

# Equity in Healthcare

Gerard Abiassaf, Brian Bahr, Jasey Diaz, Holly Hammons

## A. IDENTIFYING THE VULNERABILITY FACTORS

There are many socioeconomic and environmental factors that can lead to an increase in vulnerability against disease outbreaks and pandemics. According to the Washington Post [1], those who commute to work by means of public transportation are at a higher risk of spreading infectious diseases, and those who commute by bus or train remain within close proximity to others during prolonged periods of time. The article mentions, while both wealthy and poor people may commute by means of public transportation, wealthier people have more opportunities to take paid sick leave or work remotely [1]. However, those who work at lower income jobs may be more likely to commute by public transportation when they are sick [1]. Indeed, a recent study [2] investigated the relationship between density and pandemic spread and found that there are higher COVID-19 morbidity rates in areas that contain lower levels of car ownership.

Another factor to consider is the employment rate prior to and during a pandemic. For instance, the authors in [2] found a relationship between the employment rate of a population sample and the infection rate of COVID-19. They noted, as the employment rate increased within districts, the infection rate for COVID-19 decreased [2]. However, because many people lost their jobs during the COVID-19 pandemic, special care should be taken when considering employment rate as a factor [2]. Even so, a high income prior to a pandemic may have allowed people to save for an emergency fund and, therefore, comply more easily to stay-at-home orders and social distancing rules [2]. In addition, those with employment history may be eligible to receive health insurance or unemployment benefits, which could further assist them in the event of a pandemic.

The highest amount of education obtained is an important factor to consider. A recent study [3] explored determinants of COVID-19 infection and found a correlation between obtaining higher education and infection rate. It was found that counties with a higher percentage of adults with any education beyond high school completed generally had significantly lower infection rates of COVID-19 [3]. The authors [3] suggest that those who have a better understanding of the virus

are likely to take social distancing and shelter-in-place orders more seriously. Moreover, a local Denver news article [4], which explored COVID-19 infection rates between Denver neighborhoods, noted that the Montbello neighborhood, which had a large number of COVID-19 cases, has some of the fewest high school graduates in the surrounding area [4].

Another factor that should be considered is the crime rate in particular areas. For example, it is important to consider the consequences that incarceration has on being vulnerable to infectious diseases. A recent article [5] was aiming to identify a theoretical framework for outlining the harms of incarceration associated with pandemics. For instance, those who are incarcerated are more likely to be less healthy at a baseline than the general population before being detained, and serving time behind bars will only worsen their health [5]. Upon release, formally incarcerated people may suffer from diminished immune system functionality due to their poor health behaviors in prison [5]. These health problems, combined with residing in disadvantaged and under-resourced communities, could make formally incarcerated people more vulnerable to infectious diseases [5].

Minority status should be considered a factor when determining vulnerability. A recent study [6] aimed to examine temporal trends among counties with differing social vulnerability levels to quantify disparities in trends over time. Among many different factors that they examined, they found that the strongest disparity between the most and least vulnerable counties occurred for their category of “Speaks English Less than Well”. In fact, the CDC uses minority status to measure social vulnerability in their index, in which they quantify levels of social vulnerability for certain geographic regions [7].

Age can be considered a vulnerability factor, as the young and elderly are disproportionately affected by disease outbreaks. For example, COVID-19 proved to be devastating to the elderly population because of their weakened immune system, especially if they’re experiencing additional underlying health issues, whereas young people have yet to build antibodies that are required to fight viruses [8].

Access to clean water and proper sanitation equipment should be considered when assessing population

vulnerability. In a recent study [10] regarding COVID-19 sanitation impacts, a lack of access to clean water, sanitation, and proper hygiene services was found to contribute to COVID-19 public health risks. The article [10] emphasizes that many low-income communities, without access to adequate plumbing, will fail to maintain hand hygiene as a means to prevent disease spread. Additionally, a person who has been living without reliable access to clean water and adequate plumbing will be more likely to suffer from pre-existing health conditions, prior to a global pandemic [10].

Many factors were narrowed down to five, to represent data that heavily contributes to population vulnerability. The factors were chosen in part due to their numerous crossovers with other factors. For example, those who are, or were, recently employed, usually correspond to central age groups, and those who reside in high crime areas are more likely to live in poverty. Additionally, these factors tend to limit opportunities for people to become less vulnerable during an unforeseen pandemic. For instance, those who prefer to comply with social distancing measures are more limited if they must commute by a crowded bus. Moreover, the lack of healthcare opportunities, disproportionately affecting minorities, is inherently limiting one's progression of well being during a crisis. Finally, we felt it was important to recognize the increased amount of social and economic opportunities that result from obtaining higher education. The chosen top five factors can be found in the list below.

- 1) Education Level Obtained
- 2) Living in Poverty/Sanitation Conditions
- 3) Employment Status and Benefits
- 4) Transportation Means
- 5) Race and Language

#### B. CONSULTING US CENSUS DATA

Colorado Data was collected from the American Community Survey from the US Census Bureau website [11] from the years 2015 to 2020. The specific geographic regions analyzed were Adams, Arapahoe, Denver, and Jefferson counties, which corresponded to a total of 463 tracts. The list below describes the attributes chosen from these datasets that expand on the five factors identified in section A.

- Highest Education Earned
- Has Health Insurance
- Race
- Transportation Method to Work
- Language Spoken at Home
- Employment Status
- Has Plumbing Facilities
- Income Relative to Poverty Level

#### C. DATA PRE-PROCESSING

Before data can be properly analyzed, it must first be cleaned and pre-processed. The datasets obtained from the American Community Survey consisted of multiple spreadsheets with different population samples for each tract. In order to use different sized datasets in a single model, the population percentage of each tract was calculated with regards to the sample size of the attribute column. By doing so, each tract could be compared against different attributes as a percentage of each sample group.

Four census tracts contained missing values for crucial selected attributes. Because missing values will cause errors when performing dimensionality reduction, and due to the sufficiently large ( $n > 30$ ) sample, the identified four tracts were removed.

As a result of the standard formatting of the American Community Survey, and it being an official government source, the dataset was free of duplicate values and all values were considered to be accurate. The data set was normalized before performing Principal Component Analysis (PCA) for dimensionality reduction.

#### D. DIMENSIONALITY REDUCTION

The final dataset was very large with a size of 459 x 25, causing a reduction in interpretability of our model. So, the variances of each attribute, with regards to each census tract, were observed. However, due to each attribute column containing different subjects, each column had vastly different means. As a result, it was decided to compare variability by comparing each coefficient of variation (CV) [12].

CV is defined as the percentage quantity of the ratio of the standard deviation to the mean. This statistic provides a method for measuring the variability in relation to the mean of each attribute [12]. The maximum calculated CV was used as the metric to decide which attributes to discard. It was chosen that any attributes, which contained less than 2% of the variability of the maximum CV, would be discarded. These attributes held little information compared to the maximum CV obtained. The following attributes listed below were discarded due to their individual CV being less than 2% of the maximum CV.

- No Health Insurance
- In Labor Force
- Income At Or Above Poverty Level
- Commute Own Vehicle

The American Community Survey contained columns that had the inverse of itself, such as does or does not have health insurance, is or is not in the labor force,

and income at or below poverty level. Therefore, given the low variability in these columns listed above, and due to them containing complement columns, they were deemed acceptable to remove. In addition, commuting by vehicle, compared with any other mode of transportation, had very low variability. Because the information with regards to commuting by public transit is more concerning to the vulnerability model, it was acceptable to remove the column that described commuting by one's own vehicle. These removals reduced the size of the dataset to 459 x 21.

After discarding attributes with low variability, the dataset was normalized and PCA was performed (see Appendix A). Figure 1 below displays the principal components from the dataset, sorted by the percentage of variability in data explained.

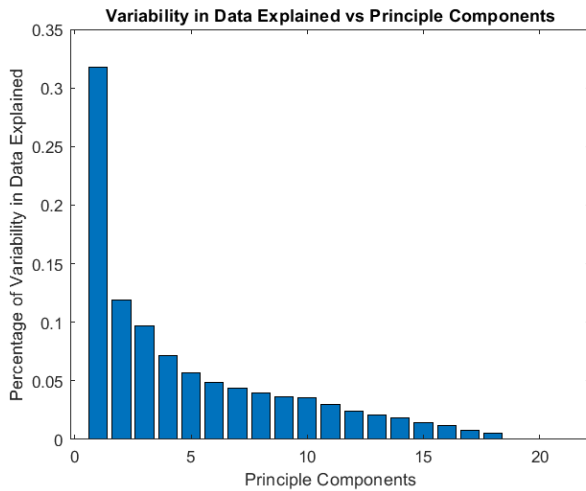


Fig. 1. Principal Components Sorted by Percentage of Variability in Data Explained by Each One

12 principal components were identified that explained more than 90% of the variability. This translated to reducing the size of the dataset from 459 x 21 to 459 x 12.

#### E. COMPUTING VULNERABILITY LEVEL

Using the approach described in [13], the reduced dataset was used to compute vulnerability levels by finding the sum of all the values of the principal components. The following list below consists of the 5 most vulnerable tracts, based on their computed vulnerability level.

- Tract number: 8031000600
- Tract number: 8001007801
- Tract number: 8031001800
- Tract number: 8005007301
- Tract number: 8001007802

After examining the distribution (see Appendix B), a vulnerability ranking was assigned for each tract based on the value  $v$  of its vulnerability as compared to the mean and standard deviation of the distribution. Finally, a map of the census tracts of interest in the Denver area was developed. The map shown in Figure 5 is color-coded by the tract's vulnerability ranking, denoted as 1-5, which corresponds to each tract being least to most vulnerable.

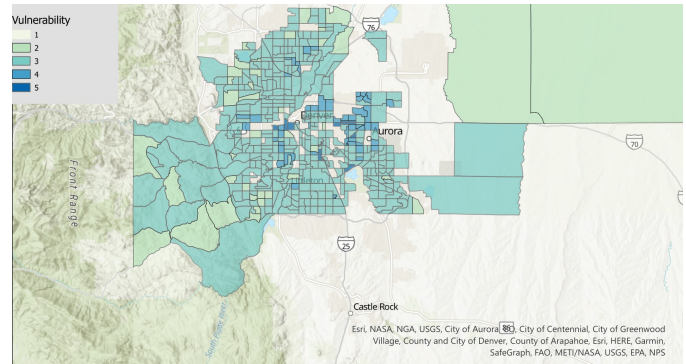


Fig. 2. Comparative Vulnerability of Census Tracts in the City of Denver

The 5 most vulnerable tracts reside around the cities of Denver, Aurora, and Commerce City. These five census tracts are in the northeastern region of the Denver Metro Area, which is known to have the highest population of minorities in the area [14]. Additionally, these areas are located within the region of Aurora Public Schools, who have been at risk of facing state action due to their low graduation rates and unsatisfactory test scores [15]. However, it is important to consider how this information can qualify the population to be vulnerable. In a recent study from the NCRC [16], the authors discuss how gentrification could provide insight into this investigation. Gentrification, or the process in which wealthier people who move to poor areas cause the suffering or displacement of current residents, can have negative impacts on diverse, undereducated, and low-income people, especially during a global pandemic. Evidently, the study identifies the five most vulnerable tracts from this paper as a part of the most gentrified tracts in Denver [16]. According to The Denver Channel [17], gentrification is changing the demographics in these areas by relocating the number of minorities and those who are living in poverty. Therefore, gentrification is an indication that socioeconomic factors can be used to quantify vulnerable people, because it is the minorities, less educated, lower-income, and unemployed residents that are negatively impacted by a seemingly positive change within local communities.

## REFERENCES

- [1] S. Tan, A. Fowers, D. Keating, and L. Tierney, "Amid the pandemic, public transit is highlighting inequalities in cities," *The Washington Post*, 15-May-2020. [Online]. Available: <https://www.washingtonpost.com/nation/2020/05/15/amid-pandemic-public-transit-is-highlighting-inequalities-cities/>. [Accessed: 18-Feb-2022].
- [2] A. R. Khavarian-Garmsir, A. Sharifi, and N. Moradpour, "Are high-density districts more vulnerable to the covid-19 pandemic?," *Sustainable Cities and Society*, 03-Apr-2021. [Online]. Available: [https://www.sciencedirect.com/science/article/pii/S2210670721001992?casa\\_token=HCzmM9HFrUwAAAAA%3AKrzRbGJAalu50KD7fRfWTDjj7WiiX60MOEXXzDzl-pMODhRCXNDU1q2apkBVKjzE4EhYqXkxMKo](https://www.sciencedirect.com/science/article/pii/S2210670721001992?casa_token=HCzmM9HFrUwAAAAA%3AKrzRbGJAalu50KD7fRfWTDjj7WiiX60MOEXXzDzl-pMODhRCXNDU1q2apkBVKjzE4EhYqXkxMKo). [Accessed: 18-Feb-2022].
- [3] S. Hamidi, "Does density aggravate the COVID-19 pandemic?," Taylor Francis. [Online]. Available: <https://www.tandfonline.com/doi/full/10.1080/01944363.2020.1777891>. [Accessed: 18-Feb-2022].
- [4] L. de Leon, "Covid-19 infections rates differ by neighborhood across the Denver Metro," *KUSA.com*, 17-Jul-2021. [Online]. Available: <https://www.9news.com/article/news/health/coronavirus/covid-19-infection-neighborhood-denver-metro/73-90d7caba-8aa5-4daf-8a5b-6e4992d1a53c>. [Accessed: 18-Feb-2022].
- [5] M. A. Novisky, K. M. Nowotny, D. B. Jackson, A. Testa, and M. G. Vaughn, "Incarceration as a fundamental social cause of health inequalities: Jails, prisons and vulnerability to covid-19," *OUP Academic*, 08-Apr-2021. [Online]. Available: <https://academic.oup.com/bjc/article/61/6/1630/6217392?login=true>. [Accessed: 18-Feb-2022].
- [6] B. Neelon, F. Mutiso, N. T. Mueller, J. L. Pearce, and S. E. Benjamin-Neelon, "Spatial and temporal trends in social vulnerability and covid-19 incidence and death rates in the United States," *PLOS ONE*. [Online]. Available: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0248702>. [Accessed: 18-Feb-2022].
- [7] "CDC SVI documentation 2018," Centers for Disease Control and Prevention, 10-Feb-2022. [Online]. Available: [https://www.atsdr.cdc.gov/placeandhealth/svi/documentation/SVI\\_documentation\\_2018.html](https://www.atsdr.cdc.gov/placeandhealth/svi/documentation/SVI_documentation_2018.html). [Accessed: 18-Feb-2022].
- [8] E. M. Crimmins, "Age-related vulnerability to coronavirus disease 2019 (COVID-19): Biological, contextual, and policy-related factors," *OUP Academic*, 07-Sep-2020. [Online]. Available: <https://academic.oup.com/ppar/article/30/4/142/5902126>. [Accessed: 02-Mar-2022].
- [9] "Factors that contributed to undetected spread," World Health Organization. [Online]. Available: <https://www.who.int/news-room/spotlight/one-year-into-the-ebola-epidemic/factors-that-contributed-to-undetected-spread-of-the-ebola-virus-and-impeded-rapid-containment>. [Accessed: 18-Feb-2022].
- [10] B. Desye, "Covid-19 pandemic and water, sanitation, and hygiene: Impacts, challenges, and mitigation strategies," *Environmental health insights*, 14-Jul-2021. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8283044/>. [Accessed: 02-Mar-2022].
- [11] "Census Data," United States Census Bureau. [Online]. Available: <https://data.census.gov/cedsci/>. [Accessed: 02-Mar-2022].
- [12] "Coefficient of Variation," European Commission. [Online]. Available: <https://datacollection.jrc.ec.europa.eu/wordef/coefficient-of-variation>. [Accessed: 26-Feb-2022].
- [13] S.L. Cutter, B.J. Boruff and W.L. Shirley, "Social Vulnerability to Environmental Hazards," *Social Science Quarterly*, vol. 84, no. 2, Jun. 2003, pp. 242–261 [Accessed: 26-Feb-2022].
- [14] "Race, diversity, and ethnicity in Denver, CO," Best Neighborhood. [Online]. Available: <https://bestneighborhood.org/race-in-denver-co/>. [Accessed: 03-Mar-2022].
- [15] M. Wingerter, "Four Denver-area schools could face state action based on low scores," *The Denver Post*, 03-Sep-2019. [Online]. Available: <https://www.denverpost.com/2019/08/30/fewest-denver-area-schools-on-state-watch-since-2015/>. [Accessed: 03-Mar-2022].
- [16] J. Richardson, B. Mitchell, and J. Edlebi, "Gentrification and disinvestment 2020," NCRC, Jun-2020. [Online]. Available: <https://ncrc.org/gentrification20/>. [Accessed: 03-Mar-2022].
- [17] J. Porter, "Study: Denver is now 2nd most gentrified city in the nation," *The Denver Channel*, 09-Jul-2020. [Online]. Available: <https://www.thedenverchannel.com/news/local-news/study-denver-is-now-2nd-most-gentrified-city-in-the-nation#:text=The%20study%20found%20over%2027,enough%20funding%20for%20affordable%20housing>. [Accessed: 03-Mar-2022].

## APPENDIX A MATLAB CODE

```
%% data pre-processing
% read in the raw data
dataIn = readtable('eeng415Proj1Data.xlsx');
% change table to array
dataIn = table2array(dataIn);

% document which rows contain missing values
[rowsSize, colSize] = size(dataIn);
nanMat = isnan(dataIn);
dataRemoved = zeros(rowsSize, 1);
count = 0;
for i = 1:rowsSize
    for j = 1: colSize
        if nanMat(i, j) == 1
            count = count + 1;
            dataRemoved(count) = dataIn(i, 1);
            break;
        end
    end
end

% this vector contains the tracts that were removed from data set
dataRemoved = dataRemoved(1:count);

% convert table to an array, only keep numerical data
rawData = dataIn(:, 3:end);

% remove rows that contain missing data
rawData = rmmissing(rawData);
%% Observe Variances/Remove Features with Low Variability
% Relative to Others
varMat = var(rawData);
coeffOfVar = zeros((colSize - 2), 1);

% calculate coefficient of variation
for i = 1:(colSize - 2)
    coeffOfVar(i) = (std(rawData(:, i)) / mean(rawData(:, i))) * 100;
end
maxVariation = max(coeffOfVar);
featToRem = zeros((colSize-2), 1);

% compare variations to max variation
for i = 1:size(coeffOfVar)
    if coeffOfVar(i) <= (0.02 * maxVariation)
        featToRem(i) = i;
    end
end
featToRem = sort(featToRem, 'descend');
```

```

% remove features that have coefficients of variance that
% are 2% or less of the maximum coefficient of variance
for i = 1:size(featoRem)
    if featToRem(i) ~= 0
        if featToRem(i) == 1
            rawData = rawData(:, 2:end);
        else
            rawData = rawData(:, [1:(featToRem(i)-1),
                                   (featToRem(i)+1):end]);
        end
    else
        break;
    end
end
end
%% PCA
% normalize the data
rawData = normalize(rawData);

% find the covariance matrix
cVar = cov(rawData);
% find the eigenvectors and eigenvalues of the covariance matrix
[eigVec, temp] = eig(cVar);
eigVals = eig(cVar);

% match the eigenvector to the order of the largest-smallest
% eigenvalues.
[rows,cols] = size(eigVals);
sortingTemp = zeros(rows, cols + 1);
% create a matrix to keep track of order of eigenvalues
for i = 1:rows
    sortingTemp(i, 1) = eigVals(i);
    sortingTemp(i, 2) = i;
end
% sort eigenvalues based on previously defined order
eigVals = sort(eigVals, 'descend');
sortingTemp = sortrows(sortingTemp, 1, 'descend');
[rowsV,colsV] = size(eigVec);
newEigVec = zeros(rowsV, colsV);
for i = 1:rows
    newEigVec(:, i) = eigVec(:, sortingTemp(i, 2));
end

% find the variances by normalizing eigenvalues
variances = eigVals / sum(eigVals);
variances = sort(variances, 'descend');

figure(1)
bar(variances)
title('Variability in Data Explained vs Principle Components');
ylabel('Percentage of Variability in Data Explained');
xlabel('Principle Components');

```

```

% sum up variances until at least 0.9
sumVar = 0;
i = 1;
numVar = 0;
while (sumVar <= 0.9000)
    sumVar = sumVar + variances(i);
    i = i + 1;
    numVar = numVar + 1;
end
newEigVec = newEigVec(:, [1:numVar]);

% transform data points into an N x numVar matrix
pcaData = newEigVec' * rawData';
pcaData = pcaData';

fprintf('The first 10 datapoints of the transformed data
after PCA are:\n');
for i = 1:10
    fprintf('%f %f \n', pcaData(i,:));
    fprintf('\n');
end
%% Part E: Computing Vulnerability Level
[rowNew, colNew] = size(pcaData);
sumMat = zeros(rowNew, 1);
for i = 1:rowNew
    for j = 1:colNew
        sumMat(i, 1) = sumMat(i, 1) + pcaData(i, j);
    end
end

% plot vulnerability sums to view approximate distribution
figure(2)
histogram(sumMat);
title('Vulnerabilities by Tract');
xlabel('Vulnerability Value');
ylabel('Percentage of Tracts Defined by Vulnerability');

% q q plot of data to check for normality
figure(3)
qqplot(sumMat);
title('QQ Plot of Vulnerability Levels');

% Use Anderson-Darling test for Normality
[adt,p,adstat,cv] = adtest(sumMat);
test = 0;
if adt == 0
    test = "normal";
elseif adt == 1
    test = "not normal";
end
fprintf('\n');
fprintf('The test statistic for the Anderson-Darling test is:

```

```

%f. \n', adstat);
fprintf('The critical value for the Anderson-Darling test is:
%f. \n', cv);
fprintf('The Anderson-Darling Test result is: %s. \n', test);

% calculate mean and standard deviation to use in
% vulnerability analysis
meanVul = mean(sumMat);
stdVul = std(sumMat);

tractVul = zeros(rowNew, 2);
for i = 1:rowNew
    tractVul(i, 1) = i;
end

% assign values 1-5 for least to most vulnerable
for i = 1:rowNew
    if sumMat(i) < (meanVul - (2*stdVul))
        tractVul(i, 2) = 1;
    elseif (sumMat(i) > (meanVul - (2*stdVul))) &&
        (sumMat(i) < (meanVul - stdVul))
        tractVul(i, 2) = 2;
    elseif (sumMat(i) > (meanVul - stdVul)) &&
        (sumMat(i) < (meanVul + stdVul))
        tractVul(i, 2) = 3;
    elseif (sumMat(i) > (meanVul + stdVul)) &&
        (sumMat(i) < (meanVul + (2*stdVul)))
        tractVul(i, 2) = 4;
    elseif sumMat(i) > (meanVul + (2*stdVul))
        tractVul(i, 2) = 5;
    end
end

sTemp = sortrows(tractVul, 2, 'descend');
[B,I] = maxk(sumMat, 5);

fprintf('\nThe 5 most vulnerable tracts are: \n');
for i = 1:5
    fprintf('Tract number: %d \n', dataIn(tractVul(I(i), 1), 2));
end

% create final table showing vulnerability level for each tract
finalTractVul = zeros(rowNew, 2);
for i = 1:rowNew
    finalTractVul(i, 1) = dataIn(i, 2);
end

% matrix containing vulnerability ranking for each tract
for i = 1:rowNew
    finalTractVul(i, 2) = tractVul(i, 2);
end
writematrix(finalTractVul,'tractsVuln23.csv')

```



## APPENDIX B

### FINDING DATA DISTRIBUTION

In order to analyze how the vulnerability data is distributed, a histogram and QQ-plot were plotted for the vulnerability levels corresponding to each census tract. It was noticed that the vulnerability level data resembles a normal distribution. However, there were large tails in the QQ-plot. This could indicate that there are more extreme values than a normal distribution to the left and right of the mean. The histogram and QQ-plot are displayed below in Figures 3 and 4 respectively.

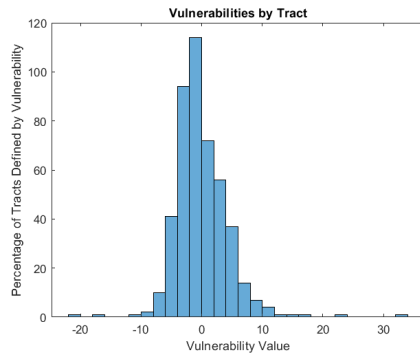


Fig. 3. Histogram of Vulnerability Levels

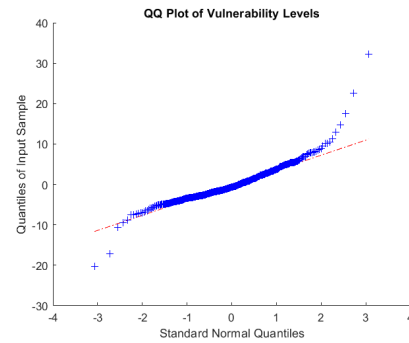


Fig. 4. QQ Plot Comparing to Normal Distribution