Quantifying the COVID Situation in Peru:

An Analysis of Case Ratios, Reproductive Numbers, and Herd Immunity

Jacob A. Unterbrink

Spring Semester, 2023

DSA 5900

4 Credit Hours

Supervised by Dr. Charles Nicholson

1. Introduction

The COVID-19 pandemic has had a pervasive impact on people and counties around the globe. It permeates all facets of everyday life, and the need to understand it – and its propagation dynamics – are imperative; not just for a return to normal, but to save lives. This is especially true for the country of Peru, who experienced the highest COVID-related death rate in the world: 5.89 fatalities per 1,000 residents [1]. For reference, this is more than twice the ratio found in the United State where there have been 2.38 fatalities per 1,000 residents. Neighboring South American countries such as Brazil (2.86), Columbia (2.48), and Argentina (2.54) also experienced much lower fatality rates than that found in Peru during the same period. The reasons for this are still largely unknown, as Peru has implemented some of the strictest preventative measures in the world including quarantines, masking, and social distancing (measures that are still in place to this day).

To unravel and remedy this issue, a team from the Universidad Nacional de San Agustin de Arequipa, Perú (UNSA) reached out to the University of Oklahoma (OU) to form a research collaboration. Specifically, members of the UNSA team requested guidance on how to produce journal-quality research and lay a foundation for a stronger healthcare system in the future through partnerships with local and national governing bodies.

The OU side of the collaboration is composed of multiple teams, and the project that follows was completed as a task under team two, whose primary focus is on modeling and analysis. This project, in part, serves to address a key hypothesis held by the research team: COVID cases in Peru are severely underreported. To do this data were collected that reflect COVID cases and deaths in the United States, Oklahoma, Peru, and Arequipa. Then death-to-case ratios were computed and compared across dominant COVID strains. Additionally, an expert's COVID models for Oklahoma and Arequipa were recreated, and forecasted values were extracted to derive herd immunity and effective reproductive numbers in Oklahoma and Arequipa.

2. Objectives

A critical issue in studying the spread of COVID-19 in Peru relates to the quality of data available. Since the healthcare system was largely unable to account for the sheer number of

cases, there are likely many infections that went undocumented. Not to mention the swaths of missing data in key features such as ICU beds available and in use, that might shed light on hospitals' capacity to assist new patients. Thus, it has been challenging for researchers to paint a realistic picture of the COVID landscape. By investigating death to case ratios, coupled with an examination into herd immunity and effective reproductive numbers, evidence can hopefully be uncovered relating to whether cases in Peru are truly underreported.

This project will meet various technical and personal learning objectives. Technically, it will demonstrate mastery of the data ingestion, exploration, and cleaning process. It will also display high-level understanding of modern data science programming languages and tools such as R and Python. Death to case ratios will be computed for each region and broken down by dominant COVID variants. These ratios will be used to quantify what we might expect COVID-related deaths to look like in Peru and the Arequipa region. An expert's COVID models for Oklahoma and Arequipa will be reproduced, and values will be extracted to derive population immunity, as well as effective reproductive numbers for both regions. These actions will help address the hypothesis held by the research team that cases are severely underreported in Peru. Finally, a research paper will be produced for publication in a peer-review journal.

Regarding individual learning objectives, this project will allow for a hands-on application of the data science method from start to finish (deployment may be excluded until future work is complete) including data understanding, ingestion, exploration, preparation, process validation, modeling, evaluation, and deployment. It will also include reading and discerning relevant information from current research and applying this knowledge to the specific problem instance. This project will teach, in writing, how to tell a story with data that can be understood by individuals without prior domain knowledge (this includes being able to explain technical topics clearly and concisely). It will also allow for the exploration of data science tools, such as Tableau and Power BI, which will possibly be used on the problem instance if possible/practical. It will also be beneficial in learning how to explain the uses and limitations of the methods used.

3. Data

3.1. Ingestion

There were four datasets used in total. Oklahoma and U.S. data were originally pulled from a GitHub repository run by Johns Hopkins University. Peru and Arequipa data were originally gathered from Peru government agencies. Copies of these data are currently being stored in a GitHub repository owned by OU Analytics. While there were not many obstacles in acquiring the raw data, the data sets had to undergo numerous transformations prior to use on the project. These modifications are outlined in the sections that follow.

3.2. Exploration and Cleaning of the U.S. & Oklahoma Data

The U.S. and Oklahoma data are both subsets of a larger parent set (this original data frame had entries for each state and the U.S. as a whole) and therefore contain the same features. These include a measure for cases, deaths, and the date of each entry. It's worth noting that cumulative measures, as well as cases and deaths per 1,000 capita were also present but provided essentially the same information as pure cases and deaths and were therefore not referenced. In the U.S. data, values were reported daily throughout the period. In Oklahoma, values were reported daily until November 3rd, 2021, but were reported weekly afterwards. There were no missing values.

The U.S. and Oklahoma data both contain 1,142 rows and entries began and ended on the same dates: January 1st, 2020, and March 9th, 2023, respectively. In both regions, the feature reflecting cases seemed reliable, and followed a trend that one might expect. However, upon inspecting deaths, it was observed that some values were almost certainly misreported (Figure 1). Specifically, in the U.S., a particularly large spike occurs in January of 2021 and shortly after there is a negative value reported. There were two similar spikes in value when examining the Oklahoma deaths (both occurring in 2021). While a few other spikes are present, it is not as clear if these are misreported or simply unusual realizations. For this reason, these values were left untouched.

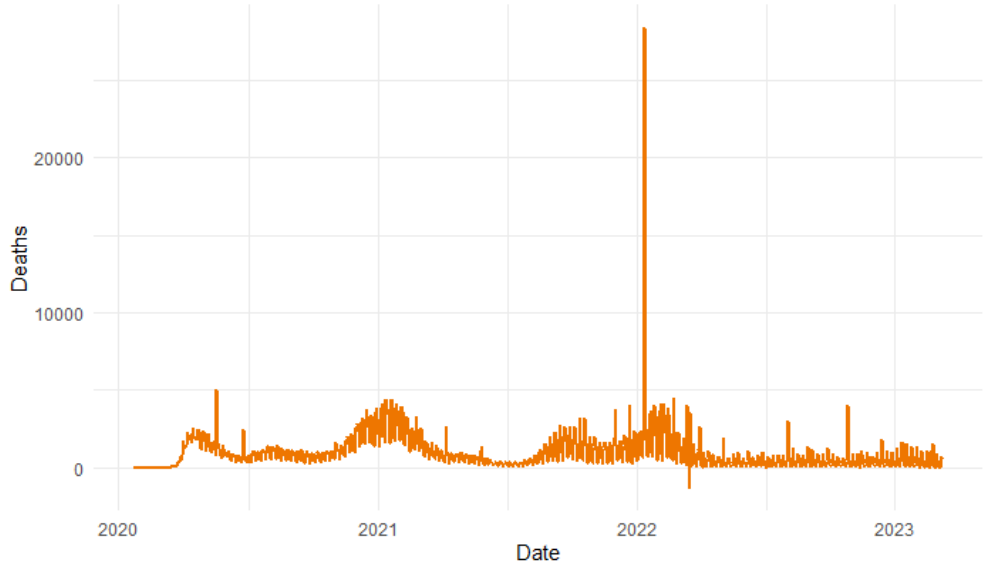

Figure 1: U.S. COVID-19 deaths.

The negative value was set to zero, while the spikes were set to the average of the two nearest neighbors. For example, the largest spike in Oklahoma happened on April 7th, 2021, so it was replaced with the number of deaths on April 6th, plus the number of deaths occurring on April 8th, divided by two. This seemed reasonable, given that deaths generally don't vary widely day-to-day.

As mentioned previously, a few spikes still existed in the death data. For this reason, it was desirable to smooth the curves somewhat. A common approach is to compute the seven-day average, but because Oklahoma values were already reported weekly after November 3rd, 2021, it seemed logical to find the weekly sums instead. Since the reported weekly values fell on Wednesdays, the derived weekly values also fell on Wednesdays to avoid any potential issues with overlap. Seven-day sums were computed for both cases and deaths and completed for both U.S. and Oklahoma data sets (sample of results shown in Figure 2).

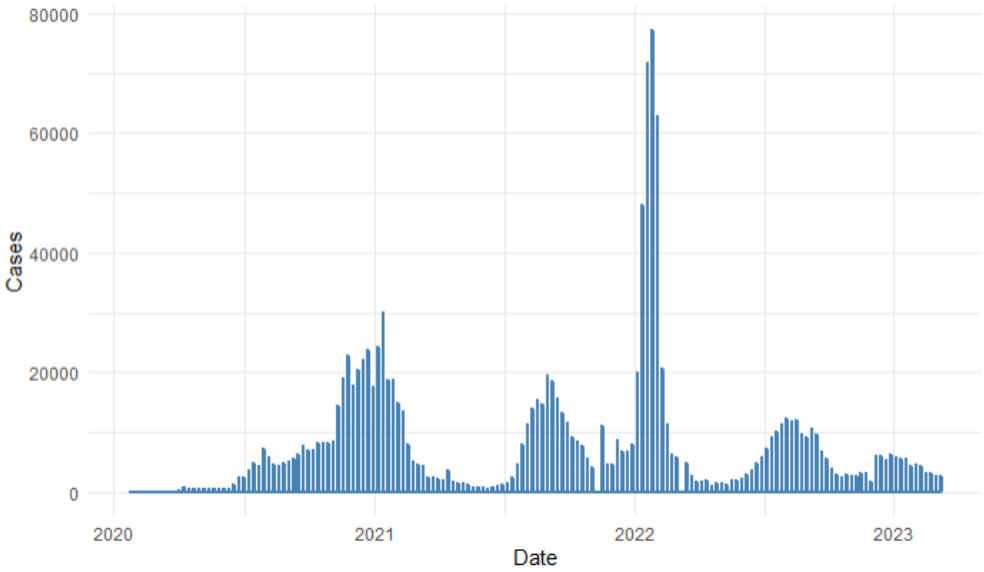

Figure 2: Seven-day summed values for Oklahoma cases.

When someone becomes infected with a fatal case of COVID, some amount of time passes between the initial infection and death. If this time is not accounted for, it could lead to inaccurate conclusions when comparing death to case ratios (since deaths at time *t* are not caused by cases as time *t*, but rather cases at time *t-n* where *n* is some lag). To correct for this, deaths were shifted back so that peaks and troughs roughly aligned among cases and deaths. This shift accounted for a two-week lag – implying that in the U.S. and Oklahoma, on average, there is a 14-day window between initial infection and death.

Finally, a feature was engineered to represent the dominant COVID variant in the U.S. and Oklahoma through time. This was done so that death to case ratios could later be computed by dominant variant and provide a more accurate representation of COVID cases and deaths than an aggregated score. To establish variants and date ranges, CoVariants.org was referenced. The dominant strains in the U.S. are as follows: original strain (prior to March 15th, 2021), alpha (March 15th, 2021, through June 21st, 2021), delta (Marth 15th, 2021 through December 20th, 2021), and omicron (after December 20th, 2021). For Oklahoma, the dominant strains were similar: original strain (prior to March 15, 2021), alpha (Marth 15th, 2021 through June 6th, 2021), delta (June 6th, 2021, through December 20th, 2021), and omicron (after December 20th, 2021) [2].

3.3. Exploration and Cleaning of the Peru & Arequipa Data

Like before, the Peru and Arequipa data were derived from a parent set. These original data were reported at the regional level (Arequipa is one such region, but there are 25 in total). The data set contained measures for the date, daily cases, cumulative deaths, and cumulative features for a variety of COVID-related tests and test results (PCR tests, serological tests, etc.). The earliest observation was recorded on March 13th, 2020, while the last entry for each region was on April 5th, 2022. However, not all regions began collecting data on the same day. For instance, in Amazonas, the first entry was on April 5th, 2022. A subset was taken to form the Arequipa data set, while all the regions were aggregated (using "date" as the grouping variable) to form the country level Peru data frame. While there were no missing values, the data required cleaning prior to use. The methods that follow were applied to both the Arequipa and Peru data, unless otherwise specified.

It was desirable to convert the cumulative features into daily measures to later compare ratios, and to better observe trends over time. Ordinarily this would be a straightforward task, but there were clear issues with the current values (see Figure 3). There were days when values would drop to zero, before immediately rising to expected levels. On other days, there would occasionally be unusual jumps in value along the curves.

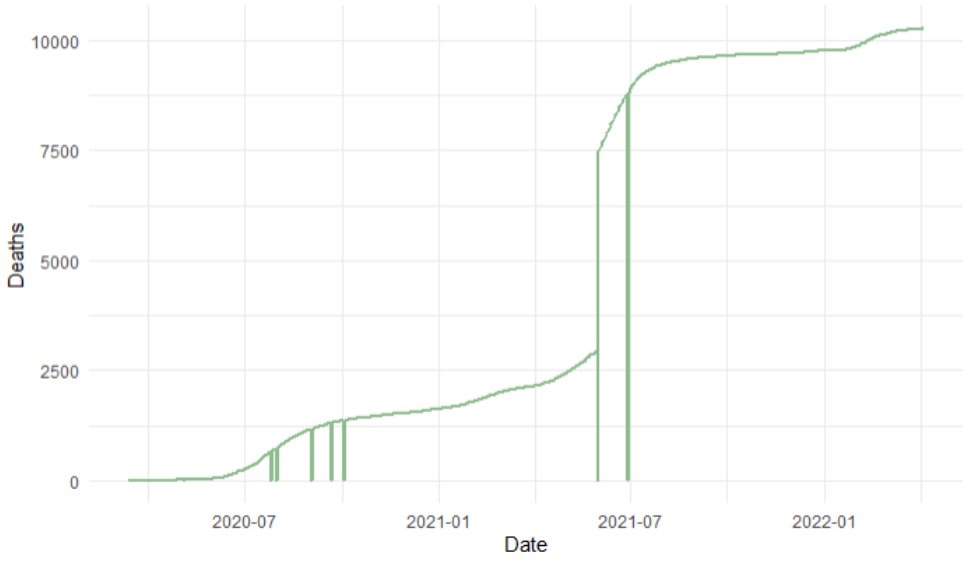

Figure 3: Original cumulative feature for Arequipa deaths.

Some of these problems are easier to explain than others. For instance, towards the beginning of the period jumps would be expected in the Peru data because not all regions began reporting on the same day. Moreover, late in 2021, the country of Peru changed how they handled reported values, which would explain the more drastic shift later in the period. The dips to zero are somewhat more mysterious but could be caused by errors in the reporting software used. It is known that the Peruvian healthcare system struggled to keep up with the volume of incoming data, and it is possible that some systems crashed resulting in zero values. Another possibility is that these instances were originally missing from the data but were coded as zeros for the sake of avoiding computational errors.

Whatever the reason, these zero occurrences were certainly inaccurate. To correct this, a function was written that could be applied to each cumulative variable. It would take as its argument some feature, then compare the i^{th} observation, with the value at i-1. If the value at location i was less than the value at i-1, then the value at observation i was replaced with the value at i-1. This approach was taken for two reasons. First, it would ensure that each observation was either increasing or remaining constant (thus following the laws of cumulative distributions). Second, the average-of-neighbors approach used earlier would have been problematic if there were two consecutive decreases since the average used to replace both points would still be less than the i-1 observation.

While this method handled the clearly erroneous dips in value, it doesn't account for the jumps along the cumulative plots (the ones likely caused by reporting irregularities, or by regions beginning to report values on different dates). For the purposes of this project, these issues could instead be addressed after converting the cumulations to daily measures.

From these cumulative features, daily measures were computed by subtracting the cumulation's first lag. However – because of the jumps in cumulative value – the resulting distributions have periodic spikes that do not reflect reality (see Figure 4). To address this issue, a function was written. It would take two arguments: the variable in question, and the number of spikes present. Since a spike is simply the maximum value along a distribution, it was easy for the function to identify the instance. After identifying, the point would be replaced by the average of its two nearest neighbors in the same manor explained previously. If the number of spikes was specified to be greater than one – as was the case in the daily deaths feature – the function would replace the first maximum value, then re-evaluate and rerun the algorithm for the specified number of spikes. This method was used on the daily features for deaths (three spikes), AG tests, total tests, and negative tests (all one spike).

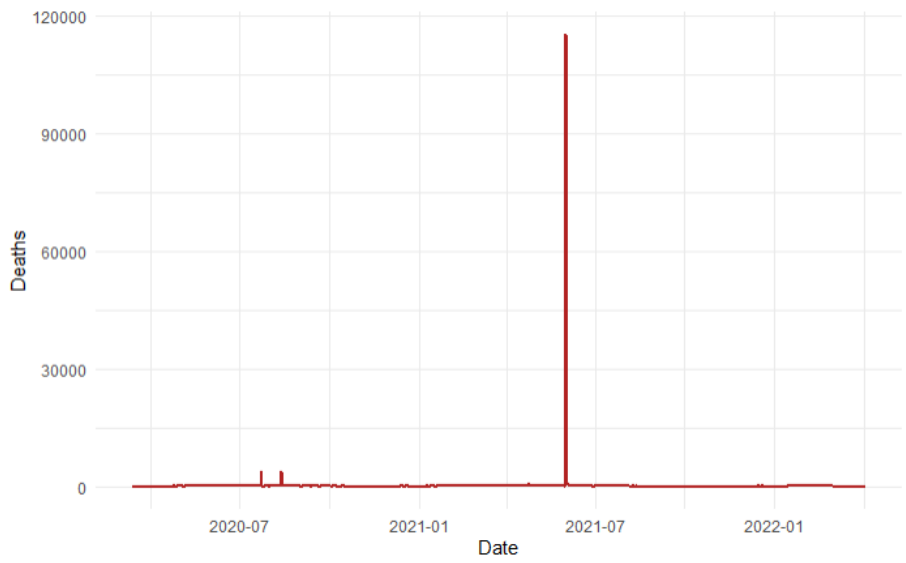

Figure 4: Peru daily deaths prior to handling spikes.

Admittedly, the above approach does have limitations. While some reporting issues are obvious and clearly documented (regions beginning to report data on different days, and the massive influx of reported deaths later in the period) others may be more subtle. Take daily Peru deaths for instance (shown in Figure 5); there exist a few notable jumps in value towards the beginning of the distribution but there is currently not enough information to know whether these were caused by a change in reporting or if these days genuinely experienced an unusually high number of deaths. Due to this uncertainty, these points were not replaced.

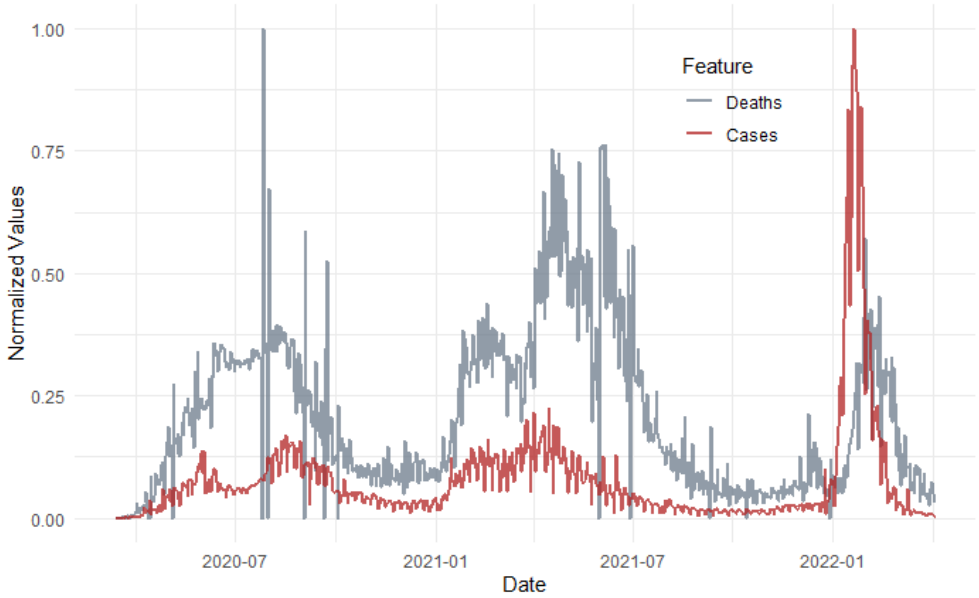

Figure 5: Peru normalized cases and deaths after handling spikes in deaths.

Once daily features were engineered, seven-day sums were calculated for cases and deaths to help smooth the curves. The summed values were always reported on Wednesdays to be consistent with the U.S. and Oklahoma data. Again, deaths were shifted back so that peaks and troughs aligned, and cases matched their associated deaths. In the U.S. and Oklahoma this shift accounted for two weeks, however in Peru and Arequipa, deaths were only shifted back one week – implying that, on average, there is only a 7-day window between initial infection and death in Peru. This might be evidence of a weaker population health relative to the U.S. or, as the research team suspects, potential evidence that Peruvian citizens occasionally wait longer to seek medical treatment after becoming sick than U.S. citizens. In the latter case, a patient would already be well into their battle with COVID by the time their infection is accounted for. Thus, if their death is reported a week later, the timeline would seem much shorter than in the U.S.

Lastly, a feature was engineered to denote the dominant COVID strain in Peru and Arequipa through time. CoVariants.org was again referenced for this task, however Arequipa numbers were not available. Because of this, Peru strains and date ranges were used as a proxy. The dominant strains are as follows: original strain (prior to February 15th, 2021), lambda (February 15th, 2021, through August 30th, 2021), delta (August 30th, 2021, through December 6th, 2021), and omicron (after December 6th, 2021) [2].

3.4. Further Exploration of Peru Data

For the purposes of this project, cases and deaths were the primary features of concern. However, the Peru data is known to be somewhat unreliable for reasons listed previously and therefore it was deemed valuable to explore all features in further detail. This was requested by the Peru team and serves to identify potential pitfalls in the data.

To better understand the variable distributions, density plots were generated. In some cases, feature values were normalized so that multiple distributions could be compared side-by-side. As a result of this, it could be seen that AG tests, Serological tests, and PCR tests followed similar spreads; as did total tests and negative tests, and finally cases and deaths. Bimodal distributions are apparent in the deaths, total tests, and negative test features. A possible explanation is the introduction of new COVID strains across time. However, it seems peculiar that, if this were the case, the distribution of cases would remain unimodal.

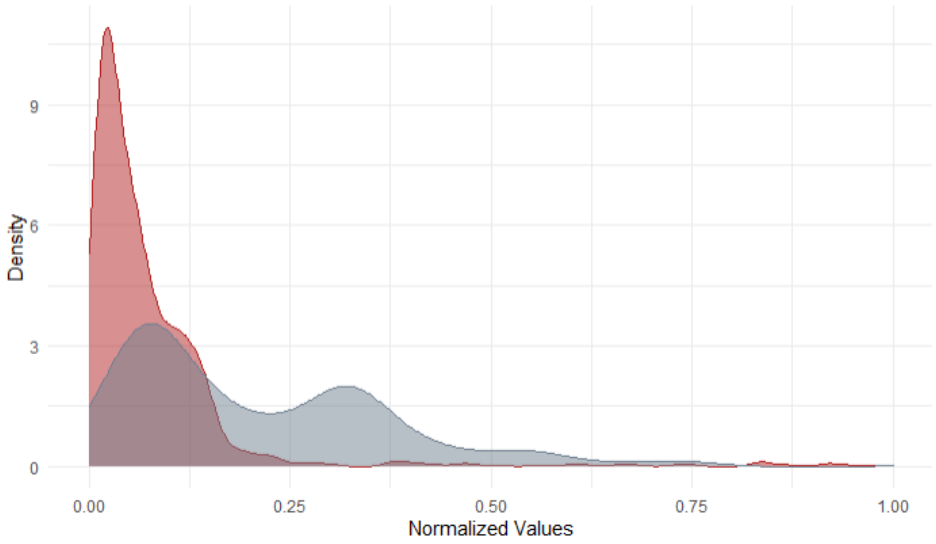

Figure 6: Density plot of Peru cases and deaths.

A correlation analysis was also conducted, and the results can be seen in the heatmap below. Cases have a strong (greater than 0.500) positive correlation with all features relating to testing, with the exception of serological tests. The relationship was still positive, but the Pearson correlation coefficient was only 0.056. Deaths were also correlated with the test-oriented features, with a coefficient hovering around 0.300 for each variable. Strangely, total tests were negatively correlated with positive serological test results (-0.202) and may warrant further investigation.

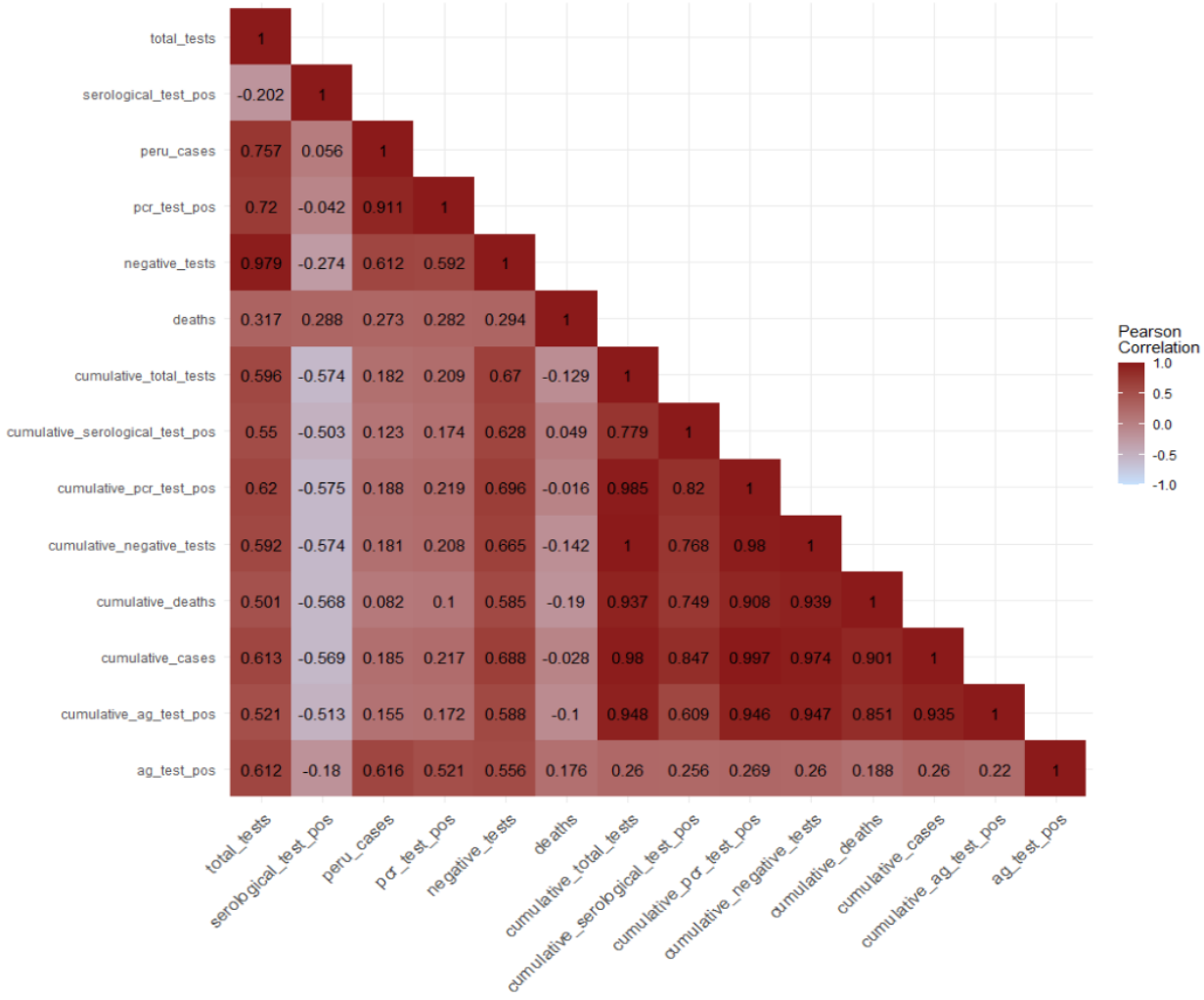

Figure 7: Heatmap of Peru features.

To assess the prevalence of potential outliers, a function was written that would analyze each feature. The function utilizes the IQR method and determines how many points fall above (below) the upper (lower) inner fence for the feature in question. This is similar to how outliers might be identified by examining a boxplot but has the benefit of automatically counting the number of possibly problematic observations. The output from this function can be seen in the table below.

Table 1: Potential Outliers in Peru Features

Variable	No. Outliers	Share of Whole
peru_cases	36	4.77%
deaths	12	1.59%
total_tests	27	3.58%
negative_tests	16	2.12%
pcr_test_pos	65	8.62%
serological_test_pos	61	8.09%
ag_test_pos	73	9.68%
cumulative_cases	0	0.00%
cumulative_deaths	0	0.00%
cumulative_total_tests	0	0.00%
cumulative_negative_tests	0	0.00%
cumulative_pcr_test_pos	0	0.00%
cumulative_serological_test_pos	48	6.37%
cumulative_ag_test_pos	49	6.50%

Outliers are seemingly the biggest issue for the PCR, Serological, and AG test variables. In each of these, at least 8.09% of instances are flagged for being potentially problematic. Oddly, the cumulative serological test variable was found to have 48 outliers, and again may warrant a deeper look into the serological feature. Looking at cases and deaths, it is seen that 4.77% and 1.59% of instances are possible outliers respectively. There is no strong evidence to suggest that any of these outlying points are recorded erroneously, and therefore they were left untouched.

4. Methodology

4.1. Comparing Ratios Across Time

In the sections that follow, it was assumed that reported deaths were a more reliable and accurate measure of the impact of COVID in Peru than reported cases. This is for several reasons. First, when someone contracts COVID, it is possible that they never seek medical assistance and therefore their case is never registered. Second, as mentioned previously, the Peruvian healthcare system has struggled to keep up with the volume of cases introduced by the pandemic, and therefore it is possible that some cases were never officially logged. On the other hand, when someone experiences a COVID-related death it is almost certainly recorded; even if the individual did not seek medical help through the duration of their sickness, the COVID pathogens would be apparent in the post-death analysis.

Under this assumption, an attempt was made to address the research team's hypothesis: cases in Peru are severely underreported. To do this, death to case ratios were computed based on a region's dominant variants. This allows for a direct comparison of reported cases and deaths and gives an idea of the reported impact each strain has on the associated region. Moreover, if the additional assumption is added that each strain's mortality rate in Peru and Arequipa should be roughly equivalent to that found in the U.S. and Oklahoma, then these ratios can be used to derive expected case numbers in these regions.

4.2. SEIR Modeling

Before herd immunity and reproductive numbers could be estimated and assessed (described in detail in the next section), a susceptible and immune population had to be estimated. An expert in the field of epidemiology, who the research team consults on occasion, donated two Susceptible-Exposed-Infected-Recovered (SEIR) models. One model was fit to Oklahoma data, and the other to Arequipa. These models were originally composed in Microsoft Excel but were recreated using a Python-backed web application developed by another team member. This web application is used throughout the team for a uniform model building and optimization framework that can be understood by everyone involved. These models were recreated using this application so that other team members had a baseline model for both Oklahoma and Peru to test their own algorithms on. Previously no models had been built for these regions under this framework.

A SEIR model is a type of compartmental model commonly used in epidemiology to track the spread of infectious diseases. These types of models were first developed by Kermak and McKendrick between 1920 and 1940 [3]-[5]. The idea is that a population can be divided into various compartments and flow from one compartment to another based on a set of parameters. These types of models are useful to understand the mechanisms underlying disease propagation, but ignore factors like geography, population heterogeneity, etc.

Under the SEIR framework, a population flows from a susceptible state (S) to exposed (E), infected (I), recovered (R), and over time back to susceptible (S). In the model provided, individuals can also flow from susceptible (S) to recovered (R) immediately by becoming vaccinated. The initial compartment values are set to zero, save for susceptible and infected.

The number of susceptible people at the beginning of the period is equal to the population total, *N*, while the initial number of infected people is set to 1.

Individuals move through these compartments based on a series of parameters. These include λ (the rate from susceptible to exposed), *f* (rate from exposed to infected), *r* (rate from infected to recovered), *w* (rate from recovered to susceptible), and *v* (rate from susceptible to recovered). People are born into the population at rate μ and can become diseased naturally while in any compartment at rate μ (the model assumes a constant birth and death rate). An example of this process can be found in the figure below.

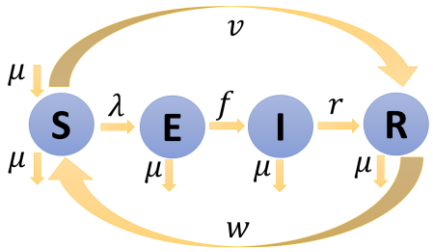

Figure 8: Population flow in a SEIR model.

The Oklahoma model starts on February 21st, 2020, and forecasts values through June 30th, 2023. The Arequipa model begins on March 6th, 2020, and ends on January 15th, 2023. The parameter values were updated by the expert throughout time as needed. The full set of parameter values and parameter value changes can be observed in the tables below.

Table 2: Oklahoma SEIR Model Parameters

Parameter	Initial Value	W1:W4	W Dates	V1:V6	V Dates	β1:β12	β Dates
N	3943000	0.002739726	2020-01-10	0.002321429	2020-12-08	5.49497E-08	2020-04-01
β	1.26807E-07	0.003030303	2021-01-10	0.005222222	2021-02-01	4.2269E-08	2020-07-01
F	0.5	0.004444444	2021-12-25	0.002576923	2021-05-01	5.07228E-08	2020-08-16
R	0.1666667	0.001666667	2022-03-01	0.002692308	2021-08-01	6.6785E-08	2020-11-05
W	0			0.002307692	2021-12-15	8.03111E-08	2020-12-07
μ	3.65297E-05			0.002076923	2022-01-01	7.81976E-08	2021-02-01
V	0					9.08783E-08	2021-03-01
						1.14126E-07	2021-05-01
						1.60622E-07	2021-07-07
						1.7753E-07	2021-11-01
						3.17017E-07	2021-12-25
						3.80421E-07	2022-06-15

Note: λ is a function of β, and is equal to β * I_t where I_t is the number of infections at time *t*.

Table 3: Arequipa SEIR Model Parameters

Parameter	Initial Value	W1	W Dates	V1:V4	V Dates	β1:β5	β Dates
N	1383000	0.002222222	2020-10-01	0.000214286	2021-04-01	1.56664E-07	2020-04-01
β	3.61533E-07			0.000611111	2021-08-01	1.6269E-07	2020-07-01
F	0.5			0.000615385	2021-11-15	1.26537E-07	2020-09-01
R	0.1666667			0.000384615	2022-03-01	1.50639E-07	2021-01-01
W	0					2.41022E-07	2021-12-01
μ	3.65297E-05						
V	0						

4.3. Basic & Effective Reproductive Numbers

Once a susceptible and immune population (immune individuals are those residing in the recovered SEIR compartment at time *t*) was estimated, the basic and effective reproductive numbers could be estimated. The basic reproductive number of a disease is a theoretical value that describes the number of secondary infections one might expect from introducing one infected person to a completely susceptible population (this value is represented by R0, or R-naught). The effective reproductive number is the number of secondary infections caused by a disease in a population that is not completely susceptible (represented by Rt) [6].

R0 can be found for each COVID variant but remains constant throughout time. However, Rt changes and can be used to analyze the effectiveness of government interventions over time. If Rt decreases, that means that the number of secondary infections is decreasing, and interventions are likely working as intended. If Rt drops below one, then a disease is expected to die off over time [6].

Researchers have estimated the following values for R0 based on COVID strains: 3 for the original strain, 5.08 for delta, and 9.5 for omicron [7] - [9]. A R0 value of 3.38 was used for alpha, as alpha was estimated to be between 30-50% more infectious than original COVID (a more scientific measure of alpha's R0 value was difficult to come by, as alpha seems to have less research surrounding it) [10]. Lambda also had limited research since it primarily affected Peru, however, an R0 value of 4.04 was estimated based on claims that it was more contagious than the original strain of COVID, and less contagious than delta [11]. The value 4.04 is the midpoint between COVID's original R0 and delta's R0. Finally, an R0 value of

14.4 was included for omicron subvariants BA4 and BA5, since it was much larger than the R0 value for the original omicron variant [12]. Originally, researchers estimated BA4 and BA5's R0 value to be 18.5 but this has since been debunked. Scientists still debate what the true R0 value should be but agree that it falls between 9.5 (from original omicron) and 18.5. For this reason, 14.4 was chosen since it is the midway point between these two values. Rt can be estimated by multiplying R0 by the fraction of susceptible individuals in a population [6]. R0 was changed throughout time to reflect the dominant COVID variant.

4.4. Herd Immunity and the Herd Immunity Threshold

Herd immunity is denoted as the fraction of a population residing in the recovered compartment at time *t*. The herd immunity threshold represents the percentage of a population needed to be immune for a disease to eventually die off. This threshold is calculated by the equation below, and R0 was updated based on the dominant COVID strain at the time [6].

$$H.I.T. = \frac{(R0 - 1)}{R0}$$

4.5. Process Validation

The above processes were validated in two ways. First, weekly meetings were held with the project sponsor to ensure that the work completed aligned with the goals of the project. Second, an expert epidemiologist – Dr. Aaron Wendelboe – was consulted to validate the approach taken and make recommendations based on what he believed would add the most value to the Peru team.

5. Results & Analysis

5.1. Ratio Outcomes

Some interesting patterns can be found by examining the death to case ratios listed in Table 4. Unsurprisingly, U.S. and Oklahoma numbers tend to be similar, as do Peru and Arequipa values. The highest ratio for the U.S. was associated with the alpha variant with three deaths reported for every 100 infections. In Oklahoma, the original strain was the most lethal with almost two deaths per 100 infections. In both Peru and Arequipa, the most dangerous variant

by far was lambda. This strain resulted in roughly five deaths per 100 infections, a devastating number considering how long lambda remained dominant.

Unfortunately, these lambda results cannot be compared to the U.S. directly since the variant largely avoided domestic soil. However, one-to-one comparisons can be made with the remaining strains. For instance, when examining the original strain and delta, Peru and Arequipa suffered nearly three times as many deaths per infection than what was reported in the U.S. and Oklahoma.

Table 4: Death to Case Ratios By Dominant Variant

Dominant Variant	U.S.	Oklahoma	Peru	Arequipa
Original Strain	0.0185493	0.0112726	0.0306500	0.0342808
Alpha	0.0133796	0.0343505	NA	NA
Lambda	NA	NA	0.0414992	0.0549355
Delta	0.0116058	0.0134179	0.0340658	0.0375343
Omicron	0.0062640	0.0070175	0.0082101	0.0056314

Omicron, although highly infectious, did not cause many fatalities and less than one death was reported per 100 infections in each region. It is important to remember however, that cases are expected to be underreported in Peru and Arequipa and therefore the real ratios in these regions may be lower than what is listed above.

With this in mind, plots were made to show what expected cases and deaths might look like in Peru and Arequipa (see Figures 10 and 11) if the ratios were the same as those found in the U.S. and Oklahoma. Only the plots reflecting U.S. ratios applied to Peru and Arequipa were included below, but Oklahoma ratios resulted in the same overall trend.

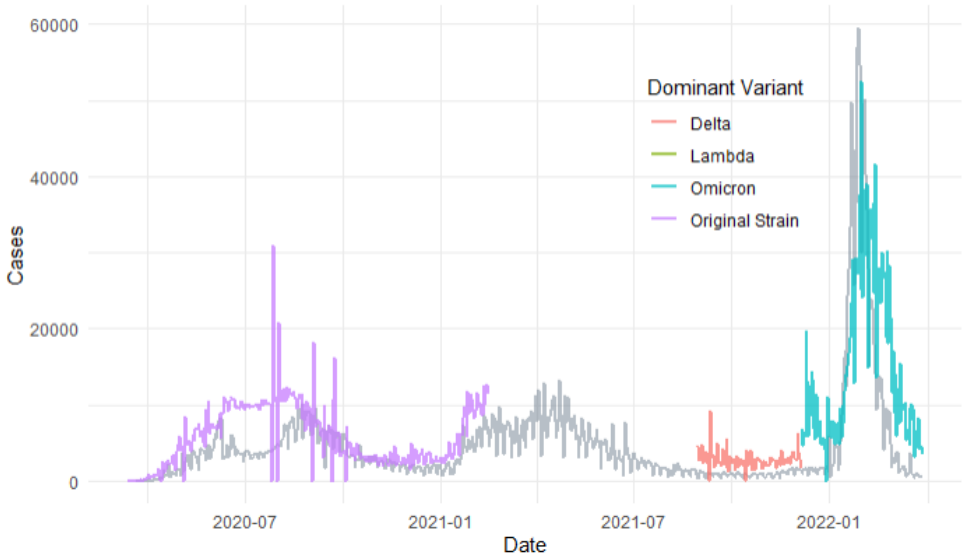

Figure 10: Peru cases (gray) vs. expected cases (colored).

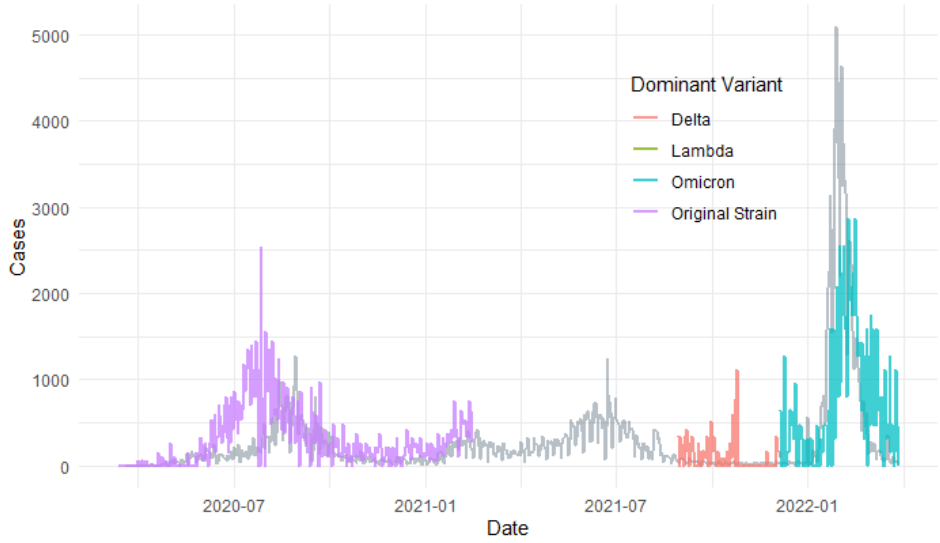

Figure 11: Arequipa cases (gray) vs. expected cases (colored).

From the above, it is seen that the number of expected COVID cases quickly surpassed the reported number of cases early in the period. As time passed, the number of expected cases more closely followed reported numbers but were still generally greater in value. It was not possible to calculate expected cases based on a lambda ratio, since this variant was not present in the U.S. but when delta takes over, a similar pattern is observed. Expected cases

were again much higher than what was reported. Interestingly, in Arequipa the number of expected and observed delta cases are nearly one-to-one later in the variant's dominance.

This changes quickly however, once omicron takes over. Arequipa's reported omicron cases were surprisingly greater than what the U.S. ratio suggests would be expected. A possible explanation for this phenomenon could be that since Peru's variants and date ranges for dominance were used as a proxy for Arequipa's variants and date ranges for dominance, the results could be somewhat unreliable in this instance.

5.2. SEIR Model Fits

The SEIR models were implemented because data on susceptible and recovered populations were not readily available and estimates were desired. But because these observed values were not present, it was impossible to compare the quality of these two forecasted compartments directly. However, the models also predicted infections, and by comparing the predicted number of cases with the observed number of cases, an idea can be formed of how well the model fits the data. Below are figures reflecting predicted vs. observed values.

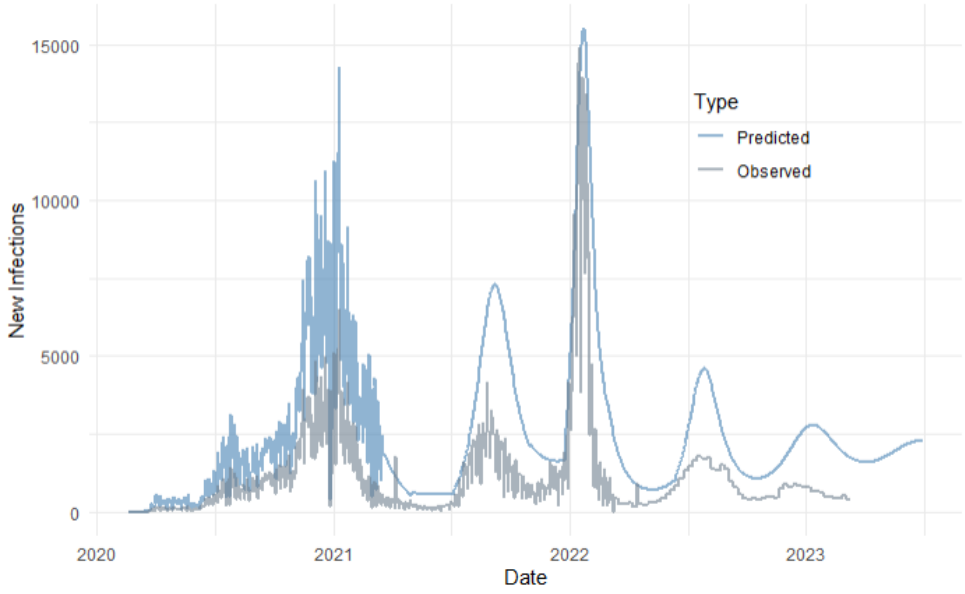

Figure 12: New Oklahoma infections, predicted vs. observed.

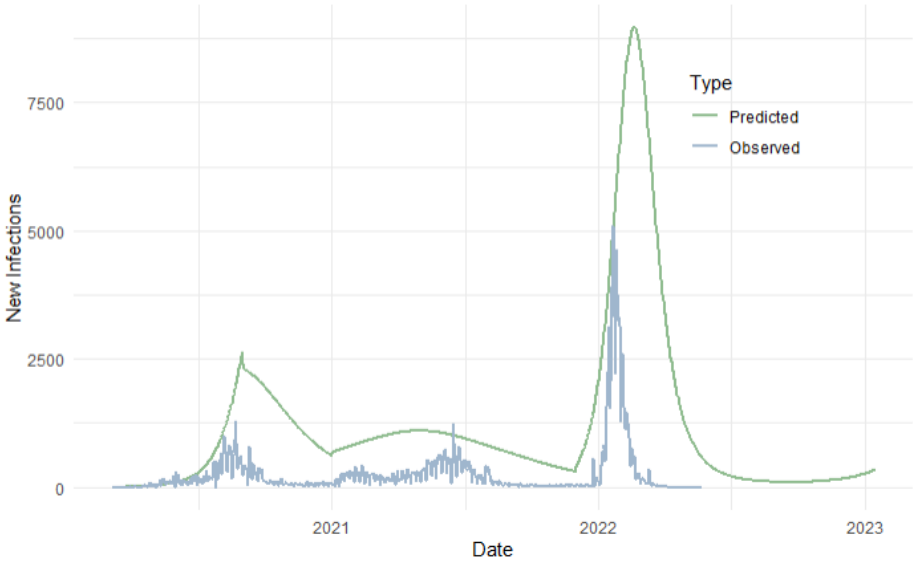

Figure 13: New Arequipa infections, predicted vs. observed.

Based on these results, the Oklahoma model does a pretty good job of predicting new infections. Moreover, this model has an RMSE of 2,071.806. The Arequipa model on the other hand captures the overall trend, but almost always overestimates the true number of new infections. This was suspected to be an issue from the start given how difficult Arequipa and Peru data has been to model in the past. While not ideal, it at least provides a starting point to work from. Given these results, the true susceptible population is likely underestimated and that will need to be considered in the next section. The RMSE of this model is 2,102.709.

5.3. Effective Reproductive Number Through Time & Herd Immunity

In Oklahoma, after the onset of COVID the basic reproductive number steadily decreased from an initial value of three down almost to one (again, if the effective reproductive number drops below one the disease is expected to die off). However, once alpha became dominant, Rt increased somewhat but kept constant. Delta introduced the first big jump in Rt – reaching about two secondary infections per case – but was short lived, as Rt again dropped near one. When omicron became dominant, a significant jump was observed and Rt reached an all-time high. After this jump, the values decreased and oscillated until they settled around 1.5. What all this seems to imply is that although Rt never touches one, the preventative measures that were put in place in Oklahoma seemed to work well. Each time a new strain was introduced,

the effective reproductive number increased as expected, but always seemed to come back down over time. This is shown in Figure 14.

Further success can be seen by examining the herd immunity threshold (Figure 15). The share of Oklahomans who have become immune to COVID, either through vaccination or by surviving the illness, is relatively high. Not high enough to surpass this threshold, but to approach it. By doing so, COVID has become less severe in Oklahoma and most people's lives have returned to normal.

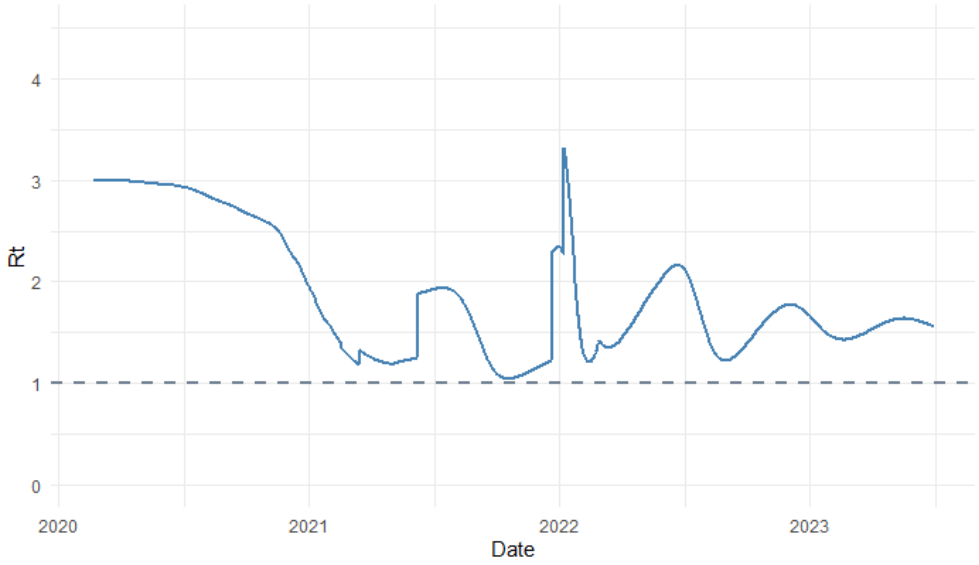

Figure 14: Approximation of the Oklahoma effective reproductive number through time.

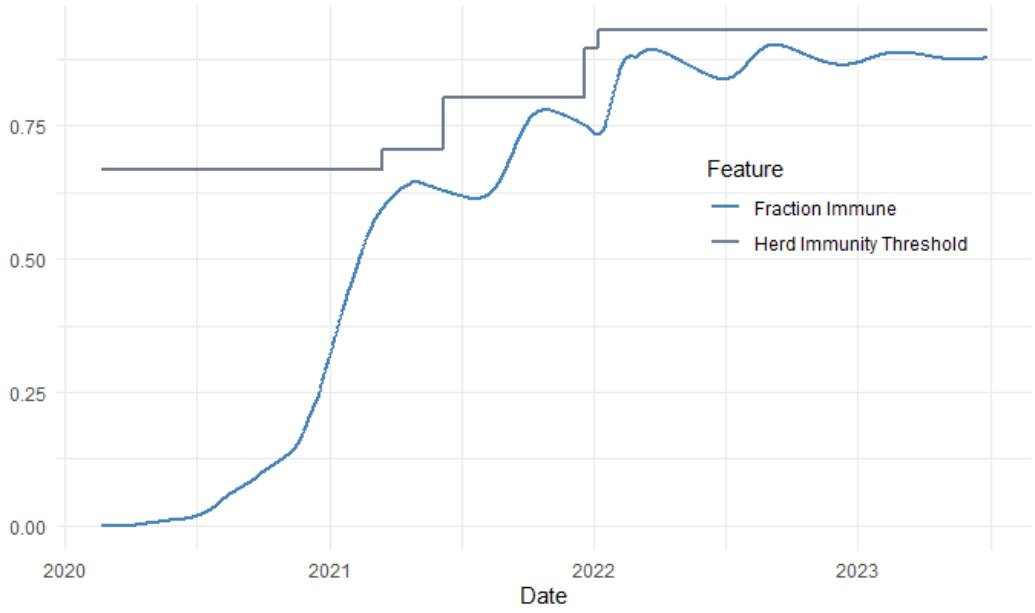

Figure 15: Predicted fraction of immune vs. herd immunity threshold in Oklahoma.

In Arequipa, a very different landscape can be seen (Figure 16). Rt decreases slightly during the initial strain, but only gets as low as 2.5 before lambda becomes dominant and Rt rises back above three. There are some periods where slight decreases can be observed, indicating that preventative measures are working somewhat. But these gains are vastly overshadowed by the introduction of new variants. It seems what whatever measures were in place were not effective enough to make any real difference, and after omicron was introduced Rt never dropped below five.

Examining herd immunity seems to tell the same story (Figure 17). While the population did make some strides in approaching the threshold, especially in the beginning of 2022, it was never enough to make a lasting impact. The fraction of immune individuals drops off significantly a few months into 2022 – likely owing to waning immunity – and much of the progress that had been made was lost. It is important to keep in mind that these values were derived from an imperfect model, and the reality is likely not as dire as the current plots might indicate. In addition, now that Rt and herd immunity can be examined through time it is much easier to pinpoint specific dates and compare what worked in one region with what did not work in another. Hopefully through this process, long lasting progress can be made in Arequipa and Peru as a whole.

Figure 16: Approximation of the Arequipa effective reproductive number through time.

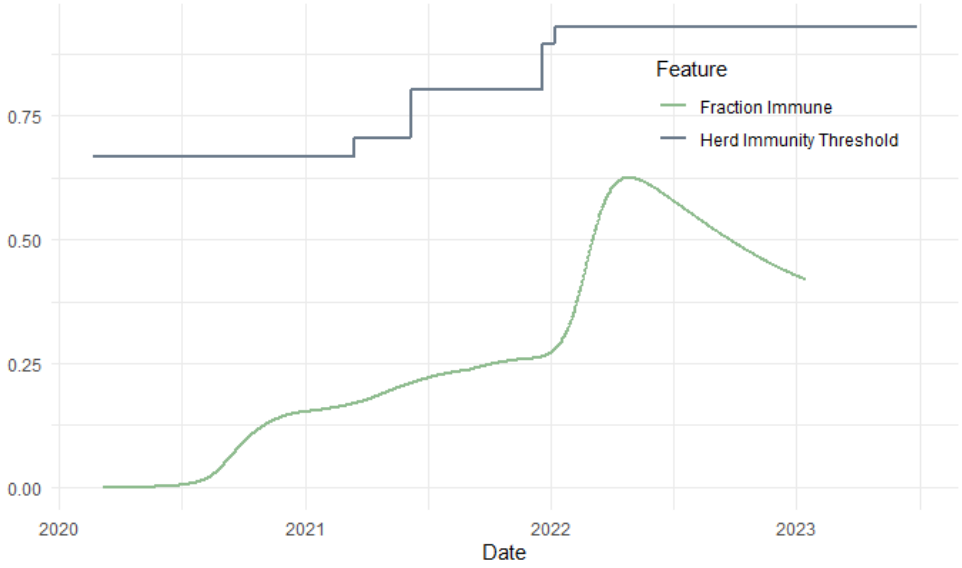

Figure 17: Predicted fraction of immune vs. herd immunity threshold in Arequipa.

6. Deliverables

The deliverables of this project included a GitHub repository of code used to ingest, explore, and clean the data. In addition, the files used to recreate the Oklahoma and Arequipa SEIR model are provided; while Python is the backend code used to generate the team's model-

building web application, once the models were created, they were exported and stored as JavaScript files (this is a constraint of the modeling software, and not a user choice). For clarity, the original Excel documents that these models are based on are also provided and contain visuals of model output. Another deliverable is a well-structured written document outlining the process used to reach the final model – as well as logic for all methods chosen – which is represented by this paper. The final deliverable is a journal quality paper to be submitted to a publication; this is currently in the work and will be completed by mid-May. This body of work helps to provide evidence for the team's hypothesis that cases in Peru are underreported, and gives Peru researchers a basis for future work once the research collaboration expires in December.

7. References

[1] "The highest COVID death rate in the world is in Peru. How did that happen? Goats and Soda: NPR." https://www.npr.org/sections/goatsandsoda/2021/11/27/1057387896/peru-has-the-worlds-highest-coviddeath-rate-heres-why (Accessed Feb. 18, 2023).

[2] covariants.org. (n.d.). COVID-19 Variants by Country. Retrieved April 27, 2023, from https://covariants.org/per-country

[3] W. O. Kermack and A. G. McKendrick, "A contribution to the mathematical theory of epidemics," Proc. R. Soc. Lond. Ser. Contain. Pap. Math. Phys. Character, vol. 115, no. 772, pp. 700–721, 1927.

[4] W. O. Kermack and A. G. McKendrick, "Contributions to the mathematical theory of epidemics. II.—The problem of endemicity," Proc. R. Soc. Lond. Ser. Contain. Pap. Math. Phys. Character, vol. 138, no. 834, pp. 55–83, 1932.

[5] W. O. Kermack and A. G. McKendrick, "Contributions to the mathematical theory of epidemics. III.—Further studies of the problem of endemicity," Proc. R. Soc. Lond. Ser. Contain. Pap. Math. Phys. Character, vol. 141, no. 843, pp. 94–122, 1933.

[6] "Epidemic Theory," Health Knowledge, [Online]. Available: https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1a-epidemiology/epidemic-theory. [Accessed: Apr. 27, 2023].

[7] "COVID-19 and Influenza Surveillance," Virginia Department of Health, [Online]. Available: https://www.vdh.virginia.gov/coronavirus/2022/01/07/covid-19-and-influenza-surveillance/. [Accessed: Apr. 27, 2023].

[8] A. M. Corson, M. Railey, and T. S. Miller, "High Throughput Severe Acute Respiratory Syndrome Coronavirus 2 Testing in a Pediatric Emergency Department during a Surge in COVID-19 Cases," Pediatr. Emerg. Care, Aug. 2021. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/34369565/. [Accessed: Apr. 27, 2023].

[9] "The Omicron variant has an average basic reproduction number of 3.28 (IQR: 2.03, 3.85)." [Online]. Available:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8992231/#:~:text=The%20Omicron%20variant%20has%20an%20average%20basic%20reproduction%20number%20of,IQR%3A%202.03%2C%203.85. [Accessed: Apr. 27, 2023].
[10] "COVID-19 Variants of Concern: Omicron." Yale Medicine. https://www.yalemedicine.org/news/covid-19-variants-of-concern-omicron (accessed Mar. 10, 2023).
[11] G. Shanmugam, S. Kumar, and M. Balakrishnan, “Zoonotic diseases and its prevention and control,” Zoonoses Public Health, vol. 68, no. 4, pp. 331-338, June 2021. doi: 10.15212/ZOONOSES-2021-0009.
[12] "Fact check: No evidence Omicron BA.5 is more infectious than measles or is the most infectious virus known," Reuters, Dec. 15, 2021. [Online]. Available: https://www.reuters.com/article/factcheck-omicron-reproduction-number/fact-check-no-evidence-omicron-ba-5-is-more-infectious-than-measles-or-is-the-most-infectious-virus-known-idUSL1N2YW1T0. [Accessed: Apr. 27, 2023].

8. Self-Assessment

My personal learning objectives included the application of the data science process, reading and applying current research, being able to tell a clear data-driven story, and being able to explain the limitations of the models used. I feel that I have accomplished these goals.

The most important skills by far pertained to R programming. Specifically, how to clean the data in a way that made logical sense, and how to visualize my results with the goal of communicating a message. Being aware of various assumptions and limitations also played a large role, especially when it came to modeling and interpreting results.

Prior to this project I had very little experience in exploring and applying research. In the beginning I wasn’t sure where to start, and often spent too much time on papers that I later found to be irrelevant to my work. Moreover, I was unfamiliar with the field of epidemiology and almost all its topics were foreign to me, including SEIR modeling. For this reason, I had to practice learning outside the classroom, and decide when I needed to ask for help. Over the semester, I feel that I have dramatically improved in both these areas and that I am better prepared for future data science work because of this.

This was completed as a 4-credit hour project and supervised by Prof. Charles Nicholson (cnicholson@ou.edu). This was a research project.