HW #4: Analyze real-world emissions data

Hayden Atchley

2022-12-02
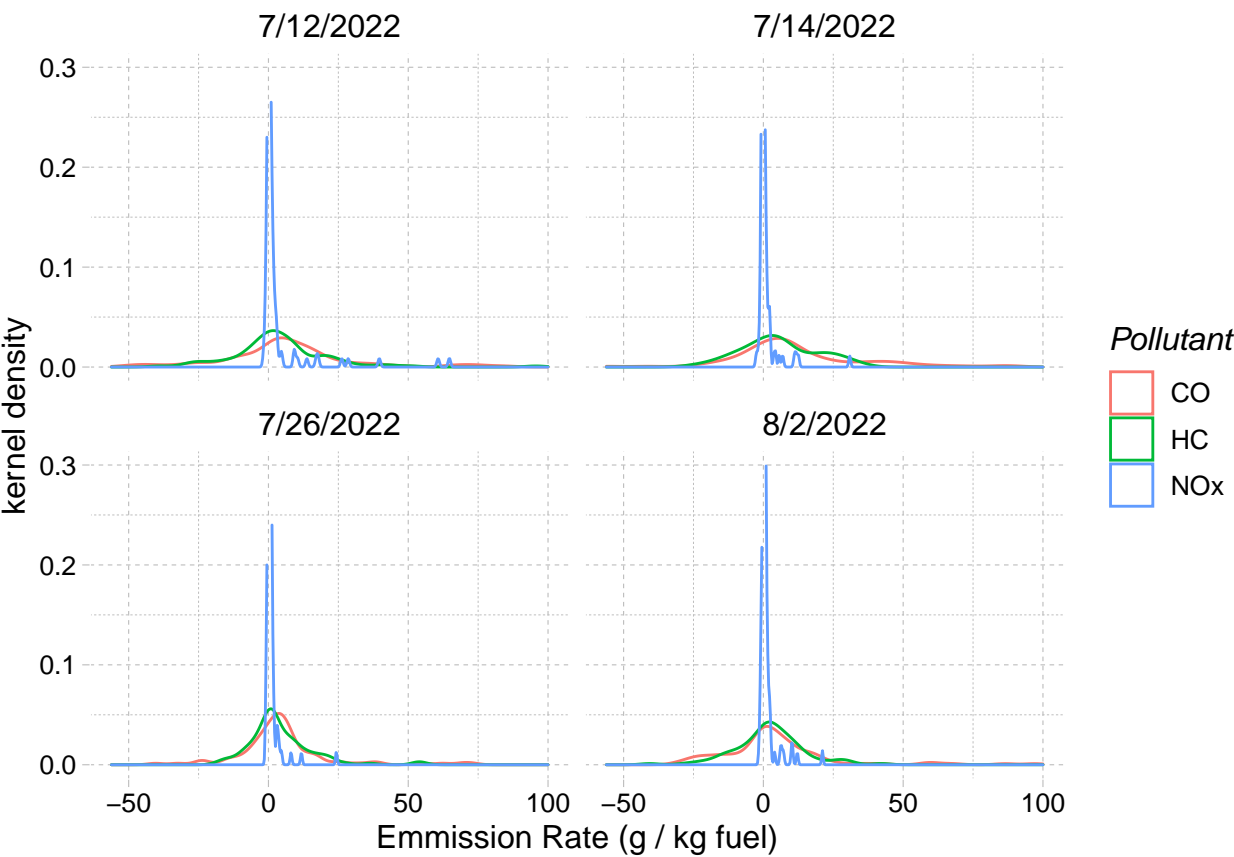
1

Figure 1: Emissions density by date.

1.1

location	pollutant	max_emisions	median_emissions	max / median
Timp Hwy East	CO	177.80	3.36	52.90
Timp Hwy East	HC	53.99	1.62	33.31
Timp Hwy East	NOx	24.31	0.50	48.38
Timp Hwy West	CO	803.31	1.59	504.24
Timp Hwy West	HC	42.14	1.89	22.25
Timp Hwy West	NOx	21.11	0.42	49.83
Univ Ave	CO	452.85	7.19	63.00
Univ Ave	HC	247.66	3.18	77.96
Univ Ave	NOx	64.70	0.25	259.12

1.2

Looking at Figure 1, NOx appears to have quite a skewed distribution, though CO has a few extreme outliers.

2
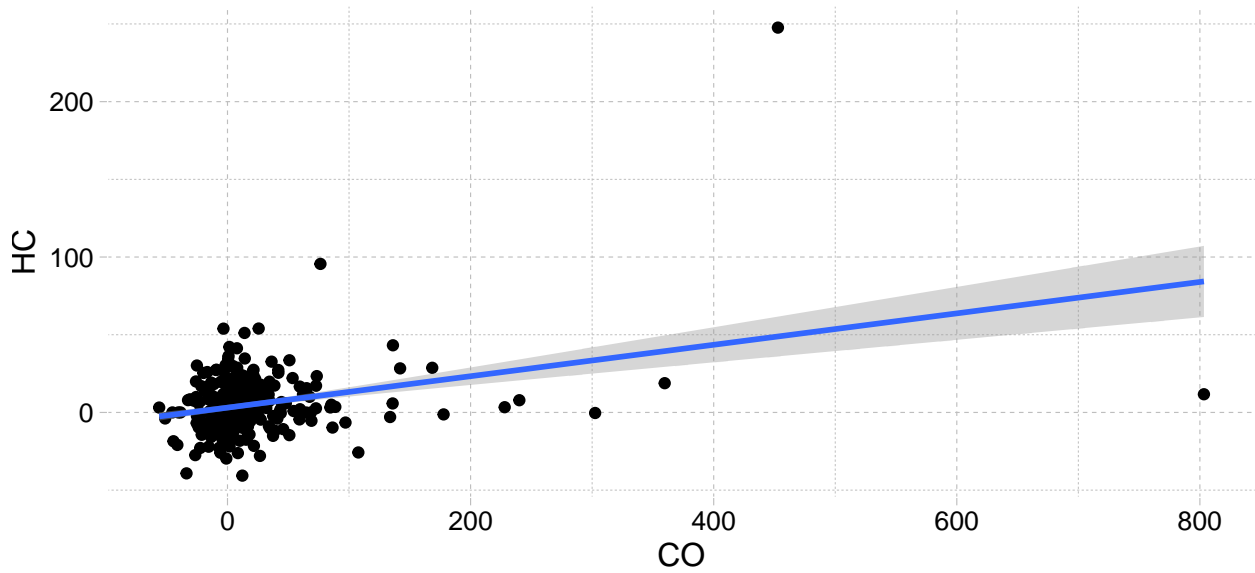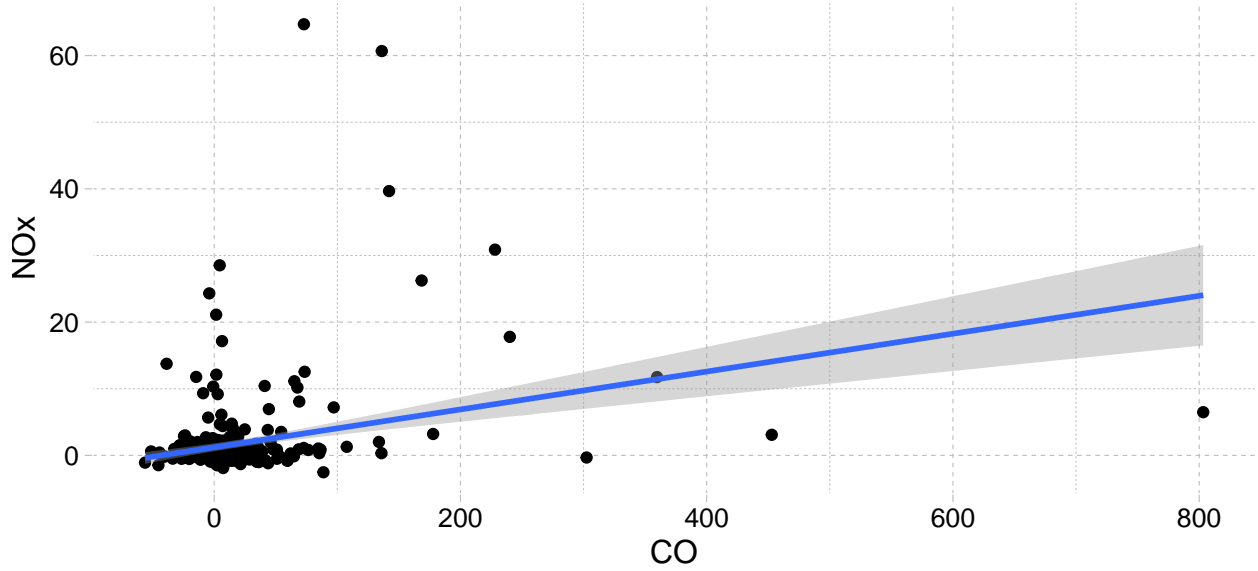
term	estimate	std.error	statistic	p.value
(Intercept)	6.74	2.91	2.32	0.021
NOx	2.46	0.46	5.38	0.000
HC	0.96	0.15	6.44	0.000
$R^2 = 0.162$				

While it appears that vehicles with more CO emissions also have more NOx and HC emissions (both of these slopes/coefficients are positive), the R^2 value is quite low. There could be many other factors explaining the variance in emission rates.

3 (3.1)

The theory behind I/M programs is good, in that vehicles with high emissions are kept from operation. However, as mentioned, this creates issues with equity, where these programs impact lower income individuals more than higher income individuals. Perhaps one of the main reasons for this is due to the age of the vehicles owned based on income. Newer vehicles are more fuel efficient as well as lower-emitting than older vehicles (at least as a rule), especially as emissions standards continue to get stricter. Newer vehicles are also significantly more expensive than older vehicles, and so are less likely to be owned by lower income individuals.

But this does not wholly discount the benefit of I/M programs. In fact, I'd argue these programs (or at least the concept of them) are a good thing so long as equity can be addressed properly. Perhaps funds that would have gone to I/M programs could be used as a subsidy for low-income individuals to purchase more climate-friendly vehicles. Though this doesn't address the question of what to do with high-emitting vehicles directly, increasing the gas/diesel tax would discourage driving as much (lowering emissions), and then the excess funds could again be used as subsidies for low- to no-emitting vehicles.

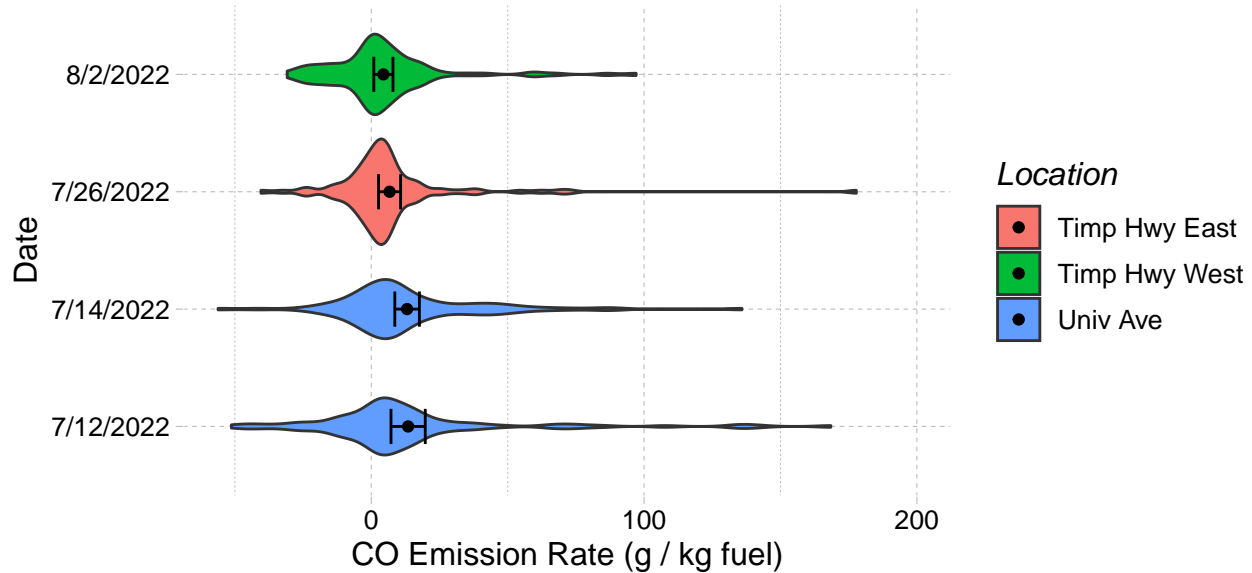| DATE	location	mean	lwr.ci	upr.ci
7/12/2022	Univ Ave	20.16	9.93	30.4
7/14/2022	Univ Ave	21.63	12.32	30.9
7/26/2022	Timp Hwy East	6.71	2.66	10.7
8/2/2022	Timp Hwy West	12.43	-1.29	26.1

Figure 2: CO emission rate for each day. The bars show the mean and 90% confidence interval for the mean.

The table below shows a Tukey-Kramer test for multiple variance based on emissions test location. None of the p-values indicate significant results, so we can't conclude that the means are different from each other.

```
## # A tibble: 3 x 5
##   contrast                 estimate conf.low conf.high adj.p.value
##   <chr>                       <dbl>    <dbl>     <dbl>       <dbl>
## 1 Timp Hwy West-Timp Hwy East    5.72   -14.2      25.6       0.778
## 2 Univ Ave-Timp Hwy East        14.2    -3.05      31.4       0.130
## 3 Univ Ave-Timp Hwy West         8.47   -8.77      25.7       0.481
```

4.1

Assuming the re-sampling method refers to a permutation/randomization test, the main advantage is that no assumptions need to be made regarding skewness, normality, or outliers. The main disadvantage is that it is computationally intensive to compare all permutations of the data points, especially when the data sets are large. Often a smaller subset of permutations is performed, which can usually offer a good approximation of the full permutation test.

5

6

Because there were multiple data points at each location, and multiple locations, the data points are not necessarily all random and independent. The measured values at each location are more likely to correlate with each other for any number of potential reasons, including measurement techniques, outside air quality, geographic distribution of vehicle types and sizes, etc. In short, there are many potential variables that a given location would have constant, but could affect the measured emissions. Therefore, each of the daily means are *not* from the same population, so shouldn't directly be compared between locations.

However, the mean emissions at each *location* is in fact random and independent, as now the effect these previously-mentioned variables can be captured, that is, the effect on the *location* means. We therefore take the mean emissions at each location:

Location	Mean Emissions
American Fork	10.2
Orem	9.7
Payson	31.8
Pleasant Grove	8.2
Provo 2	18.5
Springville	25.3
Timp Hwy	9.6
Univ Ave	20.9

and can now find the mean of *these* values to estimate the mean emission rate for all of Utah County (with 95% confidence interval):

mean	lwr.ci	upr.ci
16.7	9.4	24.1

6.1

The random, independent variable is as previously mentioned the mean CO emission rate at each *location*. These values can be assumed to be independent and random, but the *daily* emission rates cannot, as the rates at each location will be correlated with each other.