, Sadegh Bafandeh, and Mohammad Bolandraftar. "Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background." International Journal of Engineering Research and Applications 3.5 (2013): 605-610.

139- Yang, Jingli, Zhen Sun, and Yinsheng Chen. "Fault detection using the clustering-kNN rule for gas sensor arrays." Sensors 16.12 (2016): 2069.

140- Shetty, Mrs Lathika J., and Ms Shetty Mamatha Gopal. "DEVELOPING PREDICTION MODEL FOR STOCK EXCHANGE DATA SET USING HADOOP MAP REDUCE TECHNIQUE." (2016).

141- Vitola, Jaime, et al. "A sensor data fusion system based on k-nearest neighbor pattern classification for structural health monitoring applications." Sensors 17.2 (2017): 417.

142- de Castro, Pedro Afonso Paulino Ferreira. "Exploring the use of learning management systems data to early predict students' academic performance." (2018).

143- Waljee, Akbar K., Peter DR Higgins, and Amit G. Singal. "A primer on predictive models." Clinical and translational gastroenterology 5.1 (2014): e44.

144- Khaidem, Luckyson, Snehanshu Saha, and Sudeepa Roy Dey. "Predicting the direction of stock market prices using random forest." arXiv preprint arXiv:1605.00003 (2016).

145- Shaikhina, Torgyn, et al. "Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation." Biomedical Signal Processing and Control (2017).

146- Gurm, Hitinder S., et al. "A random forest based risk model for reliable and accurate prediction of receipt of transfusion in patients undergoing percutaneous coronary intervention." PloS one 9.5 (2014): e96385.

147- Mi, Chunrong, et al. "Why choose Random Forest to predict rare species distribution with few samples in large undersampled areas? Three Asian crane species models provide supporting evidence." PeerJ 5 (2017): e2849.

148- Lebedev, A. V., et al. "Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness." NeuroImage: Clinical 6 (2014): 115-125.

149- Restrepo, Carlos E., Jeffrey S. Simonoff, and Rae Zimmerman. "Causes, cost consequences, and risk implications of accidents in US hazardous liquid pipeline infrastructure." International Journal of Critical Infrastructure Protection 2.1-2 (2009): 38-50.

150- Sperandei, Sandro. "Understanding logistic regression analysis." Biochemia medica: Biochemia medica 24.1 (2014): 12-18.

151- Ioannou, I., T. Rossetto, and D. N. Grant. "Use of regression analysis for the construction of empirical fragility curves." Proceedings of the 15th world conference on earthquake engineering, September. 2012.

152- Azahara "Make data count; predict the future with machine learning" (2016) https://geographica.gs/en/blog/machine-learning/

153- JONATHAN VANIAN "4 Things Everyone Should Fear About Artificial Intelligence and the Future" (2018) https://finance.yahoo.com/news/4-things-everyone-fear-artificial-222340369.html

154- Stephen Hawking "The Usefulness—and Possible Dangers—of Machine Learning" (2017)

155- Kuusisto, Finn, et al. "Support vector machines for differential prediction." Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Berlin, Heidelberg, 2014.

156- Nigsch, Florian, et al. "Melting point prediction employing k-nearest neighbor algorithms and genetic parameter optimization." Journal of chemical information and modeling 46.6 (2006): 2412-2422.

157- Zhang, L., and R. A. Adey. "Predicting the Probability of failure of gas pipelines including Inspection and repair procedures." International Conference on Hydrogen Safety. Vol. 2007. 2007.INCLUDING INSPECTION AND REPAIR PROCEDURES: (2014)

158- Guazzelli, Alex, Wen-Ching Lin, and Tridivesh Jena. PMML in action: unleashing the power of open standards for data mining and predictive analytics. CreateSpace, 2012

159- Lam, Chio, and Wenxing Zhou. "Statistical analyses of incidents on onshore gas transmission pipelines based on PHMSA database." International Journal of Pressure Vessels and Piping 145 (2016): 29-40.