**Final Project Report: CSCI 8715**

**Tentative Title:** Using Spatial Data Science Approaches to Estimate and Predict Pipeline Failure Risk.

**Submitted by:** Bhupinder S Juneja (junej003), Sagar Tamang (taman011)

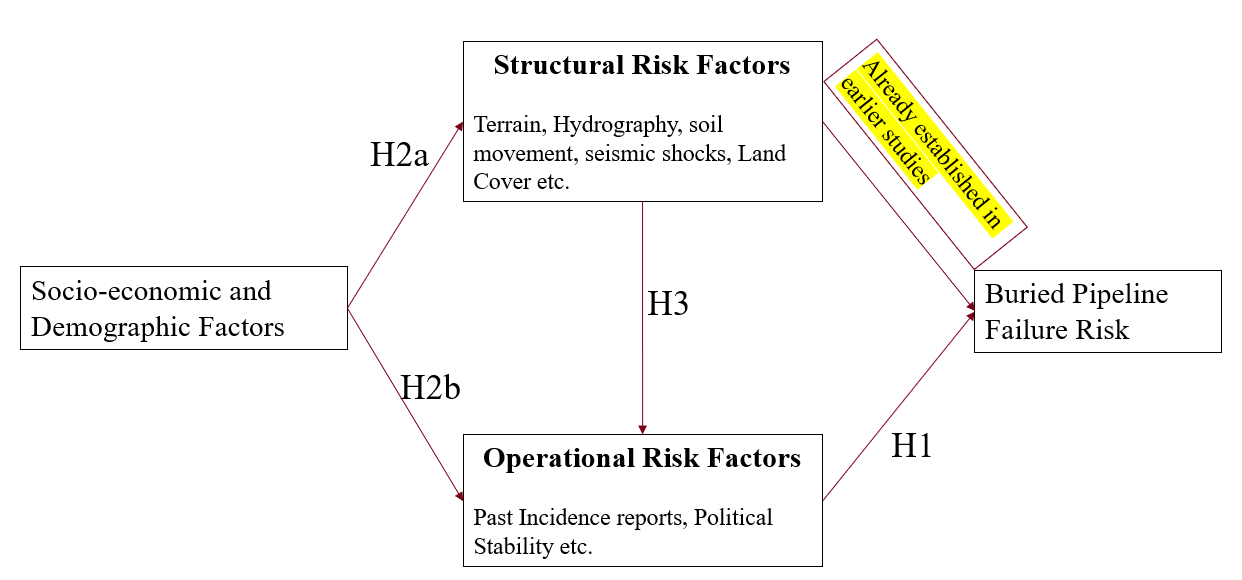
**Abstract:** Pipeline safety is vital because pipeline incidents threaten human health, biological life, and the environment. One single pipeline incident can result in a significant consequence. For instance, The Kalamazoo river oil spill that occurred in 2010 is one of the largest spills in American Industry. The oil spill caused by ruptured buried pipeline released 800,000 gallons of crude oil into wetlands near Marshall, Mich., and later into the Kalamazoo River. It took 6 years to complete the cleanup and the cleanup expense was estimated at 1.21 billion dollars. In addition, it resulted in considerable environmental damage [7]. While the pipeline industry faces considerable pressure of monitoring, scrutiny, and sanction and the stakeholders (e.g., pipeline operators, Department of Transportation) have increased investment in complying standards, preventing pipeline incidents is still very challenging because it is difficult to predict the risk of the pipeline system. First, the total pipeline mileage in the United States is more than 2 million miles and pipeline incidents are low-probability high-consequence events. Second, traditional statistical approaches cannot deal with high dimensional data from geographical characteristics (e.g. hydrography, waterway, land cover, pipeline network shape). In this respect, our study is focused on developing and validating the performance of predictive models of future pipeline risk severity and likelihood by analyzing multiple raster images created from geographical information (e.g., hydrography, population density, protected areas, and, commercially navigable waterway).

1. **Introduction:**

Pipeline including buried pipelines are subject to multiple stresses including but not limited to potentially unstable soil movements, seismic shocks, terrain topography, hydrography, population characteristics etc. These stresses can be broadly classified as Structural and Operational stresses. Manual monitoring of deformations is expensive and time-consuming intervention. The situation is further aggravated by general mindset which has been of maintaining status quo “Out of Sight, Out of Mind, Until It Leaks”. Given the complex nature of problem, the multiple contributors of stresses need to be investigated together vis-à-vis their interaction with pipelines in order to predict potential hotspots of failure and estimate associated risk.

Existing approaches are limited in a sense that they take a compartmentalized view of problem, they focus on limited set of variables belonging to structural or operational categories [8,9]. In addition, the traditional statistical or Machine Learning approaches are not well suited to deal with nuances of high dimensional data comprising of spatial and temporal characteristics of geographical data including hydrography, waterway, land cover, pipeline network shape etc. [3,10,11,12]. Most approaches treat prediction of failure hotspots and associated estimation of downside risk to rank pipelines according to their failure likelihood as two separate and independent problems when they are interlinked.Given vast pipeline network spanning 2 million miles in US, a prioritization scheme based on prediction of failure hotspots associated with estimates of downside risk is needed to reduce costs and focus efforts. In this regard we intend to propose a 2-part study as follows:

**Study 1:** An Integrative framework to empirically explore association between structural and operational stress factors related to pipeline failure risk using Spatial Regression. The stress factors comprise of both spatial features such as GIS imagery data and non-spatial features such as demographic and incidence data.



**Study 2:** Spatial prediction of hotspots.

1. **Theory and Hypothesis Development:**

**<add theory here>**

* 1. **Hypothesis for Study 1.**
* H1: Ceteris paribus, there exists positive association between operational risk factors and risk of buried pipeline failure.
* H2a: Ceteris paribus, Socio-economic and demographic factors associated with population living in the vicinity of buried pipeline network significantly accentuates the relationship between structural risk factors and risk of pipeline failure.
* H2b: Ceteris paribus, Socio-economic conditions of population living in vicinity of buried pipelines significantly moderate the association between operational factors and risk of buried pipeline failure.
* H3: Ceteris paribus, Structural risk factors positively moderate the relationship between operational risk factors and pipeline failure risk.

1. **Empirical Setting, Data and Research Design:** 
   1. **Empirical Strategy (Partly Defined): 2 Stage Model**

***Stage 1.*** *Latent Class Analysis (LCA) to group population into two distinct socioeconomic class.*

Group the population into class A, adverse socio-economic class, and B, less adverse socio-economic class by maximizing log-likelihood of mixture distribution of population of patients as shown in equation 1.

***Stage 2.*** *Estimate “time to failure” using Cox Proportional hazard models.**(Shekhar, Shashi, and Xiong, Hui. Encyclopedia of GIS. Boston, MA: Springer US, 2008.) has a chapter on Spatial Survival Analysis).*

* 1. **Data for Study**

[Incident Data Reports](https://www.phmsa.dot.gov/data-and-statistics/pipeline/distribution-transmission-gathering-lng-and-liquid-accident-and-incident-data): Incident data reports are submitted to Pipeline and Hazardous Materials Safety Administration for various categories including Gas distribution incident, Hazardous Liquid Accident data, Gas Transmission and Gathering Incident and LNG Incident data. Reports are exhaustive and comprise of 600+ variables, Available from 1986 onwards.

Pipeline and Hazardous Materials Safety Administration says following about Incident Data Reports:

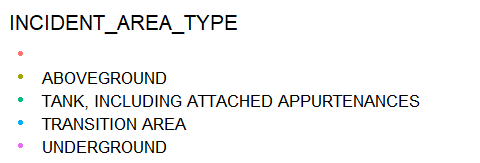
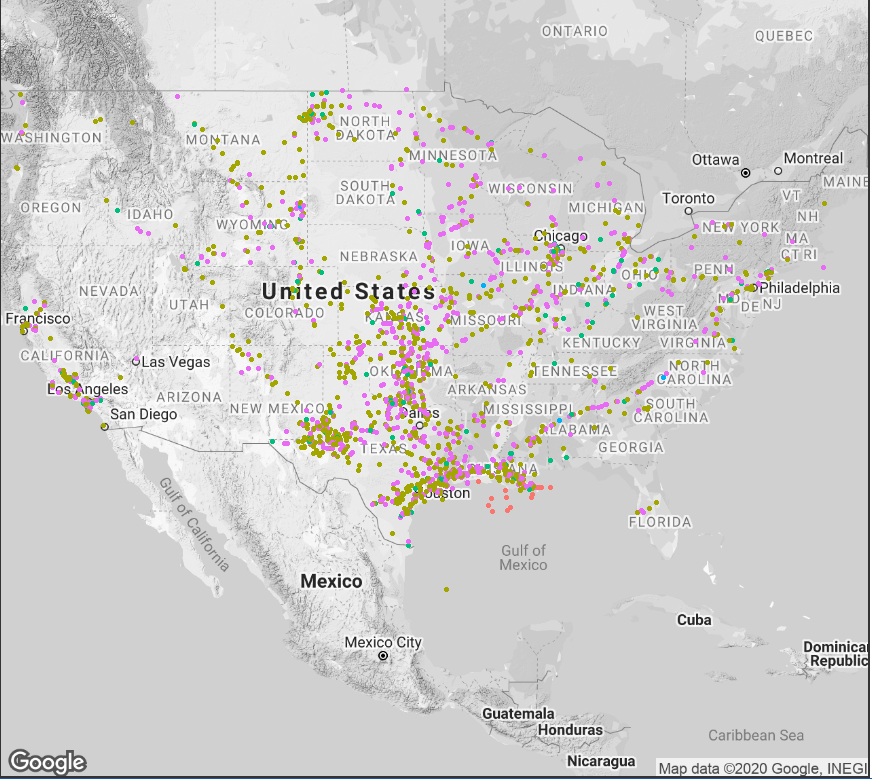
“Title 49 of the Code of Federal Regulations (49 CFR Parts 191, 195) requires pipeline operators to submit incident reports within 30 days of a pipeline incident or accident. The CFR defines accidents and incidents, as well as criteria for submitting reports to the Office of Pipeline Safety.”

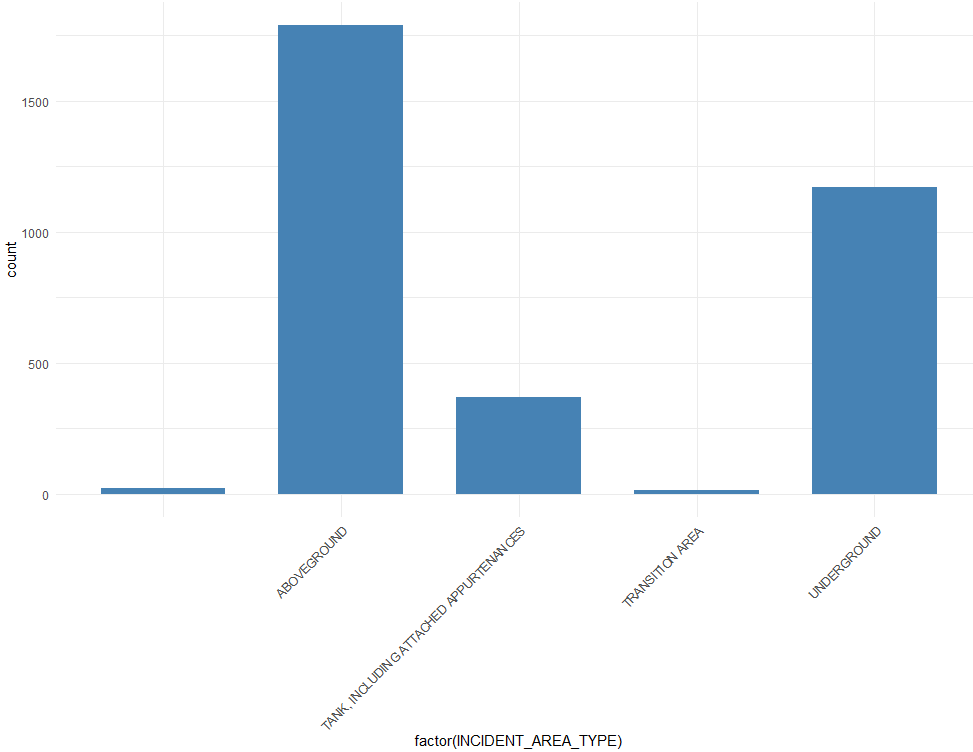
1. **Analysis:**

The Analysis presented in this section was conducted using R (v 3.6.3) spatial analysis packages from CRAN available including sp, [sf](https://journal.r-project.org/archive/2018/RJ-2018-009/RJ-2018-009.pdf), rgdal, [ggmap](http://stat405.had.co.nz/ggmap.pdf). Also two nice online books on analysis including [Geocomputation with R](https://bookdown.org/robinlovelace/geocompr/intro.html) and [Spatial Data Science with R](https://rspatial.org/index.html) was referred. The R-script for same has been included with Final Project Report. For the purpose of report we only focused on Hazardous Liquid Accident data however please note that similar analysis needs to be conducted for Gas distribution incident, Gas Transmission and Gathering Incident and LNG Incident data as well.

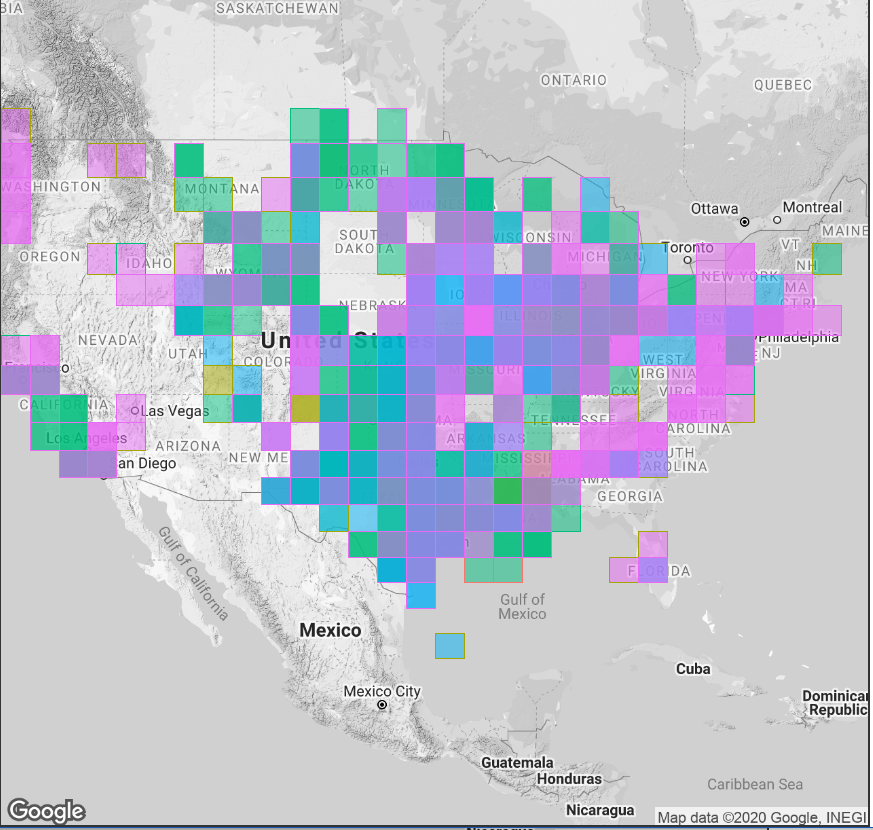
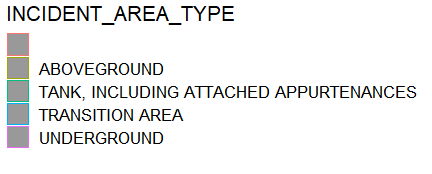
* 1. **Exploratory Data Analysis**

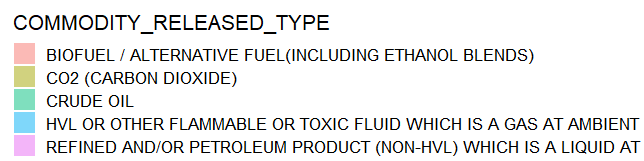
**Hazardous Liquid Accident Data: Bubble plot**

****

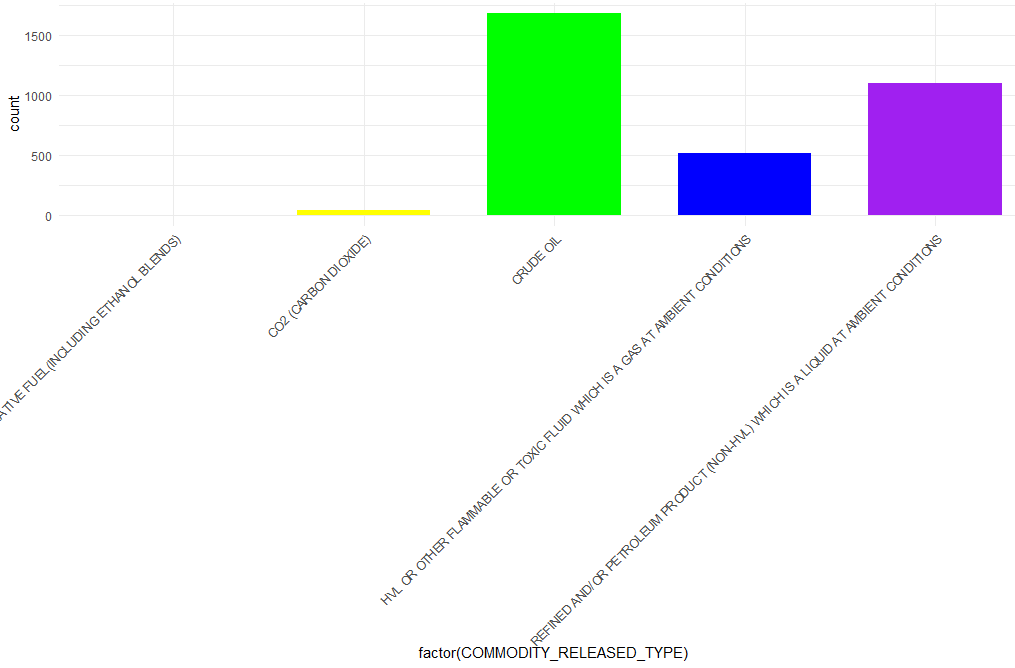
**** The bubble plot provides a visual cue that Louisiana has a high concentration of above ground pipelines.

**Hazardous Liquids Accident data - Binned plot**

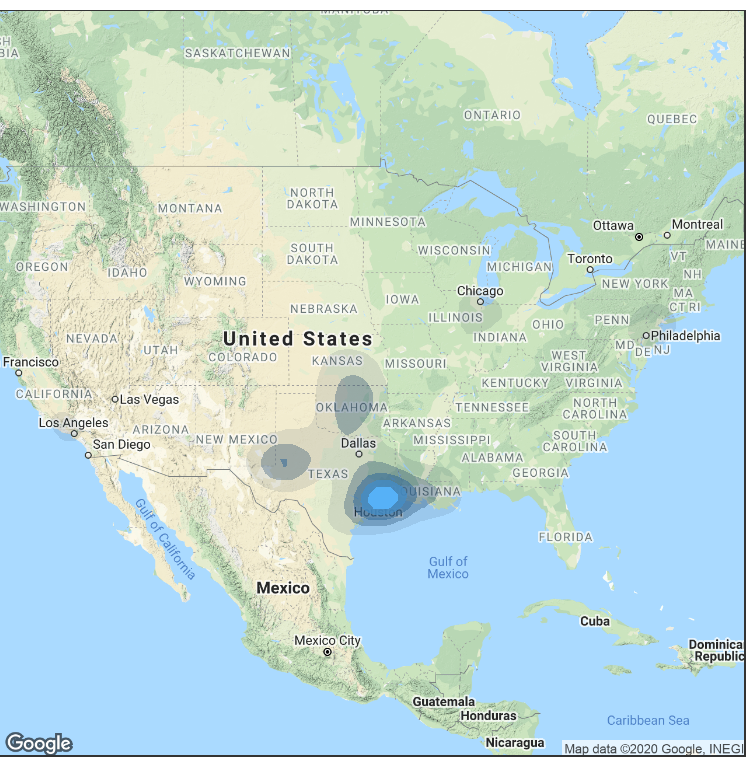
**** ****

****

The binned plot shows the areas where spill occurred along with the type of commodity released.

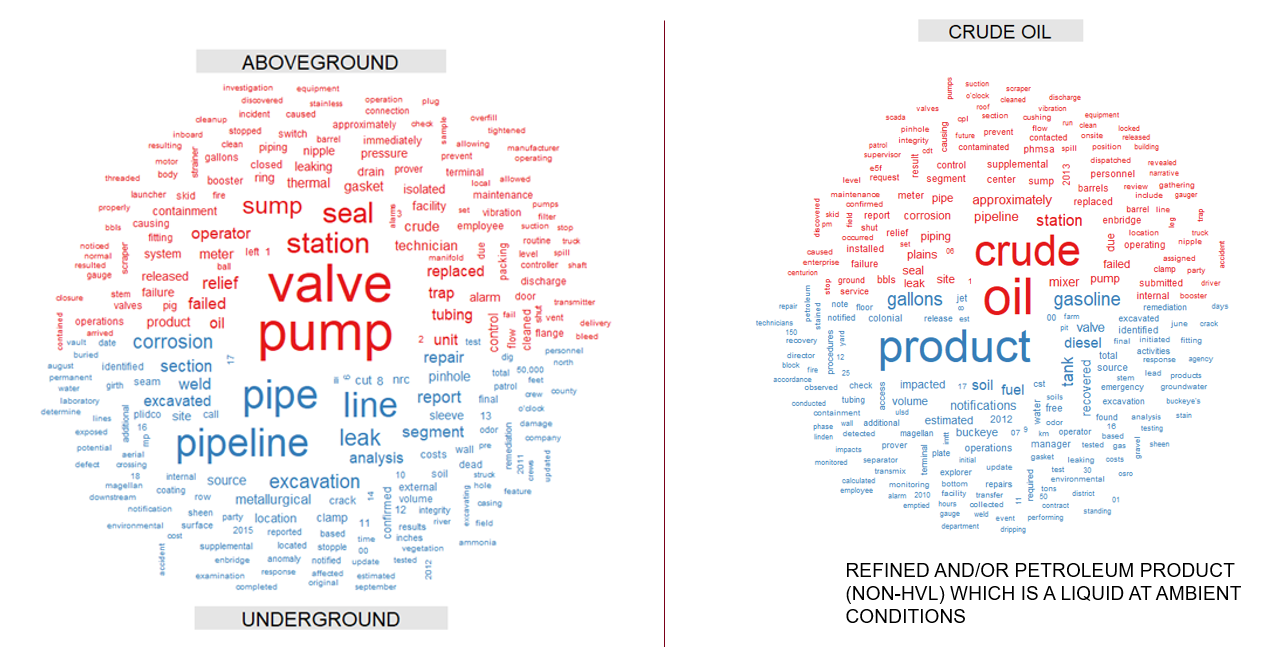
****

**Contour Plot (Needs to be fixed) :** This plot is currently generated by performing 2d Kernel density estimate on Location (longitude/Latitude) followed by displaying the results with contours. For now, it just shows the concentration of Incidents that seem to be clustered around Louisiana however what will be interesting to see is that if we normalize the incidents by length of pipeline then whether the cluster is still Louisiana or somewhere else. We tried getting this done however there is no data available regarding length of pipeline in Incident data reports. The data is very specific to the incident reported and contains measurements for Rupture Length, Length of Segment Isolated. We may have to look at some other data source to incorporate information on pipeline length.

****

**Incident Data Narrative Analysis**

Incident data provides a rich narrative in form of incident reports submitted to PHMSA. A look at summary of these incident reports will likely provide us with an intuition about which factors to consider as covariates for our analysis.



* 1. **Data Transformation**

**Step 1: Create spatial object:** Used sp (spatial package) function coordinates (…) to convert a data frame into sf[1] (simple features) spatial object. Simple features is an open standard developed by OGC (Open Geospatial Consortium). It’s a hierarchical data model that represents most used geometries and it provides a way of encoding features in data with spatial context.

coordinates(hl2010) <- c("LOCATION\_LONGITUDE", "LOCATION\_LATITUDE")

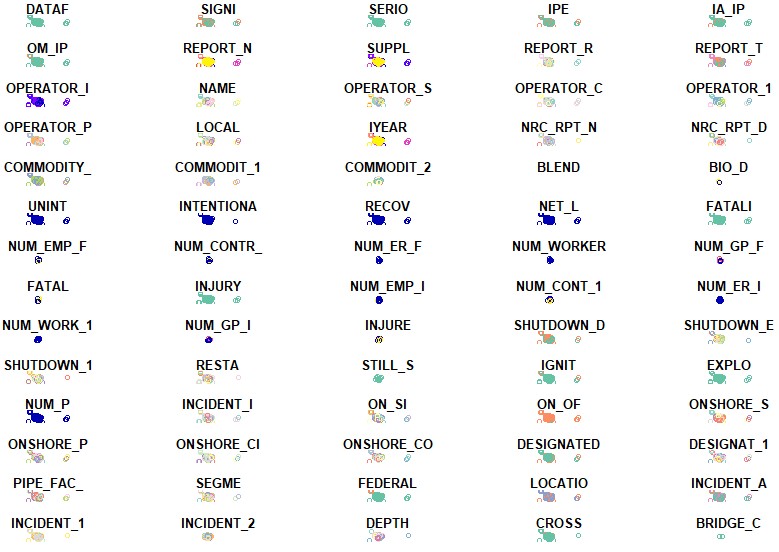
proj4string(hl2010) <- CRS("+init=epsg:4326")

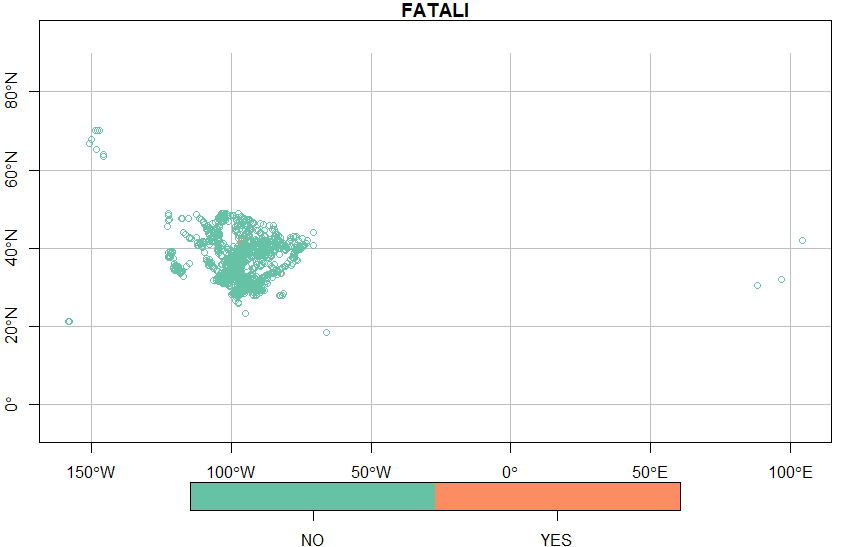
**Step2: Plot:**

* h1\_sf <- st\_read("hl\_termpaper")
* Plot(h1\_sf) ### This will create spatial distribution of all variables in the data frame.
* plot(h1\_sf[, 30], key.pos = 1, axes = TRUE, graticule = TRUE)

The results from Step 1 and 2 are shown below:

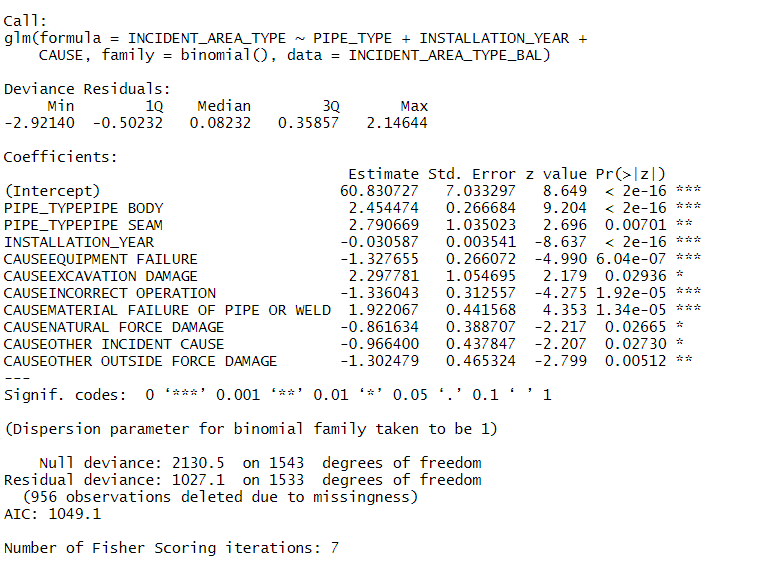
**Result from Step1:** Every column has been assigned a spatial attribute (mostly point object) in each case. Note that it doesn’t make sense to assign a spatial attribute to each variable but its just a cheap shortcut to get conversion done in one go. The plot below just shows that every variable gets a spatial attribute.

**** E.g.If we zoom into Fatalities Variable then this is what its spatial transformation looks like:



* 1. **Spatial Regression**

Once we have transformed data, we can now get going on validating our hypothesis but before that let’s look into few example spatial regressions just to see if data has any signal to perform analysis to begin with. Below we assess associations between Incident Area Type (Aboveground or Underground) which is treated as a dichotomous variable with 2 possible outcomes i-e {0,1}. The covariates inclue Pipe\_type, Cause (e.g. Equipment Failure, Excavation Damage, Incorrect Operation, Material Failure of Pipe or Welding, Outside Force Damage etc.) . The outcome of spatial model is shown below. We set the reference level of dependent Variable set to “Underground”.



The results make intuitive sense indicating a strong signal in data for further analysis including validating our key hypotheses and predicting hotspots of failure. Few observations are summarized below:

Cause like excavation damage is positively associated with Underground pipelines with significant P\_value (<0.05) while Installation year is negatively associated with same implying the old pipelines are more prone to an underground incident. On the other hand, Cause including, Outside Force Damage, Equipment Failure, Incorrect Operation are not associated with Underground incidents (for a dichotomous response variable, a negative sign of slope coefficient estimate for a covariate implies its not associated with response variable, The P-value <0.05 indicates that chances of these covariates /factors being associated with Underground incident is rather too low. On other hand we shouldn’t be surprised to see a strong positive association between these covariates / factors and incident area type = “Overground”).

**A word on R^2**: While R^2 is not a reliable measure to assess overall quality of model in case of a dichotomous response variable, a more reliable measure to compare relative quality of 2 different models and to assess contributions of factors in terms of explaining overall variability of Response would be AIC (Akaike Information Criteria). A higher AIC is an indicator of a better-quality model. For this model AIC is 1049. We would try to get a model with higher AIC by adding more relevant covariates. Note that AIC does penalize for adding too many variables or if model has higher VIF (Variance Inflation Factor).

1. **Conclusion:**

While this is not a complete work by any means, our initial investigation has provided us with a framework to carry out our specific Aims and validated the veracity and signal strength of data along the way. We plan to continue extending this study and get it to a publishable format by end of this year. We intend to send out the complete study for publication at the [Special Issue of Production and Operation Management](https://www.poms.org/Business-Analytics%20-%20Special%20Issue%20(Revised).pdf) Journal by Feb 2021.

1. **References:**
2. Okabe, Sugihara, Sugihara, Kōkichi, and Ebrary, Inc. Spatial Analysis along Networks: Statistical and Computational Methods. Statistics in Practice. Hoboken, N.J.: Wiley, 2012.
3. Barthélemy, Marc. "Spatial Networks." Physics Reports 499, no. 1-3 (2011): 1-101.
4. A. Ouadah, "Pipeline Defects Risk Assessment Using Machine Learning and Analytical Hierarchy Process," 2018 International Conference on Applied Smart Systems (ICASS), Medea, Algeria, 2018, pp. 1-6.
5. [Pipeline condition prediction using Neural Network Models](https://ascelibrary.org/doi/abs/10.1061/40800(180)61)
6. [Safety Assessment of oil and gas pipelines using satellite information](https://www.sciencedirect.com/science/article/pii/S2452321619306304)
7. Baghdadi, A.H.A, and H.A Mansy. "A Mathematical Model for Leak Location in Pipelines." Applied Mathematical Modelling 12, no. 1 (1988): 25-30.
8. <https://response.restoration.noaa.gov/about/media/investigating-environmental-impacts-kalamazoo-river-oil-spill.html>
9. Makhlouf, Abdel Salam Hamdy, and Mahmood Aliofkhazraei. *Handbook of Materials Failure Analysis : With Case Studies from the Oil and Gas Industry*. Amsterdam: Elsevier, 2016.
10. [Seismic risk assessment for oil and gas pipelines](https://www.sciencedirect.com/science/article/pii/B9780857092687500256?via%3Dihub)
11. [Pipeline condition prediction using Neural Network Models](https://ascelibrary.org/doi/abs/10.1061/40800(180)61)
12. Chinedu I. Ossai. "A Data-Driven Machine Learning Approach for Corrosion Risk Assessment—A Comparative Study." *Big Data and Cognitive Computing* 3, no. 2 (2019): 28.
13. Layouni, Mohamed, Sofiene Tahar, and Mohamed Salah Hamdi. "A Survey on the Application of Neural Networks in the Safety Assessment of Oil and Gas Pipelines." *2014 IEEE Symposium on Computational Intelligence for Engineering Solutions (CIES)*, 2014, 95-102.
14. Wilkinson, L., Wills, G. & SpringerLink, 2005. *The grammar of graphics* 2nd ed., New York: Springer.
15. Wickham, H. & SpringerLink, 2009. *Ggplot2: elegant graphics for data analysis*, Dordrecht ; New York: Springer.