Gas Leak prediction based on San Francisco Emergency Data

Fangshu Lin(fl1210@nyu.edu)

Guobing Chen(gc2300@nyu.edu)

Yixuan Tang(yt1369@nyu.edu)

Machine Learning • CUSP • 2018

1 ABSTRACT

Gas leak is a crucial risk problem in urban life. Comparing to detecting gas leak using instruments in a big city, machine learning algorithm with open data from San Francisco provides a more innovative method with less expense. However the field of gas leak detection has not been touched by researches using machine learning techniques. In this paper we applied methods as decision tree, gaussian process and location outlier factor algorithm to predict gas leak occurence and identify strong indicators of such events. Based on the result, the methods proved their worth in terms of accuracy and add value by optimizing the gas leak inspection effort. The areas with higher likelihood to have gas leak incidents and the socio-economic attributes that contribute significantly to gas leak incidents have been identified.

2 INTRODUCTION

As San Francisco's population has boomed, calls to the Department of Emergency Management's 911 center have surged. In 2007, dispatchers handled 919,908 calls. By 2016, that number was up to 1.27 million. The volumes of emergency data available are enormous for predictive analysis.

Emergency events consist of various types: fire(marine fire, outside fire, structure fire), explosion, traffic collision, medical incident and gas leak etc. There are common causes for emergencies such as illegal construction is highly related to both fuel spill and electrical hazard, while each type of emergency has specific triggers or relational contributors. Responding to those calls and handling the emergency events are time and money consumption, being reactive in this process always leads to economic losses. Given the tremendous data, there is great chance for fire department to foresee emergency happenings and being proactive rather than reactive.

In this project, we choose this topic based on the concerns of the following aspects:

1. **Significance**: gas pipeline construction has developed quickly both at home and abroad, and gas construction is a large and complex system engineering, from gas station, pipeline to users, any problem in any part could affect the whole system and hurt the normal life of million of households. With the consideration of public safety, urban development and economic savings, gas leak prediction is a meaningful project.
2. **Novelty**: based on our literature review which described detailed in the next part, we found that the majority of existing work is focus on gas leak detection, not the event prediction. There are mature techniques to measure gas leak volume in construction or chemical industries, however, the machine learning techniques appear far less in related research.
3. **Feasibility**: San Francisco is the most densely populated city in the state, and the Fire Department Calls for Service dataset provide the basis for a considerable number of studies, along with the demographic dataset, there is plenty space for data scientist to use data to identify strong indicators and discover the pattern of emergency events such as gas leak.

In conclusion, our motivation is to get better understanding of gas leak emergency events and improve the knowledge of how to optimize resources allocation to protect the lives and properties from such emergencies. The main objective is to incorporate socio-economic datasets with historical gas leak data in San Francisco to find out which factors are most strongly associated with the likelihood of gas leak emergencies and further predict such events.

3 LITERATURE REVIEW

Gas leak is an important problem in protecting public safety and ensure functionality of many products. Researchers from several disciplines have strong motivation to detect gas leak, examine gas leak causes and the efficacy of potential countermeasures. This literature comes from several academic and policy communities: construction engineers, public health specialists, urban planners, and government agencies. The most popular research around gas leak is the detection method, by measuring the pressure and flow at pipeline inlet and outlet.

A. Previous analytical techniques

Study methods have included decision tree classification, SVMs classification, DBSCN cluster analysis, and K-nearest neighbor algorithm etc. The literature employs a variety of techniques depending on whether they are investigating absolute gas leak amounts, related factors of gas leak or categorical variables such as severity, or of damage severity levels. A similar application of machine learning was done by Xiu-fang Wang et al. In this study, they performed decision tree, clustering and K-nearest neighbor algorithms to discover the gas leak information and the objective laws behind the natural gas pipeline transmission, intrinsically linking to the each parameter and development trend.

Another novel classification technique developed by Lam Hoo Lee et al in 2013, namely Euclidean-Support Vector Machines(Euclidean-SVM), make decision on the real-time condition data of the pipeline. In this study, they utilize long range ultrasonic transducers (LRUTs) to detect both internal and external corrosions in continuous environment, and process the real-time data to ascertain whether the Euclidean-SVM classifier can sufficiently detect the pipeline's failure. Through a series of experiments they further proved that the novel Euclidean-SVM approach has lower dependency on the kernel selection and values of soft margin parameter and relatively good performance, as compared to the traditional SVM classification approach.

B. City-specific studies of gas leak

In Washington, D.C. and surrounding region, one of the leading causes of gas leak emergencies are damages to the underground gas lines by individuals and contractors performing excavation work, including digging, boring and directional drilling(Washington Gas).

To protect tenants from gas leak risk, in December, 2016, the New York City Council worked to remedy that a package of bills was signed into law by Mayor Bill de Blasio. It amends the city's administrative code to require building and apartment owners to post signage alerting residents and potential tenants on the proper protocol if a gas leak is suspected.

4 DATA

4.1 Data Processing

We utilize the Fire Department Calls for Service dataset assembled by the fire department and maintained by DataSF. It contains all fire units responses to calls. Each record includes the call number, call type, incident number, address, location, unit identifier, priority and disposition. All relevant time intervals are also included: call received date, dispatch date and response date. Addresses are associated with a block number, intersection or call box, not a specific address. Location represents the geospatial locational data.

The original dataset contains 17,354 entries of gas leak call type from 2000 to 2017. We carried out several data-cleaning steps:

• Because of the temporal nature of our analysis, we transformed the column 'Received DtTm' into datetime format, and further split into hour, weekday, month variable.

• In this dataset, each call records may has multi entries due to multiple fire engine deployment, in order to conduct our analysis on exact events, we dropped the duplicated entries for each event, in the final dataset, one entry identifies one emergency event.

• Besides the location and time, we involved socio-economic attributes from American Community Survey to predict the number of gas leak. The attributes we chose were on census tract level, and related to rent, whether kitchen has complete facilities, whether kitchen is owned by tenure or house owner, income level, education and the year when the building is built.

• In order to analyze the spatial relation between gas leak incidents, we first created fishnets of 23 by 34 grid cells giving the boundary coordinates in San Francisco (excluding Treasure Island). Then create polygons for each grid cell and add to layer. The output of created fishnets is in ESRI shapefile format. Each grid is 480 meters by 380 meters. Then we were able to aggregate gas leak records to each grid cells based on the latitude and longitude information.

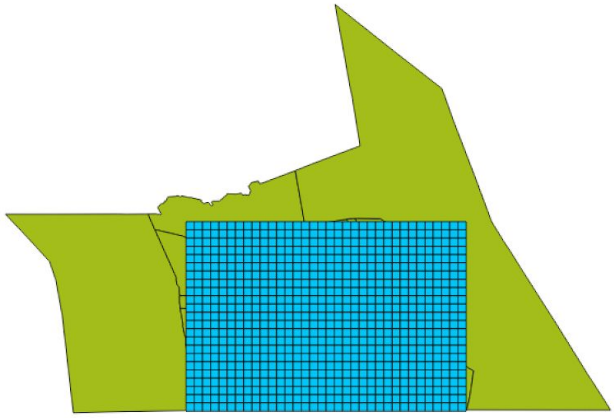
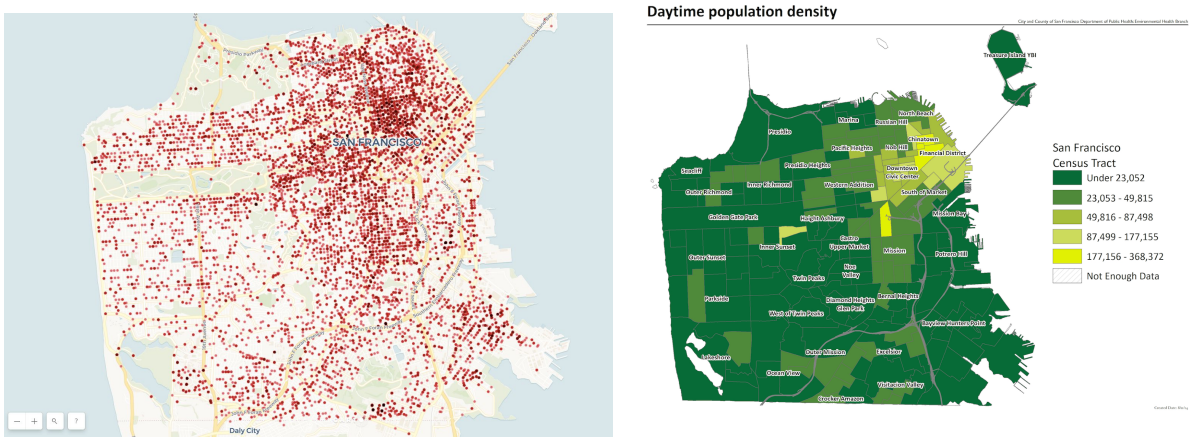


Figure 1. The census tract map and the grid map

• Since the socio-economic data we extracted are all on the census tract level, it is necessary to conduct a spatial join for the grid map and the census tract map. By using interpolation in QGIS, we obtained the socio-economic attributes for each grid.

4.2 Exploratory Data Analysis

Spatial patterns



(a) (b)

Figure 2. (a)Gas Leak Event Map (b)San Francisco Population Daytime Density[1]

The gas leak events dataset was visualized on a map of San Francisco using Carto(Figure 2(a)). We see lower event volumes in lower western areas where population density is known to be low(Figure 2(b)), and high event volumes in north-eastern areas where population density is known to be high. We also observe that the furthest areas of the city, such as the Treasure Island, event volumes appear to be relatively lower.

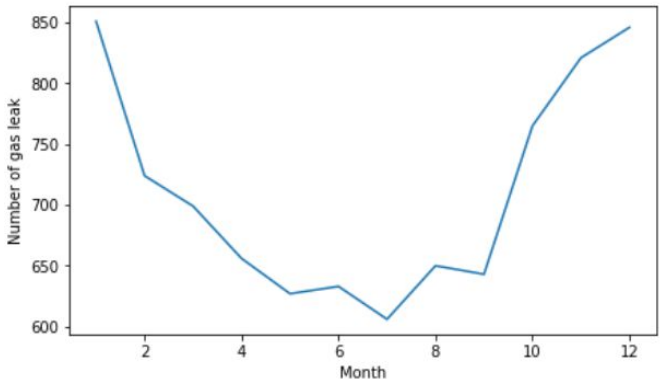


Figure 3. (a)Gas Leak Event Map

Temporal patterns

By visualizing the event counts over time at month level (Figure 3), we can see there is a clear seasonal trend of gas leak event during 2000 to 2017. Most gas leak occur during the winter season(November to January) and summer season has much lower gas leak emergency counts. This could be explained by the increasing usage of gas during winter time which has higher heating need.

[1] https://sfclimatehealth.org/portfolio/daytime-population-density/ San Francisco Department of Public Health, City and County of San Francisco

5 METHODOLOGY & RESULT

5.1 Decision Tree & Random Forest

In order to predict the number of gas leak incidents that happen in a grid per month, we built a decision tree using grid row and column number, month, and 71 demographic features as covariates. Since the socio-economic data is after 2010, therefore gas leak incidents after 2010 with 4888 records was used for decision tree modeling. We set a maximum depth equal to 8 to prevent overfitting. 70% of the dataset was used as training sample and 30% as test sample. The in sample accuracy score is around 0.69 and out of sample accuracy is around 0.56. The mean squared error is around 1.23. Later A random forest model was constructed using GridSerchCV to optimize the number of estimators. The tuned model gives a out of sample accuracy score of 0.58 and mean squared error of 1.19 with 16 estimators. A tree with max depth of 4 is visualized in Figure 4.

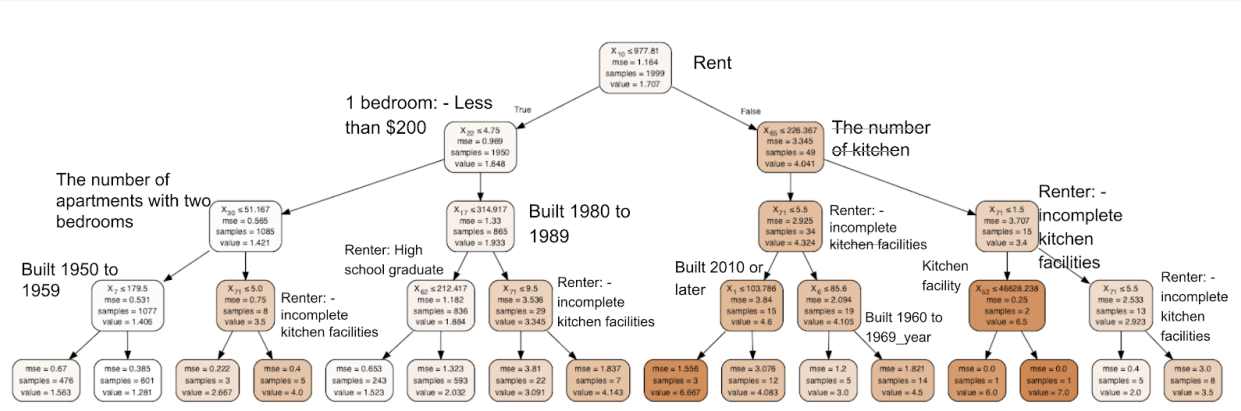


Figure 4. The Decision Tree Predicting Monthly Number of Gas Leak Using Location, Time and Demographic Data

The decision tree model has selected important attributes as rent, the number of kitchen, the situation of kitchen facilities, the year when building is built and education information. The largest monthly number of gas leak is seven while the minimum number is around 1.5. The first node is the average rent of all sizes of buildings within the grid. For grids with average rent less than 977.81, with fewer cheap one bedroom apartments of which rent are less than 200, with the number of two bedrooms less than 51.167, the gas leak incident level is low, which is around 1.5 per month. Under this condition, the less building built between 1950 to 1959, the higher the monthly number of incidents, which is around 1.56, otherwise the monthly number of incidents is around 1.2. For grids with average rent less than 977.81, with fewer cheap one bedroom apartments of which rent are less than 200 but more apartments with two bedrooms, if the apartment is occupied by renter and has more kitchens with incomplete facilities, the gas leak incidents tend to be around 4 per month, otherwise is around 2.6. For grid with average rent less than 977.81, with more cheap one bedroom apartments of which rent are less than 200, if there are more building built from 1980 to 1989 and more kitchens with incomplete facilities, the monthly gas leak incident number is around 4.1, and if there are fewer kitchen with incomplete facilities, the incident number is around 3. For grid with average rent less than 977.81, with more cheap one bedroom apartments of which rent are less than 200, if there are less building built from 1980 to 1989, and more renters holding high school degree, the monthly incident number is around 2. If there's less renters holding high school degree, the monthly incident number is around 1.5. For grid with higher rent than 977.81,the more kitchens and more kitchens with incomplete facilities, there tend to be around 7 gas leak

incidents per month. If a grid has less kitchen with incomplete facilities, the number of gas leak incidents tend to around 2.

5.2 Gaussian Process

Decision tree model is effective in identifying the important features, but it didn't take into account the spatial and temporal correlation of gas leak events. In order to predict the pattern over time and space, Gaussian Process Regression methods were performed with the dependent gas leak dataset.

First, we used the dataset from 2000-2017 to fit a Gaussian Process regression model predict the gas leak counts based on the grid row and column number. A RBF kernel is tuned for the model and a white kernel is also included to address the effects of noise. The 2-D plot of real and predict values of the events over the grids we created are shown in Figure 5. The log marginal likelihood of the data is around -2069. A map of prediction over a mesh grid map is shown in Figure 6. We found that the Gaussian Process model is effective in predicting the spatial dependent dataset such as gas leak. The grids with high likelihood of gas leak incidents are identified and the predicted values are reasonable. Using this methods, we also can obtain prediction intervals over the whole city.

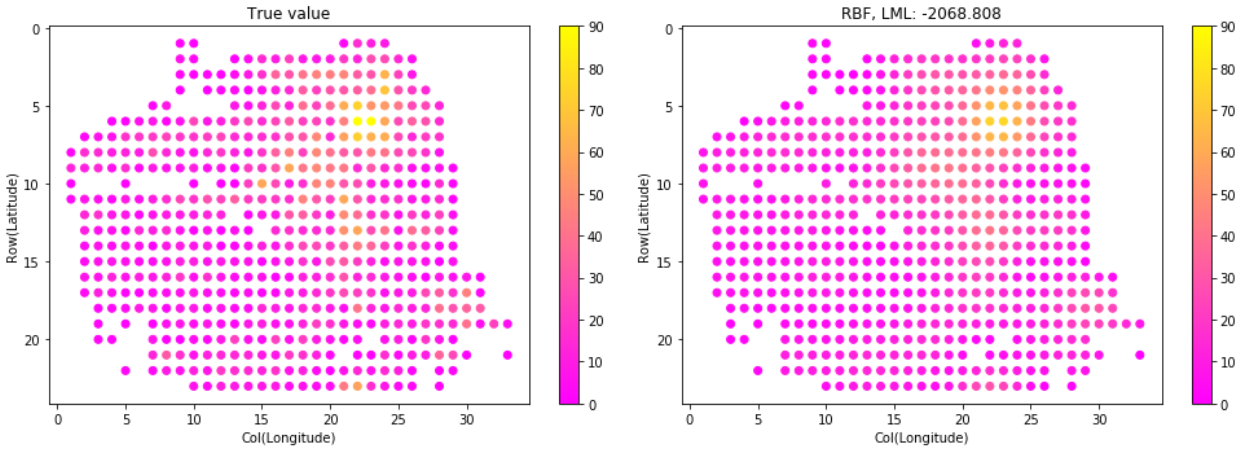


Figure 5. Gas leak events over grids in San Francisco (a) Real value (b) Prediction using GP

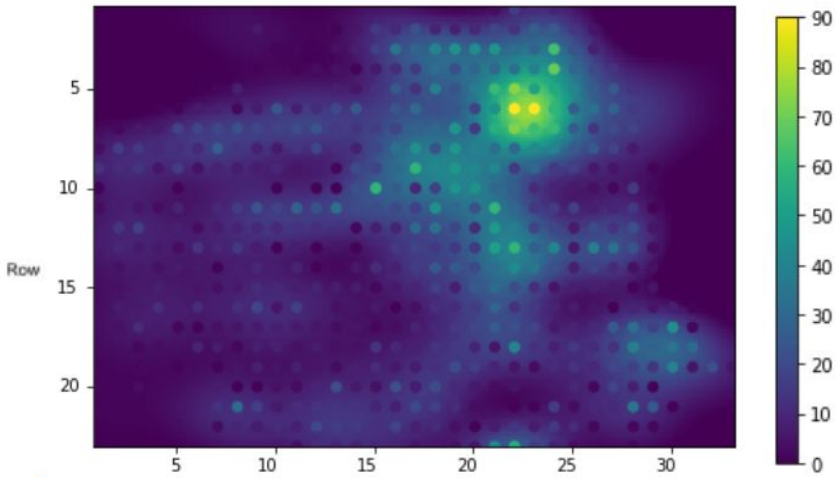


Figure 6. predictions over a mesh grid map in San Francisco overlay with the scatter plot of real value

Then we built a second model using Gaussian Process method to predict monthly gas leak events using data of 216 months from 2000 to 2017. We tuned an optimized kernel with the combination of RBF, rational quadratic and white kernel. Then we used this model to predict the monthly incidents in the next five years. The log marginal likelihood of the data is around-780. The prediction result with 95% confidence interval is shown in Figure 7. The pattern of rising trend is identified using this model.

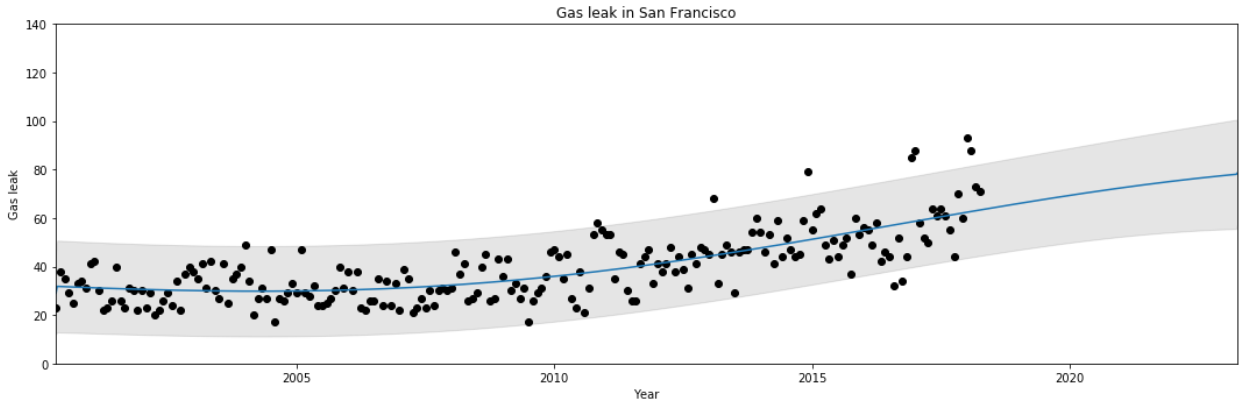


Figure 7. Monthly gas leak in San Francisco with five years prediction

5.3 Anomaly Detection

For anomaly detection, we used local outlier factor(LOF) algorithm, it finds the anomalous data points by calculating the deviation of given data point to its neighbours.

	FID	Num	Built10orlater	Built00to09	Built90to99	Built80to89	Built70to79	Built60to69	Built50to59	Built40to49	Built39orEarlier	INCOME
40	97	1	0.00	22.5	2.00	2.50	13.0	18.50	13.5	81.0	401.00	142483.50
196	293	4	4.25	14.5	55.75	77.75	79.5	162.75	374.0	304.0	1156.75	142350.25
270	372	13	0.00	32.5	24.50	62.00	179.5	204.00	139.5	149.0	1648.00	212517.50
321	425	19	0.00	27.0	49.00	50.00	302.0	198.00	208.0	135.0	1872.00	243800.00
322	426	15	0.00	40.5	49.50	66.00	203.5	145.50	150.5	102.0	1672.50	212577.00

Figure 8 . Top 5 anomalous grid

Several interesting can read from the Figure 8 above, firstly, all 5 anomalous grids have higher average household income('INCOME' feature) than the mean household income of all grids which is 104471.567402. Secondly, except of the Grid 97, rest anomalous grids have higher number of units built before 1939('Built39orEarier' feature) than average number: 766. However, when check the number of events happened in those grids, we can only state that those grids experienced slightly higher counts of gas leak events compare to other grids, so we go back and check the location of those grids, we find out that those grids located in the high population density city-centre area which reasonably explained the higher household income and older building age.

6 CONCLUSIONS

In this study, we used Fire Department Calls for Service dataset to predict gas leak incidents in San Francisco. First we constructed models using decision tree and random forest methods to predict gas leak happens in given location and month. Socio-economic features were used for the prediction. The models achieved good performance in prediction with out of sample accuracy around 0.55. The advantages of using decision tree regression is that we are able to identify important social-economic variables that affect the prediction of gas leak. It is found that the number of bedrooms, average household income, building age, and kitchen facility situation are strong indicators of gas leak. Since the gas leak incidents are highly correlated over space and time, we included Gaussian Process methods to learn this dependence structure. We built two models using Gaussian Process to identify grid cells where gas leak incidents occur more frequently and the future trend of monthly gas leak emergencies. Finally, we performed local outlier factor algorithm to identify the five most unordinary locations in San Francisco.

The methods in this study to foresee emergencies are informative and can facilitate the management of city agencies like fire department. Based on the analysis, more facilities and personnels can be allocated to the areas with high frequencies of gas leak emergencies and lower response time when emergency happens. The results can also used as evidence for city planners to inspect and renovate old buildings and infrastructures in cities.

7 LIMITATIONS AND FUTURE WORK

After rule out duplicated and wrong geo-located entries of gas leak events, our dataset ends up with 8624 entries. This volume of datasets appear not representative enough in the prediction algorithm, we ended up with around 0.69 in sample and around 0.56 out sample accuracy.

Next steps in developing the model along the lines conceptualized in this paper would include a focus on raising model accuracy through integrating additional features that contain information on emergency risk – including potentially demographic information(need a series of test) – and on testing additional analytical techniques applied to the same cleaned dataset.

Beyond the scope of the current project, the prediction models can then be applied through a Javascript interface and Geopandas to show a real-time map of predicted gas leak events in month level. The project has therefore would be a valuable urban data science contribution for the San Francisco Fire Department, city planners and local residents.

8 CONTRIBUTION OF EACH TEAM MEMBER

Fangshu Lin: Literature review, Grid creation algorithm, Decision tree, Random forest, Gaussian process

Guobing Chen: Literature review, Data processing, Decision tree, Random forest, Gaussian process

Yixuan Tang: Literature review, Exploratory data analysis, Decision tree, Anomaly detection

9 REFERENCES

[1] Xiu-fang Wang, Yan Wang, Chun-lei Jiang, Hong-wei Liang, Natural gas pipeline leak detection based on data mining https://ieeexplore.ieee.org/document/5768886/

[2] Lam Hong Lee, Rajprasad Rajkumar, Lai Hung Lo, Chin Heng Wan, Dino Isa. Oil and gas pipeline failure prediction system using long range ultrasonic transducers and Euclidean-Support Vector Machines classification approach
https://www.sciencedirect.com/science/article/pii/S0957417412011207

[3] R. B. Santos; E. O. de Sousa; F. V. da Silva; S. L. da Cruz; A. M. F. Fileti Detection and on-line prediction of leak magnitude in a gas pipeline using an acoustic method and neural network data processing
http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-66322014000100014&lng=en&tlng=en

[4] Andrew Gordon Wilson, Advanced GP methods: tutorials, papers, and Matlab code
https://people.orie.cornell.edu/andrew/code/

[5] Spectral mixture kernels: A.G. Wilson and R.P. Adams. Gaussian process kernels for pattern discovery and extrapolation. Proc. ICML, 2013.

[6] S.R. Flaxman, A.G. Wilson, D.B. Neill, H. Nickisch, and A.J. Smola. Fast Kronecker inference in Gaussian processes with non-Gaussian likelihoods. Proc. ICML, 2015a.
http://www.cs.cmu.edu/~neill/papers/icml15.pdf

[7] S.R. Flaxman, D.B. Neill, and A.J. Smola. Gaussian processes for independence tests with non-iid data in causal inference. ACM TIST, 2015b.
http://www.cs.cmu.edu/~neill/papers/TIST2015.pdf

DATA

1.KITCHEN FACILITIES FOR ALL HOUSING UNITS 2008-2012 American Community Survey 5-Year Estimates

https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_12_5YR_B25051&prodType=table

2.KITCHEN FACILITIES BY MEALS INCLUDED IN RENT

https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_12_5YR_B25054&prodType=table

3. TENURE BY KITCHEN FACILITIES

https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_12_5YR_B25053&prodType=table

4. YEAR STRUCTURE BUILT

https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_12_5YR_B25034&prodType=table

5.BEDROOMS BY GROSS RENT

https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_12_5YR_B25068&prodType=table

6.MEDIAN FAMILY INCOME IN THE PAST 12 MONTHS (IN 2012 INFLATION-ADJUSTED DOLLARS) BY FAMILY SIZE

https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_12_5YR_B19119&prodType=table

7.TENURE BY EDUCATIONAL ATTAINMENT OF HOUSEHOLDER

https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_12_5YR_B25013&prodType=table

8.Census 2010_ Tracts for San Francisco

https://data.sfgov.org/Geographic-Locations-and-Boundaries/Census-2010-Tracts-for-San-Francisco/rarb-5ahf/data