# Analysis on Time Series Meteorological Data

## Data Source and Description

The meteorological data consists of average temperature, minimum temperature, maximum temperature, total precipitation, relative humidity of each day from 1/1/2020 to 6/30/2020 at the counties. The covid data consists of confirmed cases and deaths in each day from 1/22/2020 to 7/28/2020 at the counties.

Table : Description of Meteorological Features

|  |  |
| --- | --- |
| **Feature** | **Description** |
| Mean T | Average Temperature of the day in degree Celsius |
| Min T | Minimum temperature of the day in degree Celsius |
| Max T | Maximum temperature of the day in degree Celsius |
| Total Precipitation | Total Precipitation of the day in percentage(?) |
| Relative Humidity | Relative humidity of the day in percentage |

## Preprocessing

The covid data for confirmed cases and deaths were processed separately. The covid data was aligned with available meteorological data. Then for each county initial and trailer zeros of covid data were removed.

## Spearman

For both time series of confirmed cases and deaths, spearman w.r.t. time series of meteorological data was calculated in each county. For each county, spearman rank correlation was calculated for each target variable against each meteorological feature.

(The result was used to plot in map.)

## Plot

The spearman correlation calculated at each county were plotted in a map for each case considered.

# Analysis on Socio-demographic & Comorbidities Data

## Data Source and Description

The socio-demographic and comorbidities data consist of a list of features. Each feature is either average value or number of people in each county. There were some missing data here. Along with this data, we needed population, covid cases and deaths count at each county.

Table : Description of Socio-demographic and Comorbidities features

|  |  |  |
| --- | --- | --- |
| **Feature** | **Description** | **Unit** |
| POP13\_SQMI | Population density (in 2013) of each county | population per square mile |
| WHITE | Number of white people in each county | Positive integer |
| BLACK | Number of black people in each county | Positive integer |
| AMERI\_ES | Number of Native American people in each county | Positive integer |
| ASIAN | Number of people of Asian race in each county | Positive integer |
| HAWN\_PI | Number of pacific islanders in each county | Positive integer |
| HISPANIC | Number of Hispanic people in each county | Positive integer |
| OTHER | Aggregated number of people of other races in each county | Positive integer |
| MULT\_RACE | Number of people belonging to multiple races in each county | Positive integer |
| MALES | Number of males in each county | Positive integer |
| FEMALES | Number of females in each county | Positive integer |
| AGE\_UNDER5 | Number of people below age 5 | Positive integer |
| AGE\_5\_9 | Number of people with age between 5 and 9 | Positive integer |
| AGE\_10\_14 | Number of people with age between 10 and 14 | Positive integer |
| AGE\_15\_19 | Number of people with age between 15 and 19 | Positive integer |
| AGE\_20\_24 | Number of people with age between 20 and 24 | Positive integer |
| AGE\_25\_34 | Number of people with age between 25 and 34 | Positive integer |
| AGE\_35\_44 | Number of people with age between 35 and 44 | Positive integer |
| AGE\_45\_54 | Number of people with age between 45 and 54 | Positive integer |
| AGE\_55\_64 | Number of people with age between 55 and 64 | Positive integer |
| AGE\_65\_74 | Number of people with age between 65 and 74 | Positive integer |
| AGE\_75\_84 | Number of people with age between 55 and 84 | Positive integer |
| AGE\_85\_UP | Number of people with age 85 and above | Positive integer |
| HOUSEHOLDS | Number of households | Positive integer |
| AVE\_HH\_SZ | Average household size | Number of people per household |
| HSEHLD\_1\_M | Number of 1-person male households | Positive integer |
| HSEHLD\_1\_F | Number of 1-person female households | Positive integer |
| MARHH\_CHD | Number of married couple households with children | Positive integer |
| MARHH\_NO\_C | Number of married couple households with no child | Positive integer |
| MHH\_CHILD | Number of male-headed households with children | Positive integer |
| FHH\_CHILD | Number of female-headed households with children | Positive integer |
| FAMILIES | Number of families | Positive integer |
| AVE\_FAM\_SZ | Average family size | Number of people per family |
| HSE\_UNITS | Number of total housing units | Positive integer |
| VACANT | Number of vacant housing units | Positive integer |
| OWNER\_OCC | Number of owner-occupied housing units | Positive integer |
| RENTER\_OCC | Number of renter-occupied housing units | Positive integer |
| Prevalence of obesity | Number of people with obesity problem | Positive integer |
| Hypertension | Number of people suffering from hypertension | Positive integer |
| Diabetes | Number of people affected with diabetes | Positive integer |
| CVD | Number of people suffering from Cardio Vascular Disease | Positive integer |
| HIV/AIDS | Number of people with HIV/AIDS positive | Positive integer |

## Preprocessing

The count data were normalized through division by population. The missing data were filled with median of the normalized value.

### Spearman

For both confirmed cases and deaths of the counties, spearman w.r.t. normalized features of the counties were calculated. Spearman rank correlation was calculated for each target variable against each socio-demographic and comorbidities feature.

Table 3 reports the Spearman Rank Order Correlation coefficients for COVID-19 incidence and mortality rate. Only the features with |coefficient| > 0.3 have been selected for further processing and have been reported in the table.

Table 3: Spearman Rank Order Correlation coefficients

|  |  |  |  |
| --- | --- | --- | --- |
| **Incidence** | | **Mortality** | |
| **Feature** | **Coefficient** | **Feature** | **Coefficient** |
| Black | 0.553737 | Black | 0.505527 |
| White | -0.54029 | Population Density | 0.489668 |
| Married Household (No Child) | -0.52741 | White | -0.42896 |
| Average Family Size | 0.506271 | Married Household (No Child) | -0.42414 |
| Average Household Size | 0.425235 | Average Family Size | 0.385149 |
| Population Density | 0.406421 | Female-Headed Home with Child | 0.380744 |
| Owner-Occupied Housing Units | -0.40616 | Average Household Size | 0.332255 |
| Age Group (55-64) | -0.40555 | Vacant Housing Units | -0.32871 |
| Total Housing Units | -0.40504 | HIV/AIDS | 0.324306 |
| Age Group (65-74) | -0.39865 | Total Housing Units | -0.32343 |
| HIV/AIDS | 0.397918 | Asian | 0.322075 |
| Female-Headed Home with Child | 0.38039 | Age Group (65-74) | -0.31923 |
| Household with 1 Male | -0.37942 | Age Group (20-24) | 0.317794 |
| Other | 0.375594 | Household with 1 Male | -0.31739 |
| Households | -0.36472 | Owner-Occupied Housing Units | -0.31154 |
| Vacant Housing Units | -0.36103 | Age Group (75-84) | -0.3066 |
| Age Group (75-84) | -0.36 |  |  |
| Age Group (20-24) | 0.336749 |  |  |
| Age Group (25-34) | 0.333361 |  |  |
| Hispanic | 0.323769 |  |  |
| Families | -0.30718 |  |  |

### Covariance

Covariance of the normalized features were calculated.

Table 4: Correlated feature sets identified through Covariance Analysis

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Set** | **Correlated Features** | | | |
| 1 | Average Family Size | Average Household Size |  |  |
| 2 | Total Housing Units | Vacant Housing Units |  |  |
| 3 | Hispanic | Other |  |  |
| 4 | Female-Headed Home with Child | Black |  |  |
| 5 | Married Household (No Child) | Owner-Occupied Housing Units | Households | Families |
| 6 | Household with 1 Male | Households |  |  |
| 7 | Age Group (55-64) | Age Group (65-74) | Age Group (75-84) |  |

### Feature Selection

For further analysis, features were selected based on good spearman correlation (p-value < 5% and |covariance| > 0.3). Then those features were coalesced based on covariance analysis. Some were coalesced as ratio of two features, while others were coalesced as selection of one feature from a set of correlated features.

Table 5.1: Selected Features For Incidence

|  |  |  |
| --- | --- | --- |
|  | **Feature** | **Reason** |
| 1 | *Black/White Ratio* | Incidence is significantly (spearman) correlated with *Black* and *White*. It is positive with *Black* and positive with *White*. So, we take ratio of the two. Again, *Female-Headed Households with Children* has positive (spearman) correlation with incidence and positive correlation with *Black*. So, this feature is also covered by the selected feature. |
| 2 | *Average Family Size* | Incidence has positive (spearman) correlation with both *Average Family Size* and *Average Household Size*. And both these features are correlated to each other (from covariance analysis). We select the one with higher spearman coefficient. |
| 3 | *Population Density* | *Population Density* has significantly positive (spearman) correlation with incidence and is not correlated with any other feature. |
| 4 | *Married Household (No Child)* | Incidence has negative (spearman) correlation *with Married Household (No Child)*, *Owner-Occupied Housing Units*, *Households* and, *Families*. All of these features are correlated to each other (from covariance analysis). We select the one with the highest (absolute) spearman coefficient. |
| 5 | *Total Housing Units* | Incidence has negative (spearman) correlation with *Total Housing Units* and *Vacant Housing Units*. And both these features are correlated to each other (from covariance analysis). We select the one with higher spearman coefficient. |
| 6 | *Other races* | Incidence has positive (spearman) correlation with *Other races* and *Hispanic race*. And both these features are correlated to each other (from covariance analysis). We select the one with higher spearman coefficient. |
| 7 | *HIV/AIDS* | *HIV/AIDS* has significantly positive (spearman) correlation with incidence and is not correlated with any other feature. |
| 8 | *Age Group (55-84)* | Incidence has negative (spearman) correlation with *Age Groups 55-64, 65-74 and 75-84*. All of these are correlated (from covariance analysis). We add them up into a single age group for obvious reasons. |
| 9 | *Age Group (20-34)* | Incidence has positive (spearman) correlation with *Age Groups 20-24 and 25-34*. We add them up into a single age group for obvious reasons. |
| 10 | *Household with 1 Male* | Incidence has negative (spearman) correlation with both *Household with 1 Male* and *No. of Households*. And both these features are correlated to each other (from covariance analysis). We select the one with higher spearman coefficient. |

Table 5.2: Selected Features For Mortality

|  |  |  |
| --- | --- | --- |
|  | **Feature** | **Reason** |
| 1 | *Black/White Ratio* | Mortality is significantly (spearman) correlated with *Black* and *White*. It is positive with *Black* and positive with *White*. So, we take ratio of the two. Again, *Female-Headed Households with Children* has positive (spearman) correlation with incidence and positive correlation with *Black*. So, this feature is also covered by the selected feature. |
| 2 | *Average Family Size* | Mortality has positive (spearman) correlation with both *Average Family Size* and *Average Household Size*. And both these features are correlated to each other (from covariance analysis). We select the one with higher spearman coefficient. |
| 3 | *Population Density* | *Population Density* has significantly positive (spearman) correlation with mortality and is not correlated with any other feature. |
| 4 | *Married Household (No Child)* | Mortality has negative (spearman) correlation *with Married Household (No Child)* and *Owner-Occupied Housing Units*. These features are correlated to each other (from covariance analysis). We select the one with the highest (absolute) spearman coefficient. |
| 5 | *Total Housing Units* | Mortality has negative (spearman) correlation with *Total Housing Units* and *Vacant Housing Units*. And both these features are correlated to each other (from covariance analysis). We select *Total Housing Units*, because this feature is selected for incidence. |
| 6 | *Asian race* | *Asian race* has positive (spearman) correlation with mortality and is not correlated with any other feature. |
| 7 | *HIV/AIDS* | *HIV/AIDS* has significantly positive (spearman) correlation with mortality and is not correlated with any other feature. |
| 8 | *Age Group (65-84)* | Mortality has negative (spearman) correlation with *Age Groups 65-74 and 75-84*. All of these are correlated (from covariance analysis). We add them up into a single age group for obvious reasons. |
| 9 | *Age Group (20-24)* | *Age Groups 20-24* has positive (spearman) correlation with mortality and is not correlated with any other feature. |
| 10 | *Household with 1 Male* | *Household with 1 Male* has positive (spearman) correlation with mortality and is not correlated with any other feature that mortality has correlation with. |

## OLS

For the selected features, an Ordinary Least Square(OLS) Regression model was formed for both incidence and mortality. This model performs poorly, in terms of explaining the variance of the target variable, for both incidence and mortality.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metric** | **Incidence** | | **Mortality** | |
| **OLS** | **GWR** | **OLS** | **GWR** |
| R-Squared | 0.21 | 0.54 | 0.143 | 0.66 |
| AIC | 5038.76 | 3639.86 | -13868.55 | -16304.77 |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Features** | (Intercept) | Black/  White Ratio | Average Family Size | AGE Group (55-84) | Married Household (No Child) | Total Housing Units | Other | Population Per Sq. Mile | HIV/AIDS | Household with 1 Male | Age Group (20-24) |
| **Coefficients** | -0.57733 | 0.328399 | 0.340304 | 1.796492 | -4.03359 | -0.07371 | 1.96249 | 4.10E-05 | -0.05264 | -4.52484 | 0.687415 |

## GWR

For the selected features, a GWR model was formed.

# Analysis on Unauthorized Estimates Data

## Data Source and Description

The data is given as number of unauthorized people in some counties. Some data were aggregation of several counties.

## Preprocessing

The count data was normalized through division by population. The aggregated data were divided into corresponding counties as per its population.

## Spearman

For both confirmed cases and deaths of the counties, spearman w.r.t. normalized unauthorized estimates of the counties was calculated. Spearman rank correlation was calculated for each target variable against each socio-demographic and comorbidities feature.

The result is given in the following table.

|  |  |  |
| --- | --- | --- |
| **Feature** | **Spearman coefficient** | **P-value** |
| Covid incidence | 0.200614 | 0.002137 |
| Covid mortality | 0.055512 | 0.4 |

# Analysis on Vaccination Data

## Data Source and Description

The data is given as total number of people vaccinated in each state from 1/12/2021 to 3/27/2021.

## Preprocessing

The socio-demographic and comorbidities data was aggregated for each state and concatenated with latest total number of vaccinated of that state. The features with average value were aggregated as simple average and count values were aggregated as summation. Then all count data were normalized through division by population. The missing data were filled with median of the normalized value.

### Spearman

Spearman correlation was calculated vaccination data of each state vs each of the features of socio-demographic and comorbidities data.

|  |  |  |
| --- | --- | --- |
| **Feature** | **Coefficient** | **P-value** |
| Age Group (≥85) | 0.606878 | 2.35E-06 |
| Age Group (45-54) | 0.51448 | 0.000112 |
| Age Group (75-84) | 0.466878 | 0.000553 |
| Males | 0.449412 | 0.000939 |
| Cardio-Vascular-Disease | -0.43538 | 0.001407 |
| Age Group (55-64) | 0.433394 | 0.001489 |
| Black | -0.43077 | 0.001602 |
| White | 0.342896 | 0.013769 |
| Household with 1 Male | 0.337919 | 0.0153 |
| Household with 1 Female | 0.322805 | 0.020868 |
| HIV/AIDS | -0.31647 | 0.023665 |
| Female-Headed Home with Child | -0.30262 | 0.03089 |
| Females | 0.302262 | 0.031101 |

### Feature Selection

For further analysis, features were selected based on good spearman correlation (p-value < 5% and |covariance| > 0.3). Then those features were coalesced based on their meaning. Some were coalesced as ratio of two features, while others were coalesced as summation of a set of related features.

|  |  |
| --- | --- |
| **Feature** | **Reason** |
| *Black/White Ratio* | Vaccination has correlation with *Black* and *White*. We take their ratio for obvious reason. |
| *Male/Female Ratio* | Vaccination has correlation with *Male* and *Female*. We take their ratio for obvious reason. |
| *Age Group (≥45)* | Vaccination has correlation with *Age Groups 45-54, 55-64, 75-84* and, *85 and up*. We take summation of *Age Groups 45 and up* for obvious reason. |
| *Household with 1 Person* | Vaccination has correlation with *Household with 1 Male* and *Household with 1 Female*. We take their summation for obvious reason. |
| *Cardio-Vascular-Disease* | *Cardio-Vascular-Disease* is correlated to vaccination. |
| *HIV/AIDS* | *HIV/AIDS* is correlated to vaccination. |

## GWR

For the selected features, a GWR model was formed.

* Time Series (cases & deaths)
  + Spearman of cases or deaths vs climate (of time data) per county
  + Plot of spearman across the US
* Tabular (cases & deaths)
  + Data preprocessing (normalization, merge of features etc.)
  + Spearman of cases or deaths vs other features (of each county)
  + Covariance of all features
  + Select features with good spearman and coalesce them based on covariance
  + OLS on the selected features
  + GWR on selected features
* Unauthorized
  + Spearman of cases or deaths vs fraction of unauthorized
  + Not good result
* Vaccination
  + Spearman of vaccinations vs other features (of each state)
  + Select features with good spearman and coalesce them based on meaning (i.e., common sense)
  + GWR on the selected features