Application of machine learning to the analysis of pipeline incidents in Canada

Marcelo Guarido, Daniel O. Trad, and Kristopher Innanen

ABSTRACT

Analyzing pipelines incidents in Canada is important to understand their impact on the environment and workers' safety. Data provided by the Government of Canada is confusing and incomplete but contains useful information that can be analyzed and modeled to mitigate future incidents. Most of the reports come from the province of Alberta, which contains most of the pipelines in Canada, and are mainly related to 4 companies. We could notice a correlation between the number of incidents per year with the price of the WTI crude oil price, as well as weekend effects. No seasonality is observed in the data, but we noticed some outliers - months with a larger number of reports than the average - and they are related to a single company. Clustering for dimensionality reduction and cluster analysis, applied on pipeline and maintenance information, showed 4 main clusters, each associated with different insights, such as the average volume of substance released, how long took for the occurrences discovering, and emergency level.

INTRODUCTION

Pipelines are used to transport oil and gas (among other substances) are considered the safest method of transportation for them, according to Green and Jackson (2015). However, even rare, failures causing the spill of such substances occur, becoming a hazard to the environment as there is a high probability of the contamination of water bodies, soil, and wildlife (Michel and Fingas, 2016). In Canada, accidents are usually onshore. Several studies use statistical models to estimate risk of the pipelines due to natural hazards (Badida et al., 2019), or create forecast model to evaluate oil exposure due to spills (Wang et al., 2020).

Belvederesi et al. (2018) use accident reports provided by PHMSA (Pipeline and Hazardous Material Safety Administration) to analyze the of liquid spills. The authors pointed how major failures have significantly affected the environment, such as the Plains Midstream Canada ULC pipeline failure of April 2011, when $4300m^3$ of crude oil were released into a water body close to the Traditional Territory of the Woodland Cree First Nation near Peace River, Alberta, according to the Alberta Energy Regulator 2014. However, the analysis of spills in Canada has proven to be difficult (Belvederesi et al., 2017).

In this project, we analyze pipeline incidents, that resulted on spill or not, with the goal to find patterns that can be used to prevent future incidents. We used the *pipeline incident reports* from the Government of Canada on a period from 2008 to 2019. We separated the data into clusters, and also analyzed the events over time, looking for correlations with oil production and/or oil price.

DATA ANALYSIS

The data was downloaded directly from the Government of Canada repository, and contains a large number of information, such as the incidents locations (latitude and longitude), operator's name, type of incident, type of substance leaked (if any), volume spilled, incident information, among others. Although there are lots of columns about the incident itself, their descriptions on environmental impact are limited, with just a few columns dedicated to it. Also, the provided file (*CSV*) is quite confusing and incomplete, with many columns hard to understand what they are, and many empty or missing information in most of the columns.

Summarizing the Data

Figure 1 shows the location of all the incidents report from 2008 to 2019, coloured by the type of substance the pipeline was transporting. Most of the reports come from Alberta, the province where most of the pipelines are located, but most of the Canadian provinces have had some type of occurrence.

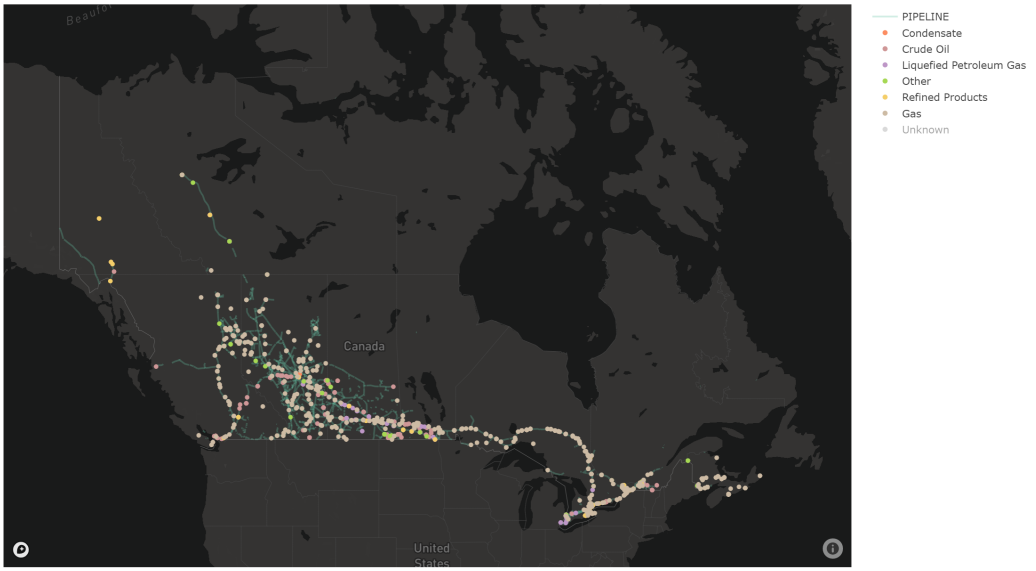

FIG. 1. Map of pipeline incidents coloured by the substance transported.

Although the confusing data, lots of information are interesting and useful for analysis. Figure 2 summarizes some of those information on pie charts. By checking the type on incident reported, 51% correspond to the *release of substance*, 22.2% are related to *operation beyond design limits*, 17.7% are associated to *fire*, while the remaining reports are divided on several other causes.

Still on Figure 2, the most common type of release is *gas* (43.5%), with *liquid* (5.61%), and *miscellaneous* (4.04%) far behind. 46.9% are *not applicable*. This means that almost half of the reports are not related to any type of release, but with some incident on the pipeline and/or the working facility, and most of the releases are related to less viscous substances (gases).

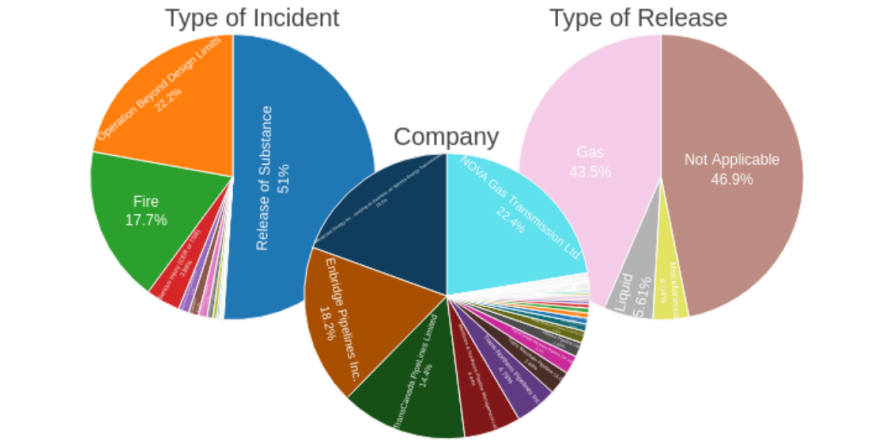

FIG. 2. Charts summarizing the types of incidents, types of release, and companies by the proportion of appearance.

The top 4 companies with the largest number of occurrences (Figure 2) are *NOVA Gas Transmission LTD.* (22.4%), *Westcoast Energy Inc.* (19.5%), *Enbridge Pipelines Inc.* (18.2%), and *TransCanada Pipelines Ltd.* (14.4%). These companies are the ones with the largest length of pipelines in Canada's territory, and they been the top 4 does not come as a surprise.

Time Series Analysis

There 3 dates for each report in the data set: the dates of *occurrence* (estimated date and time of the failure), *discovery* (date and time in which the incident was noticed), and *report* (day companies informed an incident to the Government). Each data is composed of *year*, *month*, *day*, *hour*, *minute*, and *second*. The informed time is in the $12h$ system, but with no information if it was *am* or *pm*, making it not possible to analyze if the time of the day impacts the likelihood of an incident.

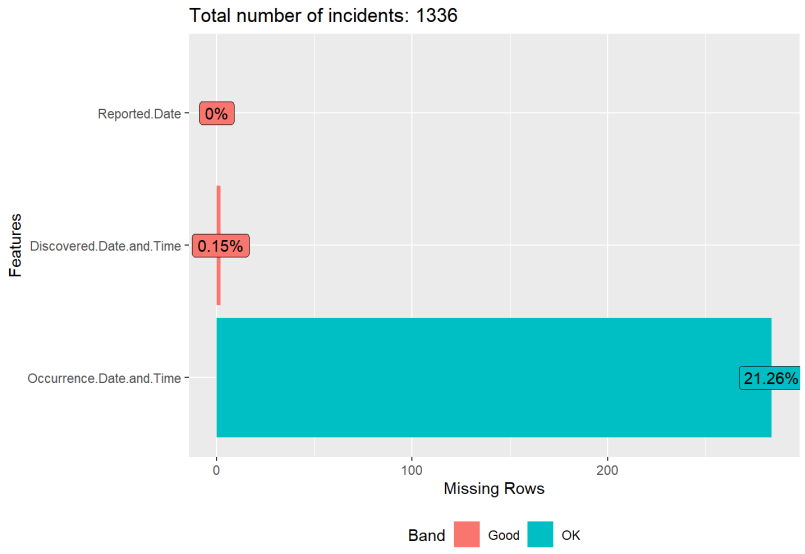

FIG. 3. Proportions of missing data for each date.

Another observation is the amount of missing data for each date (Figure 3). While dates of report and discoveries are well sampled, the date of occurrence has about 21% of missing data. This can be due to the impossibility determining when the failure happened, and some could be due to human mistake when filling the report.

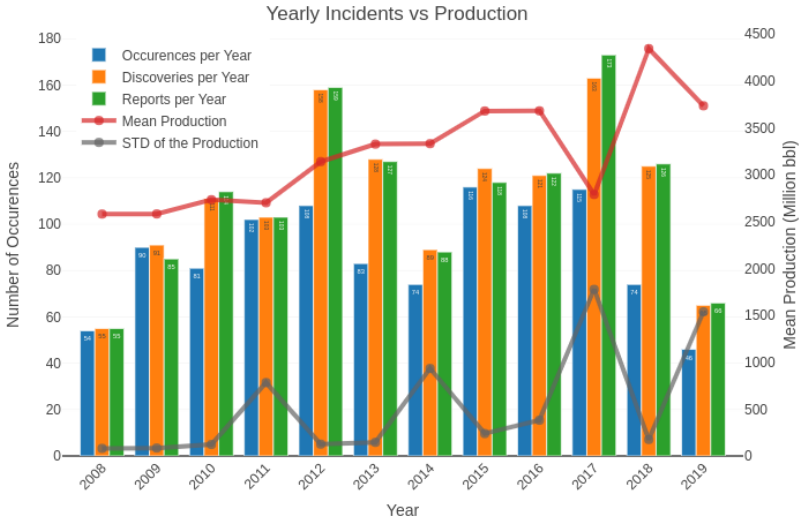

FIG. 4. Comparing the yearly number of incidents with the oil production.

On Figure 4 is presented the comparison of the yearly number of incidents (histogram) with the oil production (red line is the mean production, while gray is the standard deviation of the production). There is no apparent correlation between them. The production has a tendency to increase every year (except for 2017 and 2019), while the number of incidents oscillates. On the same plot there is the *standard deviation* (STD) of the oil production for each year, on a tentative to check any correlation between years with higher variation of pipeline usage (STD) with the number of incidents. There is no apparent *linear* correlation, but it is important to point that the years 2014, 2017, and 2019 had higher variation in production, and the number of incidents were smaller when compared to neighbours in 2014 and 2016, while there is a high peak of incident in 2017. So a *non-linear* correlation can exist.

While no direct correlation was observed between incidents and oil production, a correlation may exist when comparing the number of incidents and the average of the WTI crude oil price for each year (Figure 5). The number of incidents tends to be high when the price of the oil is increasing compared to the previous year, and lower when the price drops. Let's just remember that correlation does not mean causation, but it is interesting to see those trends.

Figure 6 shows the average of occurrences and discoveries per the day of week for all the years (2008 to 2019) in gray, and the average for all the years in red. These plots show that number of occurrences and discoveries highly drop during weekends. This may be caused by the decrease of operation in facilities during Saturdays and Sundays.

Still analyzing the dates of the incidents, Figure 7 shows the average of incidents per month for each year (gray) and the average per month for all the years (red), considering the

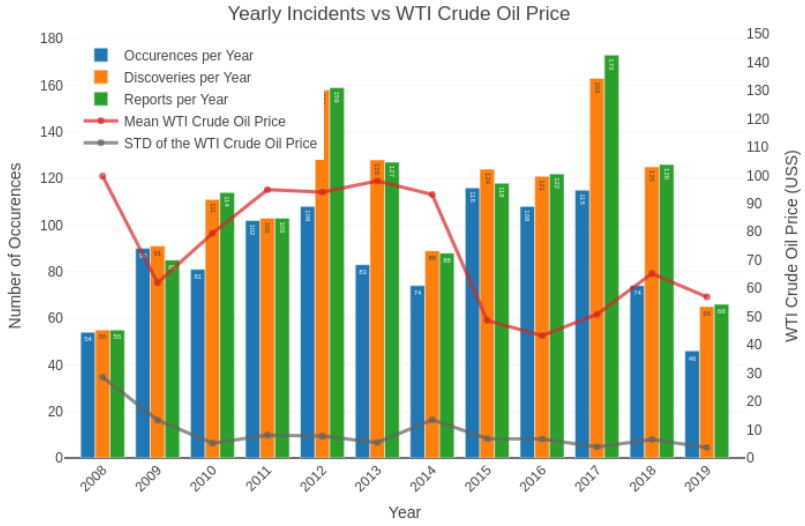

FIG. 5. Comparing the yearly number of incidents with the WTI crude oil price.

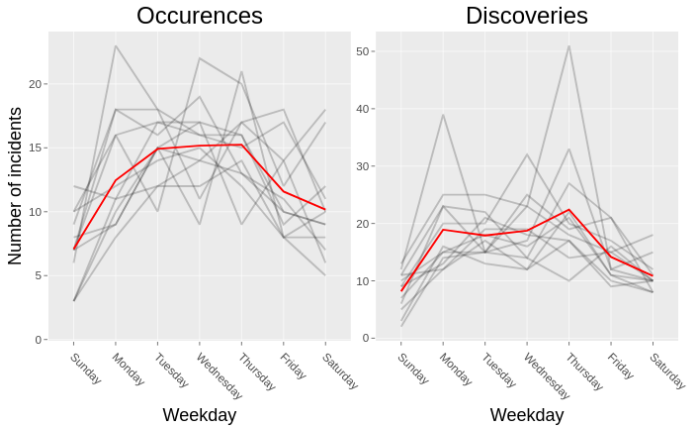

FIG. 6. Average of occurrences and discoveries per the day of week for all the years (2008 to 2019) in gray, and the average for all the years in red.

a) occurrences, b) discoveries, and c) reports in the data set. d) are the averages for all the years with error bars. There is no apparent seasonal effect on the number of incidents, but for the discoveries and reports, there are some outliers (very unusual high or low values) on *November 2010*, *February 2012*, *October 2013*, and *September 2017*. By looking at the information for these dates, most of the events are from the same company *Enbridge Pipelines Inc.*, same location, same discovery date, but different occurrence dates. Probably all the discoveries happened during the inspection date and small parts of the pipeline were checked as having some issue and, as the issues were small, they also took a longer time to be discovered. But further investigation on these events is recommended.

MODELING

This section is focused on the unsupervised learning algorithms used for the analysis, more specifically the *t-Distributed Stochastic Neighbour Embedding*, or *t-SNE* (van der Maaten and Hinton, 2008), for dimensionality reductions, and the *DBSCAN* (Ester et al., 1996) for clustering.

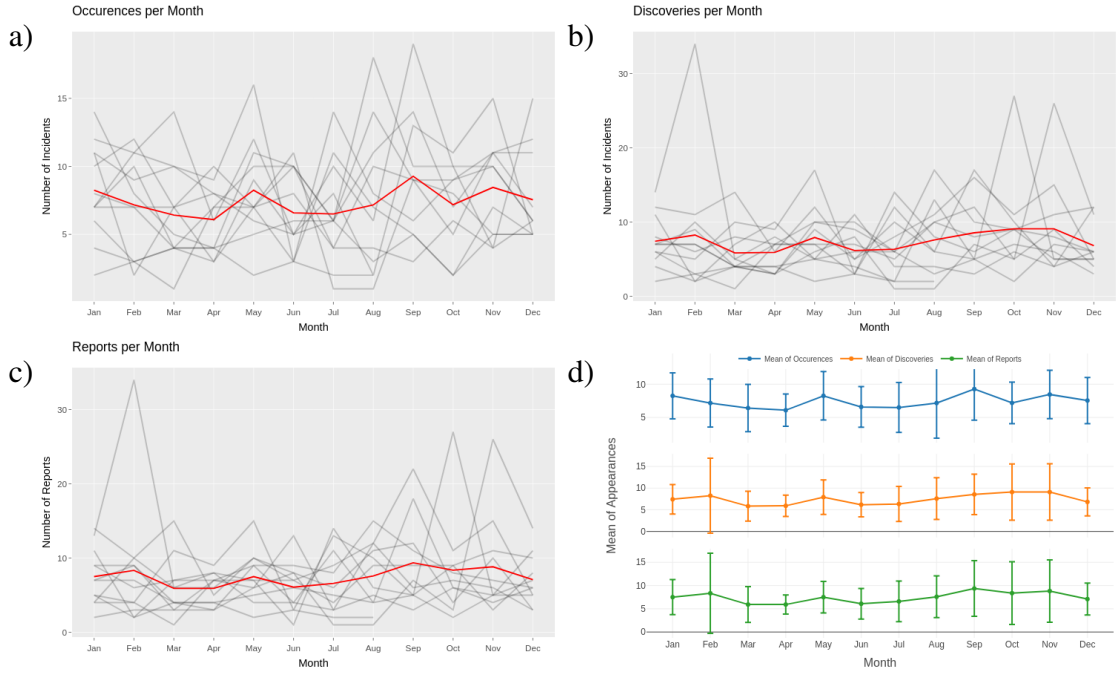

FIG. 7. Average of incidents per month for each year (gray) and the average per month for all the years (red), considering the a) occurrences, b) discoveries, and c) reports in the data set. d) are the averages for all the years with error bars.

t-SNE

t-Distributed Stochastic Neighbour Embedding (t-SNE) is a technique for dimensionality reduction to visualize high-dimensional data (van der Maaten and Hinton, 2008). The process consists on initially calculation, on the high-dimensional space, the probability $p_{j|i}$ of x_j be a neighbour of x_i for all the N observations (left part of the equation 1), given the bandwidth of the Gaussian kernel σ_i. The goal of the t-SNE methodology is to learn the similarity of the reduced d-dimensional observations y_i and y_j by preserving the probabilities $p_{j|i}$. The measure of similarity is calculated using the equation 1 (right):

$$p_{j|i} = \frac{\exp(||x_i - x_j||^2/\sigma_i^2)}{\sum_{k\neq l}\exp(-||x_i - x_k||^2/\sigma_i^2)} \iff q_{j|i} = \frac{(1+||y_i - y_j||^2)^{-1}}{\sum_{k\neq l}(1+||y_k - y_l||^2)^{-1}} \tag{1}$$

The dimensional reduced observation are estimated by minimizing the Kullback-Leibler divergence (equation 2):

$$C = \mathbf{KL}(P||Q) = \sum_{i\neq j} p_{ij}\log\left(\frac{p_{ij}}{q_{ij}}\right) \tag{2}$$

where:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N} \tag{3}$$

The minimization of the equation 2 is is achieved by the gradient descent equation 4:

$$\frac{\delta C}{\delta y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)(1 + ||y_i - y_j||^2)^{-1} \tag{4}$$

t-SNE, as a dimensionality reduction algorithm, tries to preserve the clusters (similar neighbours) on well defined groups in the low-dimension space. However it may not occur if a parameter called *perplexity* (which measures the quality of a probability distribution or probability model to predict a sample) is not set properly. In this project, we use the algorithm to visualize computed clusters on a 2-D scatter plot.

DBSCAN

Density-based spatial clustering of applications with noise, or simply *DBSCAN*, was proposed by Ester et al. (1996) as a non-parametric clustering algorithm, and it has a simple but powerful functionality. The user has to set a few parameters, the most important ones are the distance ϵ (maximum distance on what two points are neighbours), and the minimum number of neighbours $\min Pt$ for the cluster. The algorithm will set one random point of the data set as the center of the cluster, and starts to scan the points around it as neighbours until the conditions ϵ and $\min Pt$ are satisfied. When conditions fail, a point that was not marked as part of the previous cluster is set as a new cluster center and a new scan starts, now for another cluster. The algorithm repeat the process until there are no points that satisfy the conditions. The algorithm decides how many clusters are in the data, and points that are not included into any cluster is marked as outlier with the label 0.

The algorithm can fail to find different clusters with a high density data set, and is very hard to setup if the distances between the points are sparse.

Clustering and Visualizing the Incidents Reports

For the clustering step, 6 features were selected with information for the type of activity, equipment, inspection, maintenance, and substance that were in place during the incident. DBSCAN separates the data into clusters and we analyze if the clusters bring information about released volume of substance, how long it took for the incident/failure took to be discovered, and if the emergency level, that goes from 0 (no emergency) to 4 (disaster).

Figure 8 shows the t-SNE reduced dimensionality of the data (to 2 dimensions) coloured by the DBSCAN clusters. Clustering algorithm was done before the dimensionality reduction, the t-SNE was used only for visualization and to check if the clusters are reliable.

Analyzing Figure 8, the DBSCAN algorithm was able to find a total of 28 clusters, although most of them are small and contain a small number of points. The clusters are

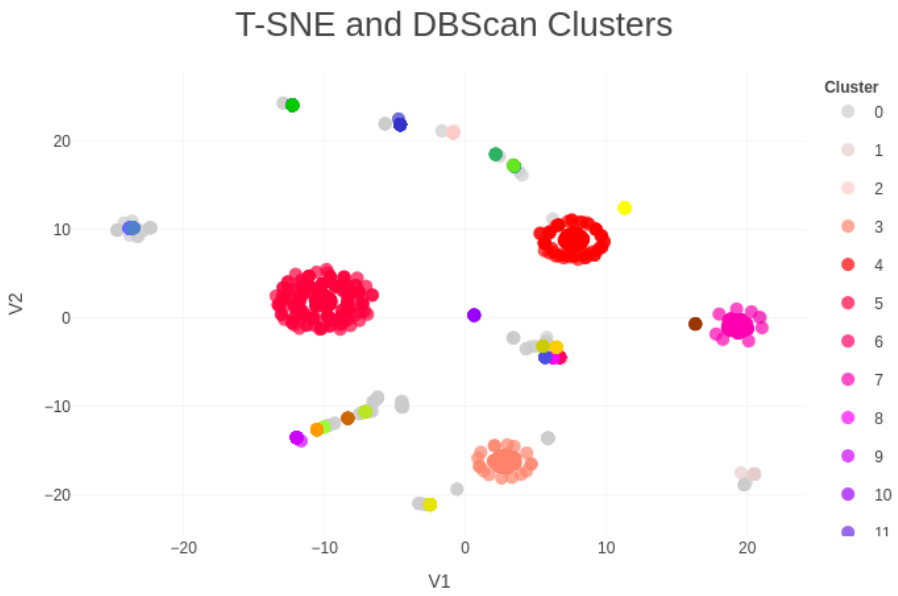

FIG. 8. t-SNE map of the data coloured by the DBSCAN clusters.

mostly far from each other, as t-SNE ended up providing a sparse representation of the points and clusters. There are 4 groups that immediately called our attention: the clusters 3, 4, 5, and 7, as they are larger and well represented in the t-SNE map.

Table 1. Most common clusters and their values.

DBScan	Count	Released.Vol.Mean	Disc.Diff.Mean	Emergency.Mean
3.00	107.00	0.00	1.44	1.02
4.00	147.00	0.00	0.48	0.99
5.00	237.00	48886.97	0.01	1.02
7.00	98.00	0.00	0.00	0.98

Table 1 summarizes the information for those clusters. The largest cluster is 5, with 237 occurrences, while cluster 7 is the smallest one (98 occurrences). Also, cluster 5 is the only one related to release of substance, with an average of $48,886.97m^3$ released. Cluster 3 contains the incidents that took longer to be discovered, with an average of 1.44 days. About the emergency level, the averages for all the cluster are virtually the same.

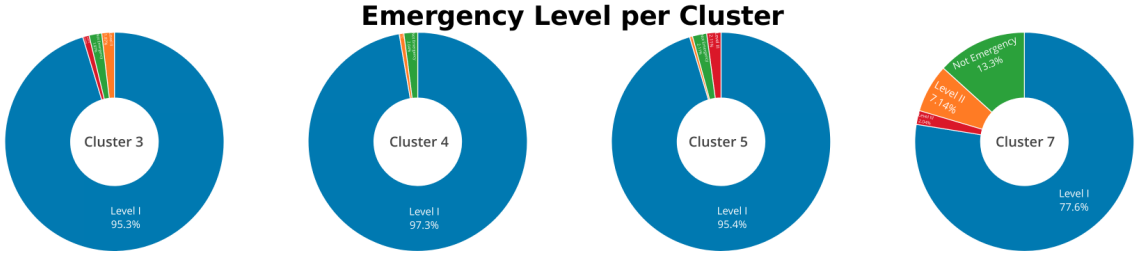

FIG. 9. Chart with the counts of emergency levels for the 4 largest clusters.

On Figure 9, the pie charts with the emergency level proportion for each cluster shows that *Level 1* is the most common type of emergency for all the clusters. Cluster 7 differs

from the others as having larger proportions of the emergencies 0 (no emergency) or 2. This means that incidents that falls into cluster 7 have a higher probability to be of a more dangerous emergency, as well not being an emergency at all.

By clustering the data set and analyzing the time series of the occurrences, we could have some insights about the pipeline incidents, and to identify abnormalities on reports the companies send to the government.

CONCLUSIONS

We analyzed Canada's pipeline incidents reports from 2008 to 2019. The data proved to be confusing and incomplete, but with a large number of useful information from which we could learn some patterns.

In the first part of the analysis, we learned that about half of the pipeline incidents reported resulted on some type of substance released, dominated by gas. Also, four companies are predominant on number of reports sent.

From the time series analysis, the number of incidents per year has no observed correlation with the average yearly production, while there is a correlation with the average price of the WTI crude oil for the year. The number of incidents tend to increase during elevations of the oil price, and to drop when the price reduces (more accentuated in 2008 and 2014, the start of the last two oil crisis). It was also observed that the number of occurrences and discoveries during the weekends. We could not observe any seasonal effect on the number of incidents, but we could identify outliers on the number of discoveries and reports. Each of these outliers are connected with a single company (Enbridge Pipelines Inc.), when discoveries and reports of several incidents were set to a same day and location.

We applied clustering methods for dimensionality reduction and cluster analysis. We could define 4 large clusters, where one of them (cluster 7) was associated to larger release of substances, another one (cluster 3) was associated to events that took longer to be discovered, while the cluster 4 was associated to incidents that can be of higher emergency, or no emergency at all.

ACKNOWLEDGEMENTS

The authors thank the sponsors of CREWES for continued support. This work was funded by CREWES industrial sponsors and NSERC (Natural Science and Engineering Research Council of Canada) through the grant CRDPJ 461179-13 and CRDPJ 543578-19, and the financial support from Canada First Research Excellence Fund. We thank Soane Mota dos Santos for the suggestions, tips and productive discussions.

REFERENCES

Badida, P., Balasubramaniam, Y., and Jayaprakash, J., 2019, Risk evaluation of oil and natural gas pipelines due to natural hazards using fuzzy fault tree analysis: Journal of Natural Gas Science and Engineering, **66**, 284 – 292.

Belvederesi, C., Thompson, M. S., and Komers, P. E., 2017, Canada's federal database is inadequate for the

assessment of environmental consequences of oil and gas pipeline failures: Environmental Reviews, **25**, No. 4, 415–422.

Belvederesi, C., Thompson, M. S., and Komers, P. E., 2018, Statistical analysis of environmental consequences of hazardous liquid pipeline accidents: Heliyon, **4**, No. 11, e00,901.
URL http://www.sciencedirect.com/science/article/pii/S2405844018356020

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X., 1996, A density-based algorithm for discovering clusters in large spatial databases with noise: 2nd International Conference on Knowledge Discovery and Data Mining, 226–231.

Green, K. P., and Jackson, T., 2015, Safety in the transportation of oil and gas: Pipelines or rail?: Fraser Research Bulletin by the Faser Institute, https://www.fraserinstitute.org/research/safety-transportation-oil-and-gas-pipelines-or-rail.

Michel, J., and Fingas, M., 2016, Oil Spills: Causes, Consequences, Prevention, and Countermeasures, chap. 7, World Scientific, 159–201.

van der Maaten, L., and Hinton, G., 2008, Visualizing data using t-sne: Journal of Machine Learning Research, **9**, 2579–2605.

Wang, D., Guo, W., Kong, S., and Xu, T., 2020, Estimating offshore exposure to oil spill impacts based on a statistical forecast model: Marine Pollution Bulletin, **156**, 111,213.