Addendum to an Analysis of the Role of Air Pollution in the Spread of COVID-19

Richard Strouss-Rooney
Drexel University
Richard.J.Strouss@drexel.edu

Additional Data Collection

The previously described data covering COVID-19 morbidity and mortality, environmental and demographic features, and EPA pollutant monitoring data were fortified with additional demographic and economic data from the 2019 American Community Survey conducted by the US Census Bureau. This was accessed through the Census Bureau API [1]. The entire dataset was able to be accessed with one api call. The following fields were requested:

- Count from the survey of those who moved into the county from in-state within the last year, those who moved from out-of-state, those who moved from abroad, and the total sample size for the migration statistics
- Count from the survey of those who used public transportation and the total sample size for transportation statistics
- Number of households , households with two or more members, family households, and non-family households with one or more residents 65+ years of age as well as the total household sample size
- Gini coefficient and mean hours typically worked in a week
- Average housing units in a structure
- Count of those with and without health insurance as well as the total sample size for health insurance statistics

These data included such things as public transportation use, immigration between counties, states, and nations, average housing units in a building, and health insurance coverage, all at the county level. While historic data such as this cannot empirically tell us the facts during our period of interest, it is reasonable to presume, for instance, that counties with high inward migration would continue to have such in 2020 and 2021 relative to other counties despite the unusual circumstances. To be useful, the census data had to be merged with both the COVID-19 and the pollution datasets. This was a relatively easy merge on FIPS codes and dates though, of course, there was a fair bit of data wrangling required.

It is only natural that counties with higher population would also have more of practically everything which would to lead finding spurious correlations. To prevent this, much of the data was converted to rates. This allowed us to find relationships between changes in proportions. Special care had to be taken to obtain the correct reference totals for each category to ensure valid rates were obtained.

Two distinct datasets were created with this merger: one that crossed the new census data with the data retrieved from *C3.ai* and another that crossed the census, COVID-19, and EPA data. This was done because only a small portion of counties have an EPA monitoring site; this allowed us to look at demographic and economic features across all counties as well as isolating counties with EPA sites.

To deal with mismatches in dates between the COVID-19 and EPA data, weekly averages were calculated for all variables for each county. This had the additional benefit of reducing the size of the dataframe by a factor of seven.

Exploratory Data Analysis

Two analyses were conducted to explore the feature space within these two datasets. The period of interest for all analyses was from 2020-03-11 to 2021-03-11, therefore the first step for both datasets was to reduce the data to that date period. All analyses only look at correlations, not causal relationships.

Census and COVID-19 Data

The first and most general analysis involved taking mean values across the entire period for each feature and county. Aggregating over a whole year this way may hide seasonality or other factors that occur differentially over time. However, it is a good starting point to guide deeper exploration. After aggregation, correlations for all features were calculated using the Pearson correlation coefficient. Table 1 shows selected correlations with COVID-19 mortality rates, selected either due to relatively strong or to surprisingly small

[1] https://api.census.gov/data/2019/acs/acs5.html

	Deaths
Avg. Daily Temp	0.20
Gini Coeff.	0.18
Staffed ICU Beds	0.17
Public Transportation	0.13
Households 65+	0.08
Population Density	0.05

Table 1: Selected correlations with COVID-19 mortality rates

correlation (see table 3 for all correlations). The average daily temperature over the one-year period had the strongest correlation ($r = 0.20, p < 0.001$). This is mildly surprising as one might expect cold weather to drive more people in-doors in close proximity to one another. While there are certainly many possible explanations and many other variables for which to control, one possibility is that retirees moving from the north to warmer locales may have driven up the proportion of the population aged 65 years or more. That proportion is weakly correlated with COVID-19 mortality ($r = 0.08, p < 0.001$) and more strongly correlated with temperature ($r = 0.14, p < 0.001$).

The Gini coefficient, a measure of income distribution, is nearly as correlated with COVID-19 mortality as daily temperatures. ($r = 0.18, p < 0.001$). However, the Gini coefficient and average daily temperatures are, relative to most of the correlations, rather strong ($r = 0.32, p < 0.001$). Income distribution might well be an indicator of likely COVID-19 severity but temperature and other correlating features must be controlled for to make any certain conclusions.

The number of staffed ICU beds per capita showed a moderate correlation with COVID-19 mortality ($r = 0.17, p < 0.001$). It was also correlated with the Gini coefficient ($r = 0.20, p < 0.001$) although no statistically significant correlation with temperature could be found. This was a slightly odd finding since the Gini coefficient correlated more strongly with COVID-19 deaths than did median household income ($r = -0.11, r < 0.001$). Rates of public transportation usage was more weakly correlated with COVID-19 mortality ($r = 0.13, p < 0.001$) and, as expected, strongly correlated with population density ($r = 0.73, p < 0.001$). So, it was surprising to find that population density itself has only a very weak correlation with COVID-19 mortality ($r = 0.05, p = 0.009$) and with much less certainty than most of the figures (though would still be considered statistically significant in many circles).

EPA and COVID-19 Data

Previous analysis explored the relationships between multiple variables and COVID-19 mortality rates at the national level as well as the relationship between pollutants and COVID-19 mortality within counties. It was of interest to examine possible relationships between counties, as well.

	NO_2	$PM_{2.5}$	Ozone
COVID-19 Mortality	0.25	0.42	-0.05*
Avg. Daily Temp	0.19	0.23	-0.30
Gini Coeff.	0.35	0.10	-0.20
Staffed ICU Beds	0.06*	0.05*	-0.18
Public Transportation	0.43	0.08*	-0.10
Households 65+	0.49	0.03*	-0.02*
Population Density	0.41	0.04*	-0.14

* $p > 0.05$

Table 2: Correlations with NO_2, $PM_{2.5}$, Ozone, and Lead

First, correlation matrices were calculated separately for each pollutant to see how the variables correlated between counties aggregated across the whole period of analysis. By and large, the same features rose the top of the correlation matrix as did for the whole nation's covid data. To identify potential difference between the average American county as well as to investigate the relationship between levels of each pollutant and COVID-19 mortality, we looked at the same features for each pollutant as were identified as of interest in the census and covid data. Table 2 shows the correlation coefficients of the pollutants under investigation with COVID-19 mortality rates and the features of interest identified in the previous section.

NO$_2$ NO_2 had a relatively strong correlation with COVID-19 mortality ($r = 0.25, p = 002$). However, NO_2 also correlated as or more strongly with most of the other features of interest. NO_2 appears to be correlated with features common to urban environments with dense populations (see table 4. Considerable care will be required to sufficiently isolate NO_2 for causal analysis in the future.

PM$_{2.5}$ $PM_{2.5}$ had quite a strong correlation with COVID-19 mortality rates ($r = 0.42, p < 0.001$). This was the strongest correlation of any pollutant to COVID-19 mortality by far. Of the features of interest other than mortality, average daily temperatures correlated most strongly with $PM_{2.5}$ ($r = 0.23, p < 0.001$). It was somewhat surprising to see that population density did not have a statistically significant correlation with $PM_{2.5}$ ($p = 0.52$) in our data as we had presumed a relationship with dense, urban environments. See table 5 for the full table of correlations.

Ozone Ozone was correlated with several of the features of interest (see table 6). However, only a very weak correlation was found with COVID-19 mortality and it failed to reach statistical significance ($r = -0.05, p = 0.139$).

Lead Lead had no statistically significant correlations with any of our features of interest including COVID-19 mortality rates.

Conclusions and Future Work

Similar to other findings in the project, these analyses found evidence of relationships between pollutants and COVID-19 mortality rates, particularly NO_2 and $PM_{2.5}$. We found some pollutants to have a stronger relationship with COVID-19 mortality rates than most social, demographic, or economic features we investigated.

This research was limited by the reliance on correlations for analysis. While this method proved insightful, it could be improved upon with methods to isolate target features and control for strongly correlated, confounding features.

With these results, we can strongly recommend further research into the effects of air pollution on rates of COVID-19 morbidity and mortality.

	COVID-19 Mortality	Median Income	Temperatures	Gini	Staffed ICU Beds	Public Transit	Households w/65+	Pop Density
COVID-19 Mortality	1.000	-0.105	0.197	0.184	0.172	0.133	0.080	0.049
Median Income	-0.105	1.000	-0.224	-0.372	-0.150	0.284	0.276	0.286
Temperatures.data	0.197	-0.224	1.000	0.315	0.018	-0.029	0.137	0.054
Gini	0.184	-0.372	0.315	1.000	0.196	0.136	0.135	0.131
Staffed ICU Beds	0.172	-0.150	0.018	0.196	1.000	0.007	-0.013	0.059
Public Transit	0.133	0.284	-0.029	0.136	0.007	1.000	0.388	0.733
Households w/65+	0.080	0.276	0.137	0.135	-0.013	0.388	1.000	0.388
Pop Density	0.049	0.286	0.054	0.131	0.059	0.733	0.388	1.000

Table 3: All counties Pearson correlation matrix

	COVID-19 Mortality	NO_2	Median Income	Temperatures	Gini	Staffed ICU Beds	Public Transit	Households w/65+	Pop Density
COVID-19 Mortality	1.00	0.25	0.19	0.01	0.07	0.05	0.23	0.10	0.08
NO_2	0.25	1.00	0.16	0.19	0.35	0.06	0.43	0.49	0.41
Median Income	0.19	0.16	1.00	-0.03	-0.20	-0.32	0.25	0.14	0.20
Temperatures.data	0.01	0.19	-0.03	1.00	0.30	0.01	-0.00	0.28	0.09
Gini	0.07	0.35	-0.20	0.30	1.00	0.38	0.33	0.29	0.39
Staff ICU Beds	0.05	0.06	-0.32	0.01	0.38	1.00	0.02	-0.00	0.12
Public Transit	0.23	0.43	0.25	-0.00	0.33	0.02	1.00	0.25	0.88
Households w/65+	0.10	0.49	0.14	0.28	0.29	-0.00	0.25	1.00	0.24
Pop Density	0.08	0.41	0.20	0.09	0.39	0.12	0.88	0.24	1.00

Table 4: NO_2 Pearson correlation matrix

	COVID-19 Mortality	$PM_{2.5}$	Median Income	Temperatures	Gini	Staffed ICU Beds	Public Transit	Households w/65+	Pop Density
COVID-19 Mortality	1.00	0.42	0.13	0.25	0.32	0.20	0.59	0.26	0.46
$PM_{2.5}$	0.42	1.00	0.06	0.23	0.10	0.05	0.08	0.03	0.04
Median Incom	0.13	0.06	1.00	-0.06	-0.17	-0.20	0.21	0.23	0.18
Temperatures.data	0.25	0.23	-0.06	1.00	0.30	0.16	0.02	0.32	0.15
Gini	0.32	0.10	-0.17	0.30	1.00	0.40	0.31	0.29	0.38
Staffed ICU Beds	0.20	0.05	-0.20	0.16	0.40	1.00	0.07	0.04	0.18
Public Transit	0.59	0.08	0.21	0.02	0.31	0.07	1.00	0.30	0.81
House w/65+	0.26	0.03	0.23	0.32	0.29	0.04	0.30	1.00	0.34
Pop Density	0.46	0.04	0.18	0.15	0.38	0.18	0.81	0.34	1.00

Table 5: $PM_{2.5}$ Pearson correlation matrix

	COVID-19 Mortality	$PM_{2.5}$	Median Income	Temperatures	Gini	Staffed ICU Beds	Public Transit	Households w/65+	Pop Density
COVID-19 Mortality	1.00	-0.05	0.13	-0.01	0.11	0.09	0.25	0.15	0.11
Ozone	-0.05	1.00	-0.00	-0.30	-0.20	-0.18	-0.10	-0.02	-0.14
Median Income	0.13	-0.00	1.00	-0.10	-0.24	-0.24	0.29	0.22	0.26
Temperatures	-0.01	-0.30	-0.10	1.00	0.24	0.07	-0.05	0.19	0.06
Gini	0.11	-0.20	-0.24	0.24	1.00	0.38	0.27	0.27	0.29
Staffed Hospital Beds	0.09	-0.18	-0.24	0.07	0.38	1.00	0.04	0.02	0.08
Public Transit	0.25	-0.10	0.29	-0.05	0.27	0.04	1.00	0.32	0.84
Households w/65+	0.15	-0.02	0.22	0.19	0.27	0.02	0.32	1.00	0.35
Pop Density	0.11	-0.14	0.26	0.06	0.29	0.08	0.84	0.35	1.00

Table 6: Ozone Pearson correlation matrix