Anna Krutsinger
Keizo Morgan
Ingrid Tsang

# Preliminary Findings

**Research Question:** How does geography affect the human impact of COVID-19?

Our project consists of two parts: first, use LDA on the set of research papers to discover geographical factors of interest (for example, the impact of humidity of the transmission rate), and second, run linear regressions on said factors with outputs such as mortality rates.

## Topic Discovery

We used the cord19 library which provides easy processing capabilities for the research paper dataset from the COVID-19 Open Research Dataset Challenge. On the subset papers published since SARS-CoV-2 (December 2019), we ran LDA on the paper summaries (distillation of the abstracts provided by the library), and then again with the paper titles. However, we were not able to get the topics from either LDA to bring about areas of interest with regards to factors that may correlate with case numbers. We may have to scrap this half of the project unless we can find a way to make this work (even trying to filter the papers by titles that include a word like "transmission" or "humidity" still is not helpful).

```
lda_summaries.print_topics()
```

```
['respiratory', 'infection', 'infections', 'positive', 'patients', 'symptoms', 'negative', 'test']
['patients', 'clinical', 'cases', 'pneumonia', 'imaging', 'lung', 'ct', 'chest']
['days', 'number', 'confirmed', 'countries', 'cases', 'case', 'data', 'rate']
['study', 'infection', 'patients', 'clinical', 'treatment', 'hospital', 'medical', 'icu']
['infected', 'number', 'epidemic', 'countries', 'cases', 'time', 'data', 'model']
['patients', 'health', 'emergency', 'medical', 'care', 'management', 'healthcare', 'pandemic']
['human', 'host', 'receptor', 'proteins', 'genome', 'binding', 'protein', 'spike']
['health', 'novel', 'cases', 'outbreak', 'china', 'ncov', 'wuhan', 'hubei']
['social', 'distancing', 'measures', 'transmission', 'epidemic', 'control', 'model', 'spread']
['cell', 'cells', 'respiratory', 'infection', 'immune', 'treatment', 'severe', 'antiviral']
['based', 'images', 'using', 'accuracy', 'proposed', 'learning', 'dataset', 'deep']
['evidence', 'patients', 'review', 'children', 'risk', 'diabetes', 'heart', 'adults']
['air', 'use', 'masks', 'protective', 'personal', 'equipment', 'face', 'water']
['expression', 'based', 'drug', 'potential', 'drugs', 'vaccine', 'angiotensin', 'ace2']
['study', 'public', 'health', 'research', 'information', 'influenza', 'data', 'pandemic']
['rt', 'samples', 'rna', 'pcr', 'detection', 'assay', 'sensitivity', 'diagnostic']
['research', 'response', 'et', 'la', 'en', 'des', 'les', 'le']
['patients', 'associated', 'severe', 'age', 'group', 'higher', '95', 'ci']
['individuals', 'based', 'population', 'testing', 'time', 'data', 'tests', 'asymptomatic']
['respiratory', 'syndrome', 'mers', 'human', 'coronaviruses', 'new', 'la', 'en']
```

```
lda_titles.print_topics()
```

```
['review', 'rapid', 'meta', 'analysis', '10', 'systematic', 'february', 'evidence']
['clinical', 'characteristics', 'study', 'patients', 'pneumonia', 'china', 'novel', 'wuhan']
['el', 'et', 'en', 'des', 'la', 'du', 'le', 'por']
['self', 'influenza', 'based', 'high', 'infectious', 'genetic', 'like', 'chapter']
['epidemic', 'lessons', 'pandemic', 'strategies', 'medical', 'period', 'protective', 'surgery']
['based', 'molecular', 'detection', 'using', 'learning', 'pcr', 'deep', 'testing']
['epidemic', 'health', 'public', 'outbreak', 'healthcare', 'care', 'workers', 'mental']
['surveillance', 'infection', 'development', 'detection', 'mediated', 'specific', 'contact', 'tracing']
['health', 'emerging', 'new', 'outbreak', 'pandemic', 'research', 'global', 'crisis']
['novel', 'treatment', 'therapeutic', 'potential', 'chinese', 'medicine', 'drug', 'diagnosis']
['novel', 'human', 'non', 'population', 'transmission', 'modeling', 'evolution', 'dynamics']
['children', 'patients', 'infection', 'pediatric', 'long', 'angiotensin', 'therapy', 'era']
['cells', 'respiratory', 'syndrome', 'infection', 'acute', 'protein', 'spike', 'severe']
['effect', 'united', 'states', 'social', 'lockdown', 'india', 'spread', 'distancing']
['respiratory', 'control', 'infection', 'novel', 'outbreak', 'impact', 'prevention', 'measures']
['acid', 'effective', 'cases', 'novel', 'rapid', 'pandemic', 'characterization', 'confirmed']
['risk', 'factors', 'mortality', 'patients', 'infection', 'associated', 'case', 'assessment']
['role', 'patients', 'response', 'cancer', 'care', 'pandemic', 'management', 'experience']
['epidemic', 'infection', 'case', 'time', 'analysis', 'rate', 'ace2', 'data']
['epidemic', 'based', 'rna', 'novel', 'prediction', 'vaccine', 'model', 'ncov']
```

We also created a search-like feature for the repository of research articles where we can extract key sentences from the abstracts related to sets of words. For example:

```
search_words = ['seasonal','transmission']
df1 = search(df, search_words)
df1 = sentences(df1, search_words)
df1.head()
```

| | pub_date | title | author | sentence |
|---|---|---|---|---|
| 0 | 2020-04-02 | Infectious diseases in children and adolescent... | Dong, Yanhui; Wang, Liping; Burgner, David P; ... | Many challenges remain around reducing regiona... |
| 1 | 2020-04-14 | Projecting the transmission dynamics of SARS-C... | Kissler, Stephen M.; Tedijanto, Christine; Gol... | We used estimates of seasonality, immunity, an... |
| 2 | 2020-02-14 | A spatial model of CoVID-19 transmission in En... | Leon Danon; Ellen Brooks-Pollock; Mick Bailey;... | Seasonal changes in transmission rate substant... |
| 3 | 2020-02-17 | Potential impact of seasonal forcing on a SARS... | Richard A Neher; Robert Dyrdak; Valentin Druel... | Here, we explore how seasonal variation in tra... |
| 4 | 2020-03-24 | Social distancing strategies for curbing the C... | Stephen M Kissler; Christine Tedijanto; Marc L... | The amount of social distancing needed to curb... |

```
search_word2 = ['temperature', 'humidity']
df2 = search(df, search_word2)
df2
```

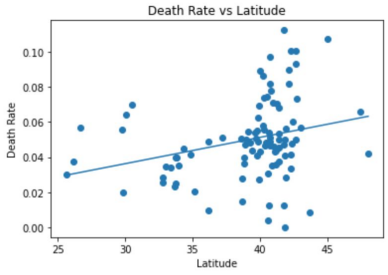
| | title | abstract | publish_time | authors | journal |
|---|---|---|---|---|---|
| 32033 | Correlation between weather and Covid-19 pande... | abstract this study aims to analyze the correl... | 2020-04-04 | Tosepu, Ramadhan; Gunawan, Joko; Effendy, Devi... | Science of The Total Environment |
| 32039 | Investigation of effective climatology paramet... | abstract sars cov-2 (covid-19) coronavirus cas... | 2020-04-17 | Ahmadi, Mohsen; Sharifi, Abbas; Dorosti, Shadi... | Science of The Total Environment |
| 32043 | COVID-19 transmission in Mainland China is ass... | abstract covid-19 has become a pandemic. the i... | 2020-04-19 | Qi, Hongchao; Xiao, Shuang; Shi, Runye; Ward, ... | Science of The Total Environment |
| 32044 | Impact of weather on COVID-19 pandemic in Turkey | abstract the coronavirus pandemic, which has n... | 2020-04-20 | Şahin, Mehmet | Science of The Total Environment |
| 32061 | Effects of temperature and humidity on the dai... | abstract the coronavirus disease 2019 (covid-1... | 2020-04-28 | Wu, Yu; Jing, Wenzhan; Liu, Jue; Ma, Qiuyue; Y... | Science of The Total Environment |
| 35899 | The role of absolute humidity on transmission ... | a novel coronavirus (covid-19) was identified ... | 2020-02-17 | Wei Luo; Maimuna S Majumder; Dianbo Liu; Canel... | unable to retrieve data |
| 35908 | Analysis of meteorological conditions and pred... | objective: to investigate the meteorological c... | 2020-02-18 | Jin Bu; Dong-Dong Peng; Hui Xiao; Qian Yue; Ya... | unable to retrieve data |
| 36220 | Role of temperature and humidity in the modula... | covid-19 is having a great impact on public he... | 2020-03-08 | Barbara Oliveiros; Liliana Caramelo; Nuno C Fe... | unable to retrieve data |
| 36391 | Effects of temperature variation and humidity ... | object meteorological parameters are the impor... | 2020-03-18 | Yueling Ma; Yadong Zhao; Jiangtao Liu; Xiaotao... | unable to retrieve data |
| 36416 | Roles of meteorological conditions in COVID-19... | the novel coronavirus (sars-cov-2/ 2019-ncov) ... | 2020-03-20 | Biqing Chen; Hao Liang; Xiaomin Yuan; Yingying... | unable to retrieve data |
| 36503 | Role of meteorological temperature and relativ... | identified in december 2019, the 2019-ncov eme... | 2020-03-23 | Jose Alvarez-Ramirez; MONICA MERAZ | unable to retrieve data |
| 36567 | The impact of temperature and absolute humidit... | objective to investigate the impact of tempera... | 2020-03-24 | Peng Shi; Yinqiao Dong; Huanchang Yan; Xiaoyan... | unable to retrieve data |
| 36605 | Climate affects global patterns of COVID-19 ea... | environmental factors, including seasonal clim... | 2020-03-27 | Gentile Francesco Ficetola; Diego Rubolini | unable to retrieve data |
| 36741 | Causal empirical estimates suggest COVID-19 tr... | nearly every country is now combating the 2019... | 2020-03-30 | Tamma Carleton; Kyle C. Meng | unable to retrieve data |
| 36795 | Temperature, humidity, and wind speed are asso... | in absence of empirical research data, there h... | 2020-03-30 | Nazrul Islam; Sharmin Shabnam; A Mesut Erzurum... | unable to retrieve data |

We would like to figure out a way to integrate LDA into this project.

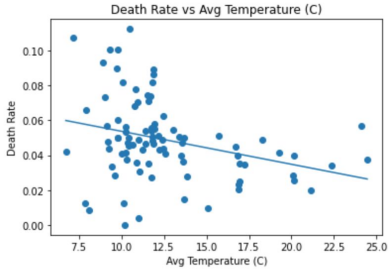
## Regression

We performed linear regressions of several geographical factors -- latitude, average temperature, and average heat index -- on mortality rates on the county-level in the US.

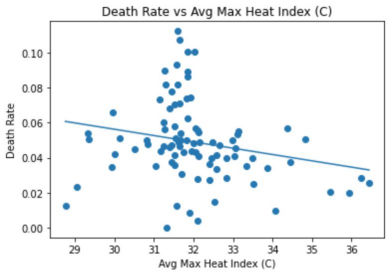
Filtering out counties with less than 5000 cases, we get the following results:




```
p-values: 0.0065560206153594 05
R^2: 0.07678758923740729
Slope: 0.001496432008797724
```

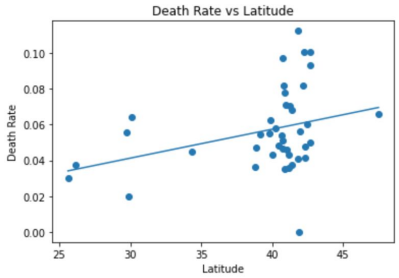



```
p-values: 0.0036884623645239895
R^2: 0.09389815023947228
Slope: -0.0018930812417575972
```

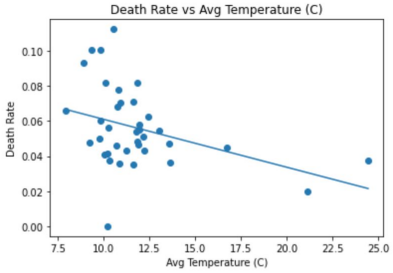



```
p-values: 0.038259971252509045
R^2: 0.048973578766576945
Slope: -0.0036135820985701067
```

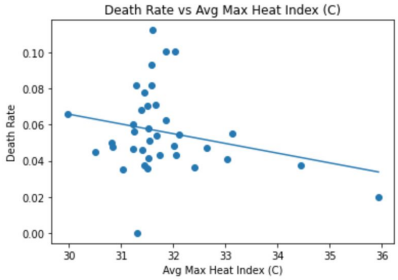
Filtering out counties with less than 5000 cases, we get the following results:




```
p-values: 0.036887760358230 32
R^2: 0.10694645418978166
Slope: 0.0016183989222999038
```




```
p-values: 0.022709039859089 32
R^2: 0.1396446930508946
Slope: -0.002731496192405999
```




```
p-values: 0.14355563056831558
R^2: 0.060118230230872
Slope: -0.005396814729671787
```

While not entirely conclusive, the positive R^2 values for the graphs of death rate against latitude and average temperature imply that there is slight positive correlation between the two variables respectively. In particular, the R^2 value for death rate vs average temperature for counties with more than 5000 cases was 14%, which was a surprisingly high correlation. Additionally, all the p-values for all but the graph of death rate against average max heat index for counties with more than 5000 cases are smaller than 0.05, which indicates statistically significant results.

We plan to extend these regressions with more geographical factors, including population density, average humidity and average pollution levels, on incident and mortality rates on the county-level. We may also go further by combining two or more factors to perform regressions on incident and mortality rates.