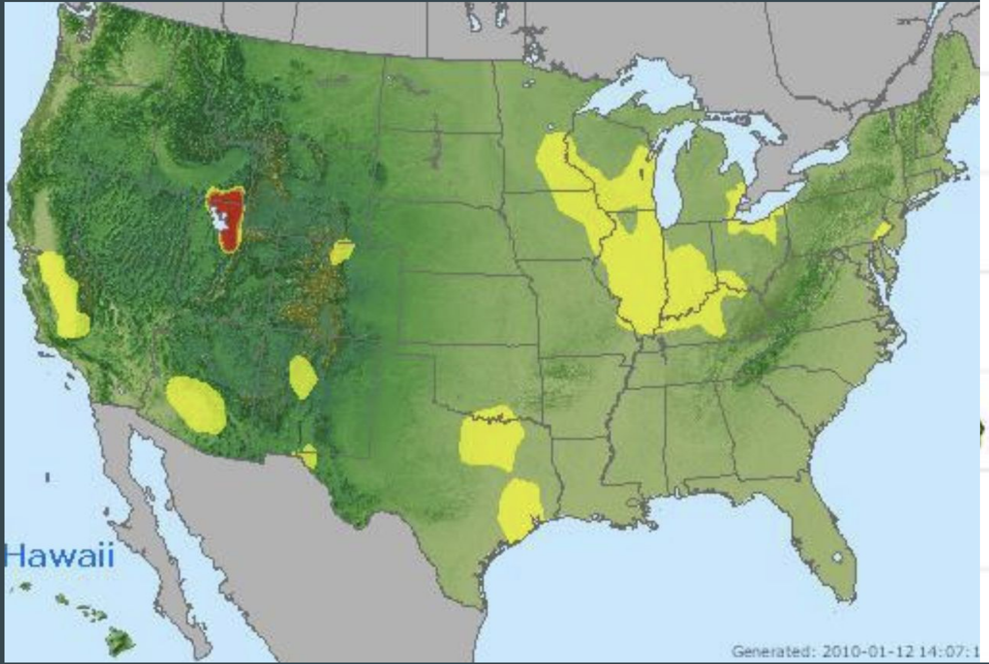
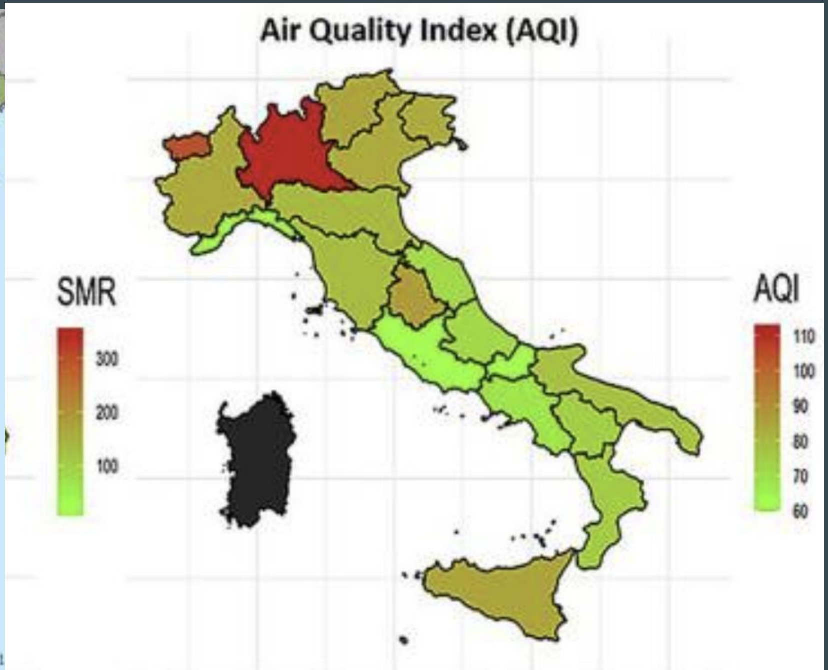
The Wranglers
COVID-19 & Air Quality

2021/4/28

Teigen Judd, Jon Barton, Yi-Jin Chen, Adriana Reyes-Miranda

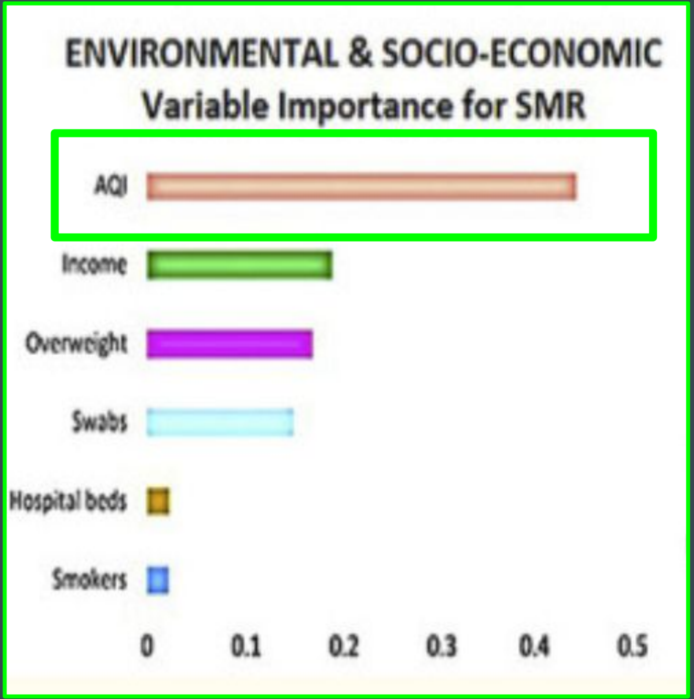
COVID-19 & Air Quality

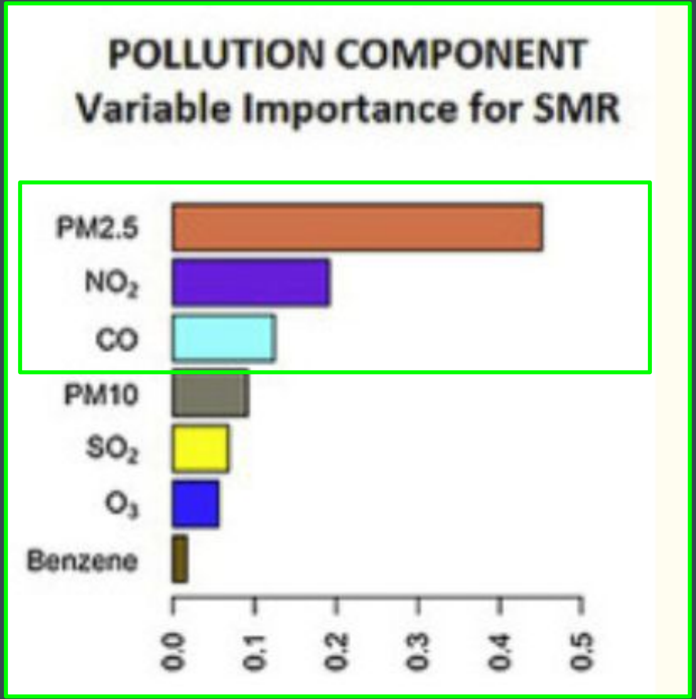




Problem

The original study used machine learning methods to reveal the prolonged exposure to air pollution associated with SARS-CoV-2 in Italy. [1]





Original Description of the Data - Air Quality

EPA https://aqs.epa.gov/aqsweb/documents/data_mart_welcome.html

In the beginning:

https://www.epa.gov/outdoor-air-quality-data/download-daily-data

Final (provided by Professor Naomi Riches):

https://aqs.epa.gov/aqsweb/airdata/download_files.html

Original Description of the Data - COVID-19

- Provided by John's Hopkins
  - https://github.com/CSSEGISandData
- In an aggregated format, with reporting down to county level.

- Data reporting began whenever individual counties began reporting their COVID data

COVID Original Data

date	county	state	fips	cases	deaths
5/1/2020	Snohomish	Washington	53061	2466	108
5/2/2020	Snohomish	Washington	53061	2492	108
5/3/2020	Snohomish	Washington	53061	2737	108
5/4/2020	Snohomish	Washington	53061	2784	110
5/5/2020	Snohomish	Washington	53061	2807	110
5/6/2020	Snohomish	Washington	53061	2830	112
5/7/2020	Snohomish	Washington	53061	2889	114
5/8/2020	Snohomish	Washington	53061	2917	114
5/9/2020	Snohomish	Washington	53061	2917	114
5/10/2020	Snohomish	Washington	53061	2932	116
5/11/2020	Snohomish	Washington	53061	2970	118
5/12/2020	Snohomish	Washington	53061	2998	119
5/13/2020	Snohomish	Washington	53061	3009	119

Data Quality Report - Air Quality (Original Data)

Lack of CO Data

- Less than 50% of counties per state with CO data

Unreasonable and Context-Inconsistent Data

- Negative Sensor Values
- Not States:
  (DC, Puerto Rico)

End-date Mismatch

- Ozone only recorded to Nov 14
- NO2 and PM2.5 both recorded to Oct 31

Data Quality Report - COVID-19 (Original Data)

- ➢ Data had to be switched from aggregated totals to daily numbers
- ➢ Some values were negative (possible mis-reporting), these were set to zero
- ➢ Some values were extremely high, we clipped these values down
- ➢ Data reporting didn’t start on the same date for every county
- ➢ Some dates had null values, we used the interpolate function to replace null values with nearest date value from same county

Air Quality Data - Wrangling Steps (PM2.5 as example)

- Subset variables

```
#subset
pm25 = pm25[["Arithmetic Mean","State Name", "County Name","AQI","County Code"]]
```

- Subset based on desired dates

```
datemask = pm25.loc['2020-05-01':'2020-12-31']
print(datemask['Arithmetic Mean'].describe())
```

```
count    157486.000000
mean          7.616998
std           8.080942
min          -4.913043
25%           4.425000
50%           6.400000
75%           9.000000
max         576.600000
Name: Arithmetic Mean, dtype: float64
```

Air Quality Data - Wrangling Steps (PM2.5 as example)

- Drop Not-states

```
datemask = datemask[~(datemask["State Name"]== 'District Of Columbia')]
datemask = datemask[~(datemask["State Name"]== 'Virgin Islands')]
```

- Impute negative values with 0

```
#Impute negative values in Arithmetic Mean with 0

pm25['Arithmetic Mean'] = pm25['Arithmetic Mean'].apply(lambda x : x if x > 0 else 0)
```

COVID-19 Data -Wrangling Steps

- Because this was the larger, more complete dataset, we adapted this dataset to merge well with the AQ data
- Unique merge id of county/state/date was created
- Had to ensure all county names matched between datasets
  - New York boroughs, Alaska boroughs, abbreviations, and capitalization of different counties/states made this more difficult
  - E.g. - St. Clair and Saint Clair would cause the mergeID to fail, so we had to adjust to the AQ data format

Pre-Merge COVID Data

5/4/2020	Snohomish	Washington	53061	2784	110	Snohomish, Washington	47	2
5/5/2020	Snohomish	Washington	53061	2807	110	Snohomish, Washington	23	0
5/6/2020	Snohomish	Washington	53061	2830	112	Snohomish, Washington	23	2
5/7/2020	Snohomish	Washington	53061	2889	114	Snohomish, Washington	59	2
5/8/2020	Snohomish	Washington	53061	2917	114	Snohomish, Washington	28	0
5/9/2020	Snohomish	Washington	53061	2917	114	Snohomish, Washington	0	0
5/10/2020	Snohomish	Washington	53061	2932	116	Snohomish, Washington	15	2
5/11/2020	Snohomish	Washington	53061	2970	118	Snohomish, Washington	38	2
5/12/2020	Snohomish	Washington	53061	2998	119	Snohomish, Washington	28	1
5/13/2020	Snohomish	Washington	53061	3009	119	Snohomish, Washington	11	0
5/14/2020	Snohomish	Washington	53061	3048	121	Snohomish, Washington	39	2
5/15/2020	Snohomish	Washington	53061	3065	125	Snohomish, Washington	17	4
5/16/2020	Snohomish	Washington	53061	3071	125	Snohomish, Washington	6	0

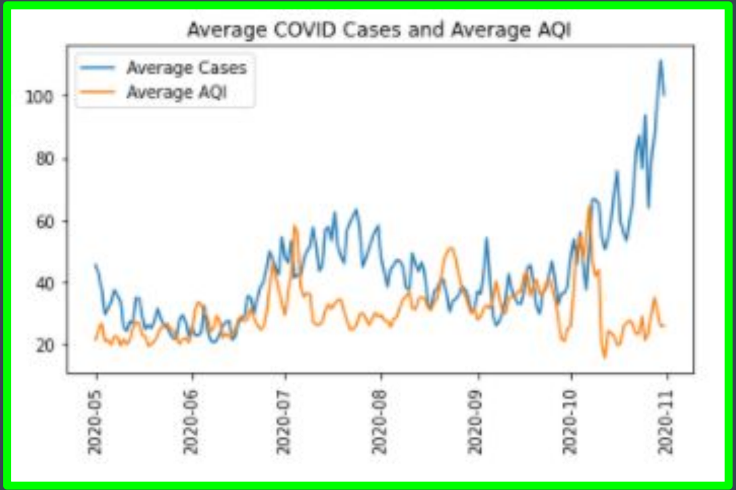
Report of the quality of merged data

The merge went well, only a little over 700 rows were left with null values

- ➢ These could all be accounted for
- ➢ Most often: AQ data was present, but COVID data had not started reporting yet
- ➢ There were some counties that did not begin reporting COVID data until August.

The merge also highlighted the known shortcomings in the AQ data

- ➢ Multiple metrics had not been reported at the time of data gathering



Merged Data

Date	State	County	Arithmeti	AQI_Ozon	Arithmeti	AQI_No2	Arithmeti	AQI	Daily_Cas	Daily_Deaths
5/1/2020	Alabama	Baldwin	35	50					1	1
5/1/2020	Alabama	DeKalb	47	58					0	0
5/1/2020	Alabama	Elmore	28	44					2	0
5/1/2020	Alabama	Etowah	34	51					0	1
5/1/2020	Alabama	Jefferson	30.33333	47.66667	15.31155	32	11.0125	46	43	2
5/1/2020	Alabama	Madison	35.5	55.5					0	0
5/1/2020	Alabama	Mobile	40	58					42	6
5/1/2020	Alabama	Montgom	27	47					18	0
5/1/2020	Alabama	Morgan	39	61					3	0
5/1/2020	Alabama	Russell	31	49			6.1125	25	3	0
5/1/2020	Alabama	Shelby	31	47					0	1
5/1/2020	Alabama	Sumter	24	47					4	1
5/1/2020	Alabama	Tuscaloos	25	46					2	1
5/1/2020	Alaska	Denali	44	44						
5/1/2020	Alaska	Fairbanks	29	34			4.733333	19.66667	1	0
5/1/2020	Arizona	Cochise	48	50					0	0
5/1/2020	Arizona	Coconino	48	50					12	2
5/1/2020	Arizona	Gila	53	67					0	0
5/1/2020	Arizona	La Paz	46	51			3.8375	16	1	0
5/1/2020	Arizona	Maricopa	42	56.69565	13.14417	24	6.546759	27	184	2
5/1/2020	Arizona	Navajo	51	58					29	0
5/1/2020	Arizona	Pima	40.25	48.75	5.00625	9.5	4.582065	19	26	1
5/1/2020	Arizona	Pinal	46	58.2			9.225	37.5	20	2
5/1/2020	Arizona	Yavapai	48	49					3	0

Data by AQ Metric (Ozone)

Date	State	County	Arithmeti	AQI_Ozon	Arithmeti	AQI_No2	Arithmeti	AQI	Daily_Cas	Daily_Deaths
5/1/2020	Alabama	Baldwin	35	50					1	1
5/1/2020	Alabama	DeKalb	47	58					0	0
5/1/2020	Alabama	Elmore	28	44					2	0
5/1/2020	Alabama	Etowah	34	51					0	1
5/1/2020	Alabama	Jefferson	30.33333	47.66667	15.31155	32	11.0125	46	43	2
5/1/2020	Alabama	Madison	35.5	55.5					0	0
5/1/2020	Alabama	Mobile	40	58					42	6
5/1/2020	Alabama	Montgom	27	47					18	0
5/1/2020	Alabama	Morgan	39	61					3	0
5/1/2020	Alabama	Russell	31	49			6.1125	25	3	0
5/1/2020	Alabama	Shelby	31	47					0	1
5/1/2020	Alabama	Sumter	24	47					4	1
5/1/2020	Alabama	Tuscaloos	25	46					2	1
5/1/2020	Alaska	Denali	44	44						
5/1/2020	Alaska	Fairbanks	29	34			4.733333	19.66667	1	0
5/1/2020	Arizona	Cochise	48	50					0	0
5/1/2020	Arizona	Coconino	48	50					12	2
5/1/2020	Arizona	Gila	53	67					0	0

Linear Regression

Sample Linear Regression

In [45]:

```
# trying to predict AQI by looking at covid cases and deaths
AQILR = sm.ols(formula="AQI ~ Daily_Cases + Daily_Deaths", data=filteredMergedData).fit()
AQILR.summary()
```

Out[45]:

OLS Regression Results

Dep. Variable:	AQI	R-squared:	0.008
Model:	OLS	Adj. R-squared:	0.008
Method:	Least Squares	F-statistic:	172.1
Date:	Mon, 26 Apr 2021	Prob (F-statistic):	3.73e-75
Time:	21:16:23	Log-Likelihood:	-1.7105e+05
No. Observations:	40261	AIC:	3.421e+05
Df Residuals:	40258	BIC:	3.421e+05
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	28.6426	0.090	318.889	0.000	28.467	28.819
Daily_Cases	0.0117	0.001	18.539	0.000	0.010	0.013
Daily_Deaths	-0.0748	0.010	-7.469	0.000	-0.094	-0.055

Omnibus:	26109.874	Durbin-Watson:	0.866
Prob(Omnibus):	0.000	Jarque-Bera (JB):	683859.655
Skew:	2.715	Prob(JB):	0.00
Kurtosis:	22.447	Cond. No.	169.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

This R-squared is extremely low, showing little-to-no predictability.

Predicting AQI with COVID cases/deaths

Very low R-squared (low correlation)

Linear Regression

```
# trying to predict daily deaths by looking at all AQI indicators
AQILR = sm.ols(formula="Daily_Deaths ~ AQI + AQI_No2 + AQI_Ozone", data=filteredMergedData).fit()
AQILR.summary()
```

OLS Regression Results

Dep. Variable:	Daily_Deaths	R-squared:	0.033
Model:	OLS	Adj. R-squared:	0.033
Method:	Least Squares	F-statistic:	200.1
Date:	Mon, 26 Apr 2021	Prob (F-statistic):	1.28e-127
Time:	21:16:23	Log-Likelihood:	-71364.
No. Observations:	17727	AIC:	1.427e+05
Df Residuals:	17723	BIC:	1.428e+05
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.6108	0.297	2.059	0.039	0.029	1.192
AQI	-0.0411	0.006	-6.764	0.000	-0.053	-0.029
AQI_No2	0.3222	0.014	23.599	0.000	0.295	0.349
AQI_Ozone	0.0010	0.007	0.149	0.882	-0.012	0.015

Omnibus:	42353.575	Durbin-Watson:	1.913
Prob(Omnibus):	0.000	Jarque-Bera (JB):	589191296.897
Skew:	24.432	Prob(JB):	0.00
Kurtosis:	894.795	Cond. No.	165.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

This R-squared is extremely low, showing little-to-no predictability.

Predicting COVID deaths with AQ metrics

Again, very low R-squared (low correlation)

Sources

1. Cazzolla Gatti, R., Velichevskaya, A., Tateo, A., Amoroso, N., & Monaco, A. (2020). Machine learning reveals that prolonged exposure to air pollution is associated with SARS-CoV-2 mortality and infectivity in Italy. Environmental pollution (Barking, Essex : 1987), 267, 115471. https://doi.org/10.1016/j.envpol.2020.115471