JSC370 - Midterm Report

Mohsin Reza

2022-03-10

Introduction

The dataset I chose contains information related to environmental spills in Ontario between 2003 and 2020 (inclusive). My research question is 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?

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.

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. 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.

The code used to perform the wrangling steps described above is shown below:

```r
cities <- c("Barrie", "Bellevile", "Brampton", "Brantford", "Brockville",
            "Burlington", "Cambridge", "Clarence-Rockland", "Cornwall", "Dryden",
            "Elliot Lake", "Greater Sudbury", "Guelph", "Hamilton",
            "Kawartha Lakes", "Kenora", "Kingston", "Kitchener", "London",
            "Markham", "Mississauga", "Niagara Falls", "North Bay", "Orillia",
            "Oshawa", "Ottawa", "Owen Sound", "Pembroke", "Peterborough",
            "Pickering", "Port Colborne", "Quinte West", "Richmond Hill",
            "Sarnia", "Sault Ste. Marie", "St. Catharines", "St. Thomas",
            "Stratford", "Temiskaming Shores", "Thorold", "Thunder Bay",
            "Timmins", "Toronto", "Vaughan", "Waterloo", "Welland", "Windsor",
            "Woodstock")

data <- data %>%
  mutate(Date.Reported = as.Date(Date.Reported, format = "%Y/%m/%d"),
         Contaminant.Name = str_to_sentence(Contaminant.Name),
         year = as.numeric(format(as.Date(Date.Reported, format = "%Y/%m/%d"), "%Y")),
         year_category = ifelse(year <= 2007, "2003-2007",
                         ifelse(year <= 2012, "2008-2012",
                         ifelse(year <= 2016, "2013-2016", "2017-2020"))),
         location_type = ifelse(Site.Municipality %in% cities, "city", "outside city"),
         consequence_score = as.numeric(substring(Health.Environmental.Consequence, 1, 1)))
```

Preliminary Results

To answer our research question, we focused on the following variables:

- Date Reported
- Site Municipality
- Contaminant Name
- Source Type
- Incident Reason
- Year Category
- Location Type
- Consequence Score

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. 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. For the consequence score variable, I was able to determine that the range for this variable was 0 (no impact) to 6 (major impact on human health) according to the data dictionary found here. Since all the consequence scores in my dataset were between 0 and 6, I concluded that data errors were probably not present for this variable. 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, displayed below are summary statistics for each variable 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

Contaminant	Number of spills	Proportion of spills
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
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	17598	0.1611
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
Equipment Failure - Malfunction of system components	2171	0.0199
Material Failure <96> Poor Design/Substandard Material	1384	0.0127
Negligence (Apparent) - Caused by lack of diligence	1262	0.0116
Other	1211	0.0111
Material Failure - Poor Design/Substandard Material	1192	0.0109
Deliberate Act	1130	0.0103
Maintenance	918	0.0084
Equipment/Vehicles	908	0.0083
Blockage	898	0.0082
Weather	882	0.0081
Fire/Explosion - Resulting from fires/explosions (Not occurrences which cause a fire or explosion)	814	0.0075
Power Interruption/Loss	786	0.0072
Power Interruption - Loss of electrical power	643	0.0059

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 16.11% 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.

Table showing summary statistics for the consequence score variable

Mean	Sd	Median
1.901	0.6146	2

This table shows us that on average, the consequence score for an environmental spill was around 2, which means that the spill caused minor environmental damage (according to the data dictionary).

Table showing the top 20 municipalities with most spills

Municipality	Number of spills	Proportion of spills
Toronto	16553	0.1515
Ottawa	6219	0.0569
Hamilton	5218	0.0478
Mississauga	4750	0.0435
Brampton	2222	0.0203
London	2137	0.0196
Vaughan	2017	0.0185
Thunder Bay	1963	0.0180
Barrie	1723	0.0158
Greater Sudbury	1669	0.0153
Guelph	1533	0.0140
Windsor	1402	0.0128
Kingston	1348	0.0123
Kitchener	1324	0.0121
Cambridge	1230	0.0113
Sarnia	1208	0.0111
Oakville	1074	0.0098
Markham	1024	0.0094
Burlington	988	0.0090
Chatham-Kent	955	0.0087

This table shows us that Toronto had by far the most number of environmental spills, with a total of 16553. We can also see in the table that the top 20 municipalities with the most spills are all classified as cities.

Table showing the years with the most spills

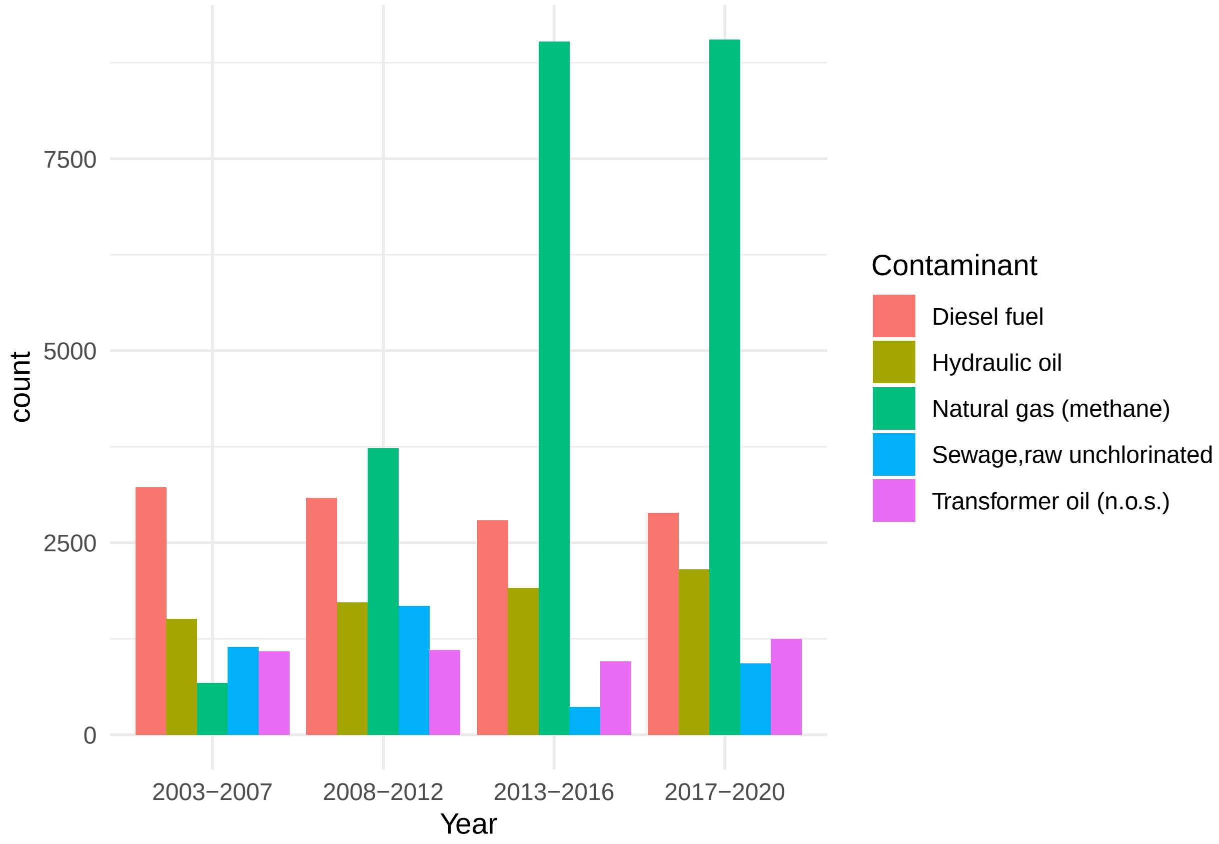
Year	Number of spills	Proportion of spills
2018	8255	0.0756
2016	8204	0.0751

Year	Number of spills	Proportion of spills
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.

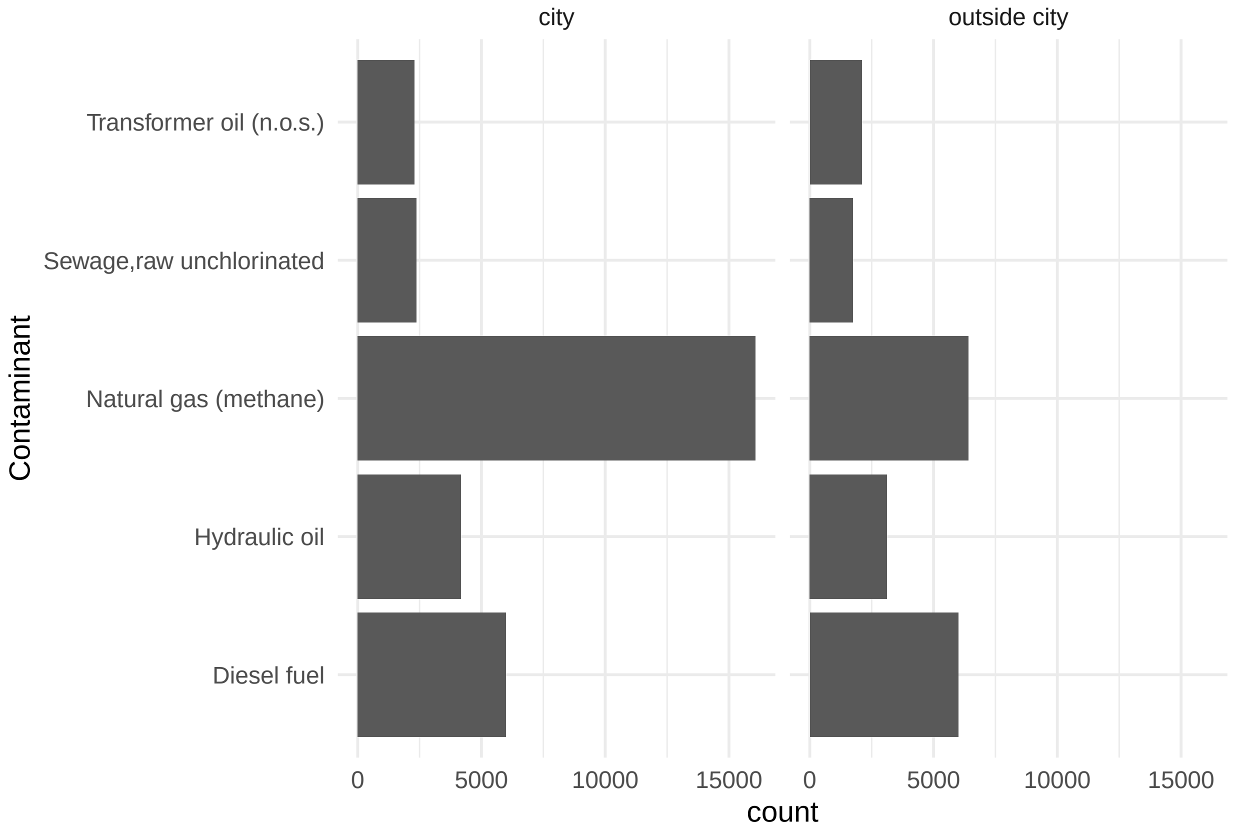
Now, displayed below are some visualizations which will help answer the question of interest.

Graph showing the number of spills for each of the top 5 contaminants by year category



One interesting thing we can see in this plot is that as the years have gone by, natural gas has become an increasingly common contaminant and has accounted for a greater number of spills. Otherwise, the number of spills with diesel fuel, hydraulic oil, sewage, and transformer oil have stayed relatively consistent with only minor fluctuations.

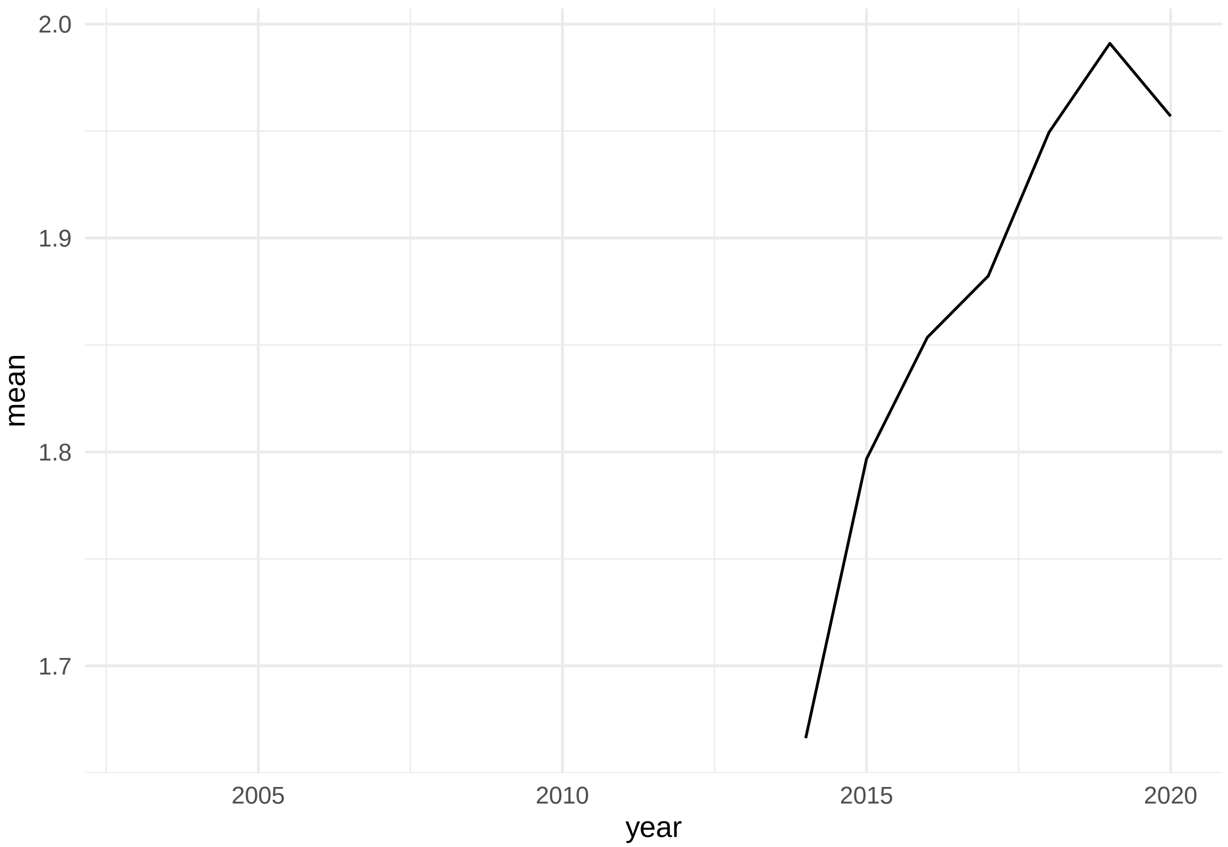
Graph showing the number of spills for each of the top 5 contaminants by location type



From this graph, one interesting thing we can conclude is that in cities, natural gas is the contaminant for a much greater number of spills than diesel fuel. However, outside cities, both substances were contaminants for a similar number of spills. We can also see that the other substances have a fairly similar number of spills inside and outside cities.

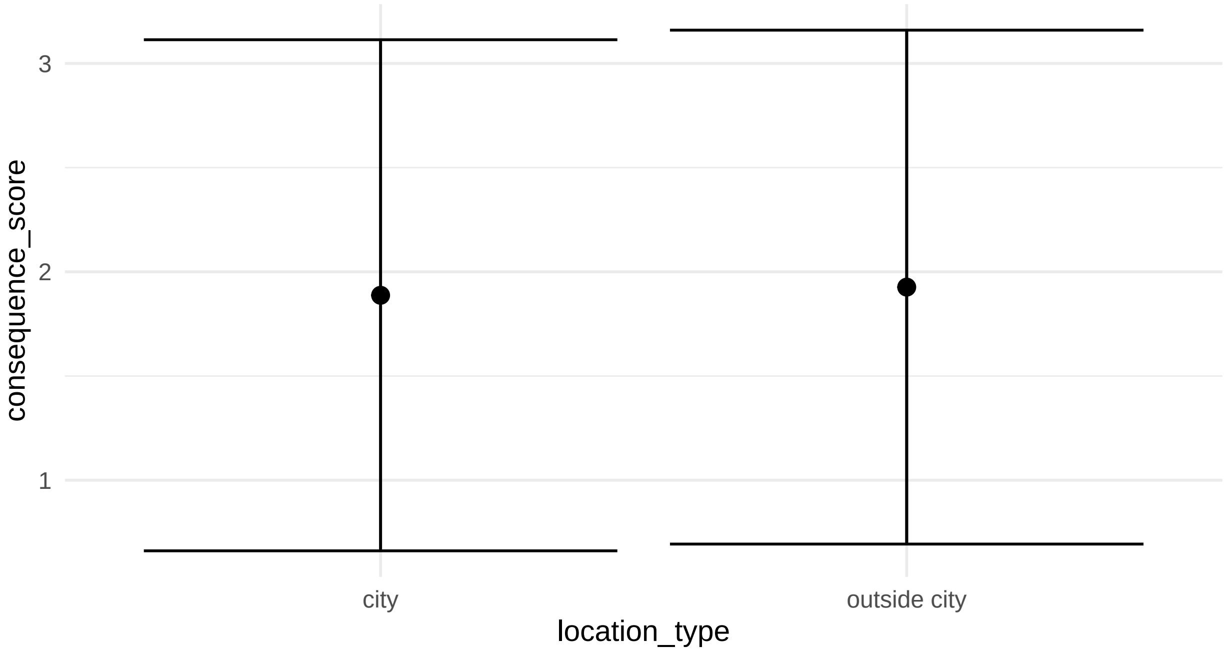
Time series plot showing the mean consequence score by year

```
## Warning: Removed 11 row(s) containing missing values (geom_path).
```



Unfortunately, the consequence score is only available for spill data since 2014. However, based on the data that we have, we can see that there was an increase in the mean consequence score as the years have gone by. This means that on average, environmental spills had more severe consequences as the years went by. The exception to this was 2019-2020, where the mean consequence score decreased. I suspect this may have been due to the lockdowns and less human activity.

Plot showing the mean and sd of consequence score by location type



From this plot, we can see that the severity of environmental spills was fairly similar both inside and outside cities.

Conclusion

Recall that our research question was:

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?

Based on our analysis thus far, we found that the major sources of environmental spills in Ontario were natural gas, diesel fuel, hydraulic oil, transformer oil, and sewage. The only notable change swe saw in the major sources was that natural gas became an increasingly common contaminant over time. Also, while diesel fuel and natural gas were similarly common outside cities, natural gas was much more common in cities. In terms of the causes, we found that operator/human error and equipment failure were the biggest causes of environmental spills, and there was no notable change in the causes across time and locations. Finally, we saw that for consequences, on average, most spills resulted in low environment damage. This damage increased on average from 2014-2019, but decreased in 2020. It was similar across location types.