JSC370 - Final Project - Written Report

Mohsin Reza

2022-04-21

Introduction

The dataset I chose to analyze for this project contains information related to environmental spills in Ontario between 2003 and 2020 (inclusive). My initial research question was as follows:

What were the major sources, causes and consequences of environmental spills in Ontario between 2003 and 2020, and how did the number, sources, causes and consequences of environmental spills change across different years and locations?

However, after beginning the project, I believed that this question was too broad and I refined it to focus on just the sources and causes. I decided to choose this aspect as analyzing the main causes would allow us to determine how to best prevent environmental spills in the future (by mitigating the causes and risks). Therefore, my final research question is as follows:

What were the major sources and causes of environmental spills in Ontario between 2003 and 2020, and how did the number, sources and causes of environmental spills change across different years and locations?

In terms of background information on the research question, I found extensive research and publications on oil spills in particular, but found comparatively little research on environmental spills from other contaminants. The major consequences of oil spills, according to the chapter on fossil fuels in the World Scientific Series in Current Energy Issues, were human error and equipment failure. It estimated that 30-50% of oil spills were directly or indirectly caused by human error, and 20-40% were caused by equipment failure/malfunction. One interesting thing I found regarding my question is that a 2020 paper published on ScienceDirect claimed that the risk of spillage had increased due to oil production, exploration, and consumption. However, according to the chapter on fossil fuels in the World Scientific Series in Current Energy Issues, the risk of spillage had decreased in the past 10 years despite the increased production and consumption of oil. After conducting this background research, I was curious to see what my dataset would say regarding the consequences of environmental spills in Ontario.

Methods

The data were acquired from the Government of Ontario's open data catalogue, and the link to the dataset can be found here. This dataset was originally created by compiling incident reports received by the Spills Action Centre (SAC) in Ontario.

In terms of tools used for data exploration, I used the R programming language for all aspects of the analysis. The dplyr package, which is part of the tidyverse library, was used to clean and wrangle the data. Additionally, the kable package was used to create the summary tables, and the ggplot package was used to create the visualizations in this report. I also used the plotly package to create the interactive visualizations that can be found on the website. Finally, the rpart and randomForest packages were used for the machine learning parts of the analysis.

To analyze the data and answer the research question, I created several summary tables and visualizations, which can be seen in the results section. I also used the chi squared test to determine if the patterns I saw in the visualizations and tables were statistically significant. Finally, I used machine learning techniques including decision trees, random forests, bagging, and XGBoost to determine which factors were most important in predicting the cause of an environmental spill.

Several steps were taken to clean and wrangle the data appropriately. Firstly, a new "year" variable was created, which extracted the year from the "date reported" column. Secondly, a "year category" variable was created using the "year" variable, which had values "2003-2007", "2008-2012", "2013-2016", and "2017-2020" depending on which year the spill occurred in. Thirdly, I also converted all the contaminant names to sentence case, as there were many observations with the same contaminant name but in different cases. Fourthly, a new categorical variable called "location_type" was created. It took the value "city" if the site municipality was a city, and "outside city" if the site municipality was not a city. I used information from this site to determine if a municipality was considered a city or not. Fifthly, I also created a categorical variable called "cause" that took the value Human error and Equipment failure if the incident reason was one of these two and other if the incident reason was something else. This variable was created primarily for the ML algorithms, as without making the number of categories smaller the algorithms were taking too long to run. Finally, the number from the "health environment consequence" column was extracted and put in the newly created "consequence_score" column. In the case of missing values, for all variables, I decided to leave them as they were rather than imputing or removing them.

Results

When data was imported, I found that it had 109247 observations of 12 variables. Additionally, by checking the header and footer of the dataset, I determined that there were no import issues present. In terms of missing values, the consequence score variable had the most number of missing values with 60349. However, this is not a concern as the variable is not to be used in the analysis. Additionally, the incident reason variable had 29329 missing values, the source type variable had 7342 missing values, and the contaminant name variable had 3802 missing values. Additionally, the website mentions that the dataset is for environmental spills between 2003 and 2020 (inclusive). Since all the reported dates were dates between 2003 and 2020 (inclusive), I concluded that data errors were probably not present for the reported date. In the rest of the variables, which were all categorical, it was extremely difficult to tell if there were data errors present, as there was no set list of normal values for them found in the data dictionary.

Now, below are several visualizations and summary tables that attempt to answer the question of interest.

Table showing the top 20 contaminants by number of spills

Contaminant	Number of spills	Proportion of spills
Natural gas (methane)	22479	0.2058
Diesel fuel	11978	0.1096
Hydraulic oil	7296	0.0668
Transformer oil (n.o.s.)	4390	0.0402
Sewage,raw unchlorinated	4112	0.0376
Unknown / n/a	3802	0.0348
Furnace oil	3463	0.0317
Gasoline	3309	0.0303
Oil (petroleum based, not specified)	2271	0.0208
Refrigerant gas, n.o.s.	2234	0.0204
Motor oil	1380	0.0126
Mineral oil	1296	0.0119

Contaminant	Number of spills	Proportion of spills
Sewage, raw unchlorinated	1269	0.0116
Coolant n.o.s.	1223	0.0112
Smoke	1064	0.0097
Water	1042	0.0095
Methane gas, compressed (natural gas)	986	0.0090
Fuel oil	883	0.0081
Natural gas, compressed (methane)	841	0.0077
Sediment(suspended solids/ sand/ silt)	836	0.0077

This table shows us that natural gas was by far the biggest culprit in terms of causing the most environmental spills. It alone was responsible for around 20% of spills in Ontario. Diesel fuel and hydraulic oil were also contaminants for a large number of spills, with diesel fuel being the contaminant in almost 11% of spills and hydraulic oil being the contaminant in 6.7% of spills.

Table showing the top 20 causes of environmental spills in Ontario

Incident Reason	Number of spills	Proportion of spills
Operator/Human Error	23084	0.2113
Equipment Failure	19769	0.1810
Unknown - Reason not determined	5936	0.0543
Spill	5009	0.0459
Other - Reason not otherwise defined	3822	0.0350
Error- Operator error	2673	0.0245
Weather Conditions	2392	0.0219
Power Interruption/Loss	1429	0.0131
Material Failure	1372	0.0126
Negligence (Apparent) - Caused by lack of diligence	1262	0.0116
Other	1211	0.0111
Deliberate Act	1130	0.0103
Maintenance	918	0.0084
Equipment/Vehicles	908	0.0083
Blockage	898	0.0082
Weather	882	0.0081
Fire/Explosion	814	0.0075
Intentional Discharge	488	0.0045
Damage By Moving Equipment - Containers damaged by moving	454	0.0042
Vandalism - Illegal/deliberate (incl. sabotage)	409	0.0037

This table shows that operator/human error and equipment failure were the two biggest causes of environmental spills in Ontario. 21.13% of spills were caused by operator/human error, and 18.10% were caused by equipment failure. This shows us that perhaps the best way to prevent environmental spills is to purchase and maintain equipment more effectively, as well as give the personnel operating the equipment better training. Another interesting thing to note is that these results are in line with our background research, which mentioned that the top two causes of environmental spills were human error and equipment failure.

Table showing the top 20 cities with most spills per capita

City	Number of spills per capita	population
Dryden	0.0449	7749
Kenora	0.0378	15096
Thunder Bay	0.0182	107909
Timmins	0.0178	41788
Sarnia	0.0169	71594
Port Colborne	0.0154	18306
Temiskaming Shores	0.0137	9920
Quinte West	0.0129	43577
Sault Ste. Marie	0.0125	73368
Brockville	0.0125	21346
Pembroke	0.0122	13882
Barrie	0.0122	141434
Guelph	0.0116	131794
Cornwall	0.0114	46589
Thorold	0.0110	18801
Kingston	0.0109	123798
Kawartha Lakes	0.0109	75423
Greater Sudbury	0.0103	161531
North Bay	0.0101	51553
Orillia	0.0101	31166

This table shows us that Dryden and Kenora had by far the most number of spills per capita. One interesting thing that we can see in the table is that the top 20 cities with the highest number of spills per capita are mostly the cities with smaller populations.

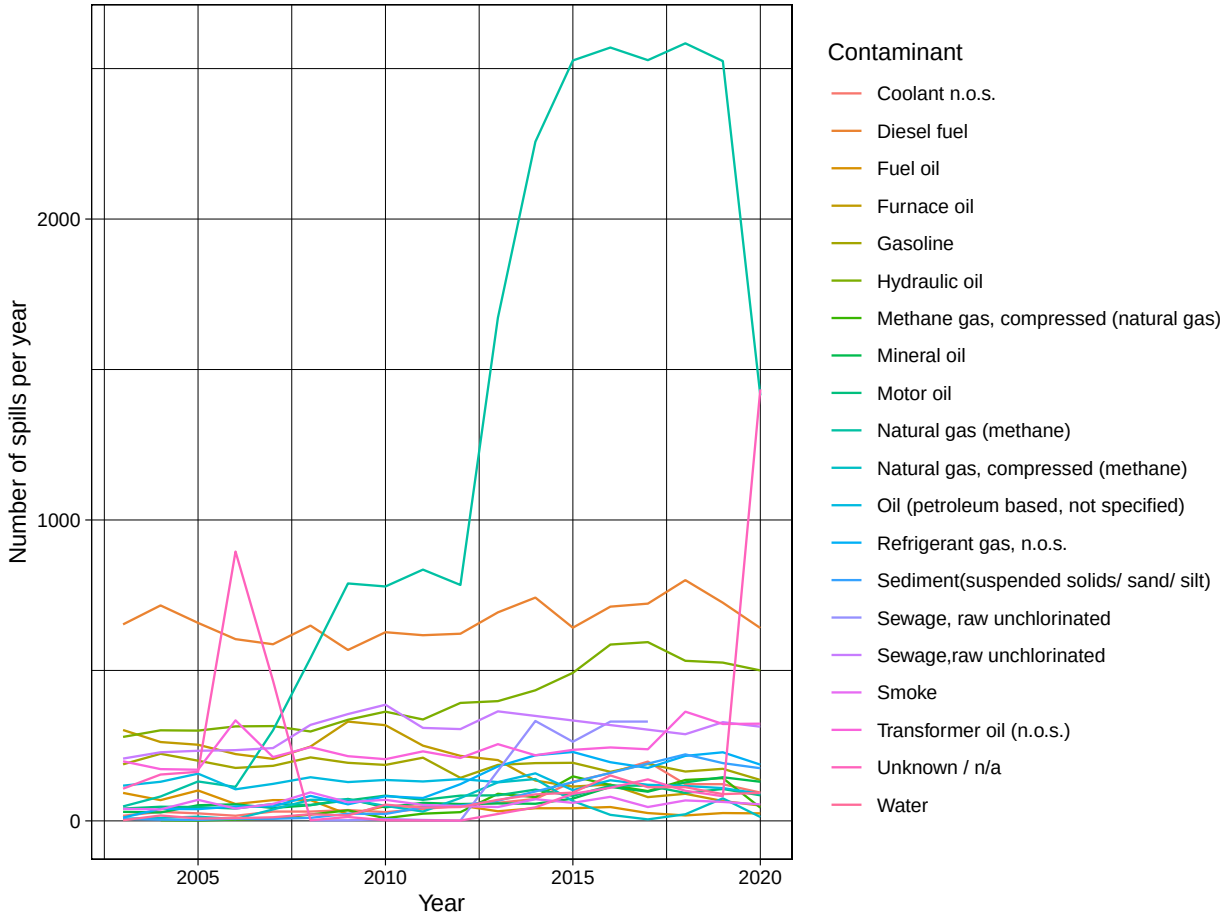
Table showing the years with the most spills

Year	Number of spills	Proportion of spills
2018	8255	0.0756
2016	8204	0.0751
2017	8170	0.0748
2015	8037	0.0736
2019	8034	0.0735
2020	7694	0.0704
2014	7512	0.0688
2013	6548	0.0599
2012	5246	0.0480
2011	5237	0.0479
2010	5207	0.0477
2009	5154	0.0472
2008	5067	0.0464
2006	4541	0.0416
2007	4450	0.0407
2005	4198	0.0384
2004	3961	0.0363
2003	3732	0.0342

From this table, we can see almost a clear trend that the number of spills has increased as the years have gone by. All of the last six years in the dataset make up the top six years with the most number of environmental spills.

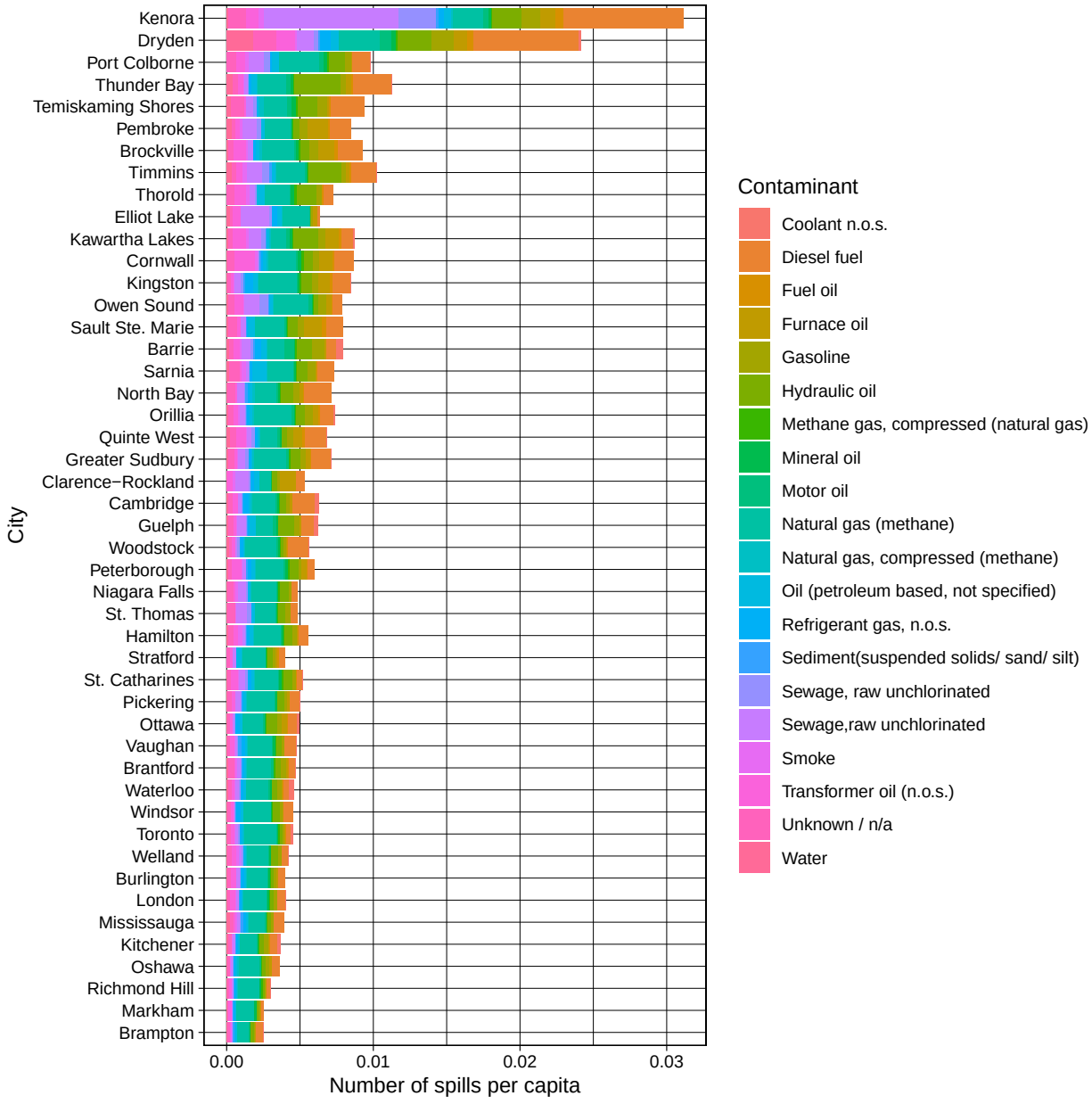
After having seen the main sources, causes, and locations of environmental spills, I then attempted to see whether there was any change in these sources and causes across different years and locations. Displayed below are some visualizations which will help answer this part of the research question.

Graph showing the number of spills per year for each of the top 20 contaminants



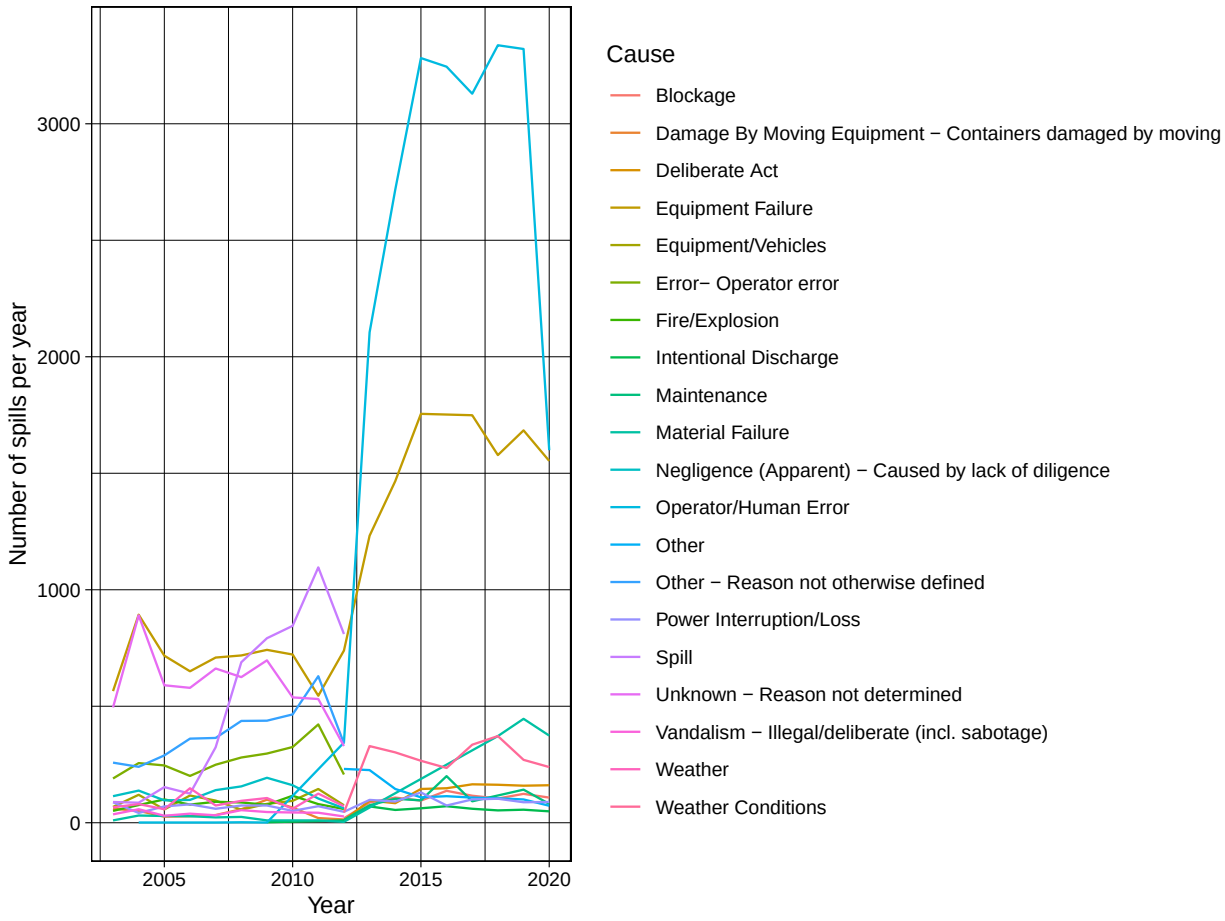
Right away, we can see in this plot that there was a huge jump in the number of spills per year caused by natural gas between 2012 and 2016, and then there was a large decline in the number of spills per year caused by natural gas between 2019 and 2020. I wonder why there was such a huge jump and then such a huge decline. There was also an increase in the number of spills per year caused by hydraulic oil. Otherwise, for the most part, the number of spills caused by other contaminants stayed consistent, with only minor fluctuations from year to year.

Graph showing the number of spills per capita for each of the 20 contaminants by city



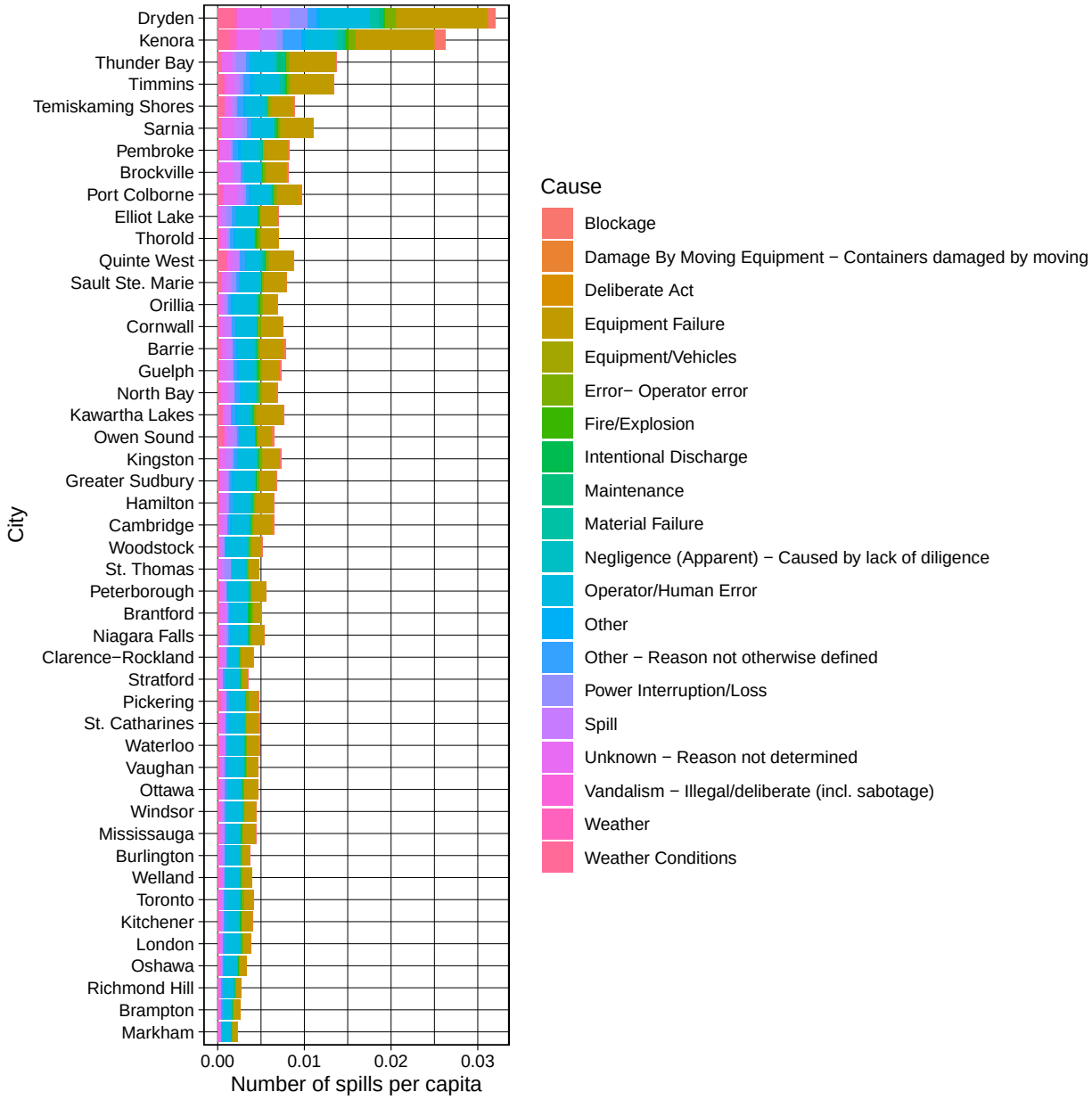
Overall, I cannot see any noticeable difference in the proportion of spills per capita for each of the top 20 contaminants across different cities. One interesting thing I can see, however, is that even though Kenora and Dryden have a much higher number of spills per capita overall compared to the other cities, they have a similar number of spills per capita for natural gas. Kenora has a much higher number of spills per capita for sewage, and both Kenora and Dryden have a higher number of spills per capita for diesel fuel compared to other cities.

Graph showing the number of spills per year for each of the top 20 causes



There are a lot of interesting patterns that we can see in this plot. Firstly, until around 2012, the biggest causes of environmental spills were either unknown or equipment failure, with human/operator error being in third place. However, from 2012 to 2015, there was a large increase in the number of environmental spills per year caused by human error, and human error far surpassed equipment failure to become the leading cause of environmental spills. Then, from 2019 to 2020, the number of spills per year caused by human error dropped significantly to be at a similar level to spills caused by equipment failure. I am interested to know what the reason was for such a drastic fluctuation.

Graph showing the number of spills for each of the top 20 causes by city



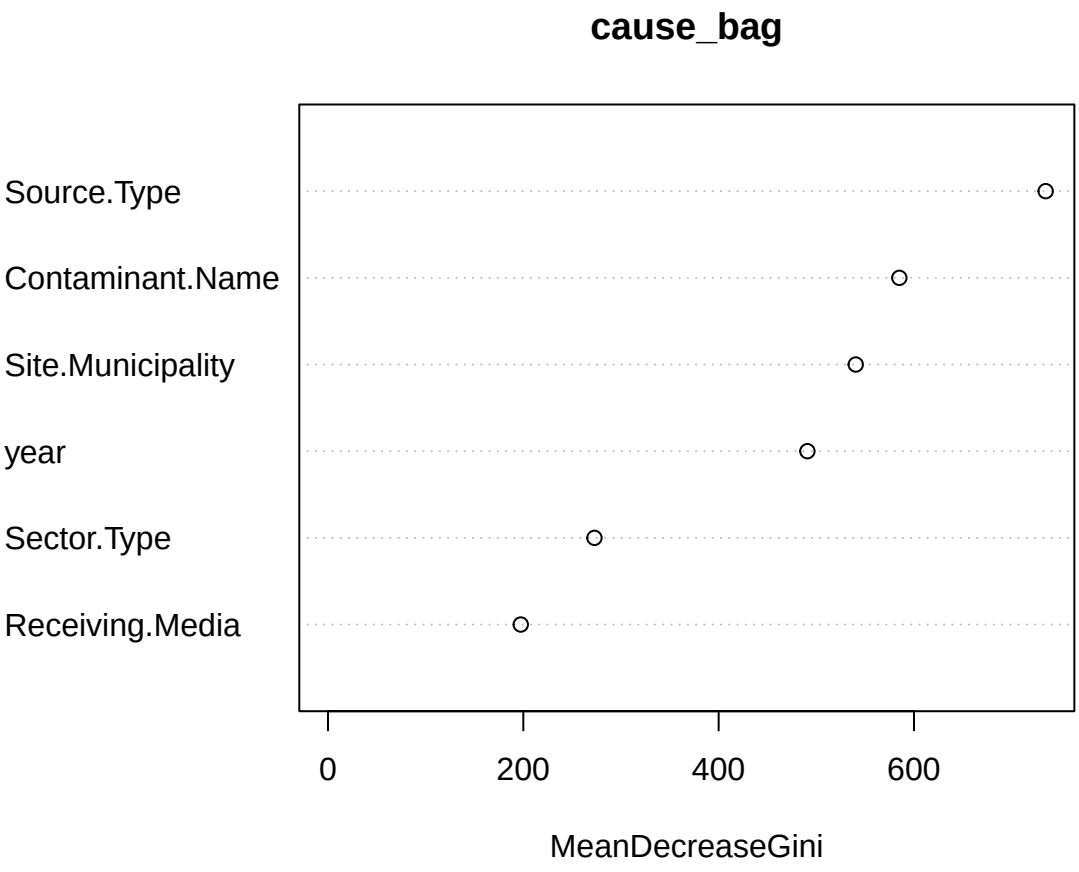
Overall, there are no noticeable differences in the proportion of spills per capita for each of the top 20 causes across different cities. Each city seems to have a similar proportion of spills per capita coming from the top two causes of human error and equipment failure, with a smaller proportion of spills per capita coming from weather, unknown, and other causes.

The final part of the analysis involved attempting to predict the cause of an environmental spill based on a number of factors. This would be especially helpful as in a future incident, it would be possible to make an educated prediction of the cause and take action accordingly while awaiting the results of a full scale investigation. To do this, I implemented a classification tree, bagging, and random forest algorithm to predict the cause of an environmental spill based on various factors. I wanted to use gradient boosting and XGBoost in addition, but unfortunately my dataset was quite large (109247 observations) and the computation was taking too long. A 70-30 training test split was used for all three algorithms. As mentioned in the methods section, to facilitate the ML algorithms, a new cause variable was created which took the value “Human

Error/Equipment Failure" if the cause of the incident was one of these two, and other if the cause was something else. These two causes were by far the most common, accounting for around 39% of incidents, and therefore they encapsulate a good portion of the data. The results for the three algorithms are shown in the table below:

ML Algorithm	Proportion of test correctly classified	Proportion of test misclassified	OOB error
Classification tree	0.7567	0.2433	6009.0
Bagging	0.7477	0.2523	123.2
Random Forest	0.7480	0.2520	116.6

When I fit the aforementioned three algorithms to predict the cause of an environmental spill, I found that all three algorithms performed fairly well, with an approximately 74-75% accuracy. There was no signifcant difference in the three algorithms in terms of accuracy, but the classification tree had a significantly higher out of bag error, which I found quite surprising. At first I believed that it had been computed incorrectly, but I do not believe that this is the case after having checked the computation over. In terms of variable importance, both bagging and random forests gave similar results. The source type variable was by far the most important predictor of the cause, followed by the type of contaminant, the year of incident, the location of the site, the sector (industry), and the receiving media (land, air, or water). The variable importance plot for bagging only is shown below (it is very similar for random forests):



Conclusions and Summary

Recall that my research question was:

What were the major sources and causes of environmental spills in Ontario between 2003 and 2020, and how did the number, sources and causes of environmental spills change across different years and locations?

Based on the analysis conducted, I found that the major sources of environmental spills in Ontario were natural gas, diesel fuel, hydraulic oil, transformer oil, and sewage. The notable change seen in the major sources over time was that natural gas became an increasingly common contaminant over time, but then saw a sharp decline in the number of spills per year it caused from 2019 to 2020. Also, natural gas caused a similar number of spills per capita across all cities even though Kenora and Dryden had a much larger number of spills per capita overall, while the number of spills per capita caused by sewage and diesel fuel was much higher and Kenora and Dryden compared to the other cities. In terms of the causes, I found that operator/human error and equipment failure were by far the biggest causes of environmental spills, and there was no notable change in the causes across locations. However, across time, the number of spills per year caused by human error increased significantly and far surpassed the number of spills per year caused by equipment failure as the years went by. Finally, when attempting to predict the cause of an environmental spill using machine learning techniques, I found that bagging, classification trees, and random forests performed similarly well in terms of their misclassification rate, but classification trees had a significantly higher out-of-bag error. The source of the spill and the contaminant were the most important variables in predicting the cause of the spill.