International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

A Visualization and Analysis of the Effect of Population Density on the Mutation Rate of SARS- CoV-2

Juno Kim*

Henry M. Gunn High School, 780 Arastradero Rd, Palo Alto, CA 94306, USA

Email: jnkim674@gmail.com

Abstract

The SARS-CoV-2 genome is prone to mutations during replication, similar to other viruses. Mutations are caused by random errors in the process of viruses replicating themselves via a host cell. With these mutations, SARS-CoV-2 changes into several different strains, often categorized by their geographical location, such as the UK variant(B.1.1.7) or the Brazilian variant(P.1). Population density is a metric detailing the number of citizens in km^2 and is used as an approximation for social interaction in this study. By comparing the two factors in a country, the relationship can be observed for SARS-CoV-2 viral mutation rate against population density. **(**The GISAID SARS-CoV-2 dataset was downloaded and analyzed for this study, along with population density data provided by the UN. After preprocessing the data, the number of distinct viral mutations in a country was analyzed by finding the unique mutations in the cases of a country. After calculating the viral mutation rate, a heatmap was assembled with several Python libraries which highlights areas of interest. Additionally, a scatter plot graph comparing the two factors was created using the Seaborn library. After taking into account the population densities of multiple countries, the results show that population density has no observable correlation with mutation rate from this dataset After analyzing the graph and the map, there is no clear correlation between population density and the mutation rate of SARS-CoV-2. However, the procedures used for this study can be applied to other factors as well, such as temperature, which may result in trends that forecast future areas of viral mutation.

Keywords: SARS-CoV-2; COVID-19; Viral mutation; Visualization; Population density.

---

* Corresponding author.

1. Introduction

The SARS-CoV-2 virus has caused the ongoing worldwide COVID-19 pandemic. With the spread of the virus through millions of patients, the virus has also mutated, resulting in various strains. Like other viruses, the SARS-CoV-2 virus mutated while replicating itself in a host cell, with errors in this process. These mutations change the nucleotide bases of the mRNA, the instructions the virus uses to replicate itself. Population density is utilized in this paper as an approximate measure for social interaction. The core rationale of a positive correlation between SARS-CoV-2 mutation rate and population density is: With higher population density, more people will interact and infect themselves with the virus. With more infectees, there will be more mutations.

This study examines the relationship between population density of geographical areas and the mutation rate of the SARS-CoV-2 virus. The GISAID SARS-CoV-2 dataset is utilized and manipulated into a more digestible form, allowing for the creation of a map and graph which display the desired relationship.

2. Related Works

2.1. Viral Mutation

Finding the mutations of viruses is extremely important in the process of developing a more complete vaccine. As viruses undergo natural selection and evolve, they could potentially build a tolerance to the antibodies that are produced by currently used vaccines [1]. These mutations can not only build a tolerance to the antibodies used to fight them with vaccines, but they can also mutate to attack a wider variety of hosts [2]. Additionally, SARS-CoV-2 is a RNA virus, meaning its mutation rates are quite significant[3].

2.2. Population Density v. Viruses

Population density can affect the time needed to develop natural immunity. With a higher population density, a quicker herd immunity can be achieved, as seen in a study performed in Thailand concerning urban and rural regions [4]. H1NI influenza viruses are a very common infection among human populations. A study examining the relationship of the H1N1 influenza virus subtype A found a positive and linear correlation with the mutation rate and population density through experimentation [5]. With the potential of having many changes on the virus, mutations of SARS-CoV-2 have been studied and there are multiple variants which could have a profound effect on the mutation rate of the virus [6].

2.3. Population Density v. Covid-19

A recent study examining the impact of population density on COVID-19 infection and mortality rates in India found positive correlation between the spread of the virus and population density [7]. As this relationship shows, there is an increase in infection rate with a higher population density, which points to a probable positive correlation between population density and mutation rate, as mutations take place during infection. A study of Chinese lockdown policies and population density as a factor in the spread of COVID-19 shows that while population density may be a factor in the spread, when quarantine is implemented correctly, it becomes less

important to the infection rate [8]. This potentially shows that social interaction along with population density may be a clearer indicator of virus spread. Finally, a similar study on Bangladesh that examines the impact population density has on SARS-CoV-2 mutation frequency with a positive correlation as a result [9].

2.4. Visualizations

The mission of this study is both to confirm the relationship suggested by previous studies between population density and covid-19 mutation rate, and to create simpler visualizations that clearly illustrate this. This will be done by performing analyses on covid-19 mutation data alongside population density statistics, then producing multiple visuals via Python that clearly map the two variables – population density and mutation rate – against each other. This slightly simplified and more isolated approach addresses existing limitations by creating easily understandable comparisons and solely examining one possible cause as opposed to a multitude of combinations.

3. Background

The COVID-19 pandemic started in December of 2019 with the first case identified in Wuhan, China. With 176 million cases, and a quarantine approaching a year and a half, the virus has impacted the daily lives of everyone globally. Like other viruses, the SARS-CoV-2 virus has various strains, such as the UK strain, or the Indian strain. In order to combat viruses, mutations must be taken into account when developing vaccines. Mutations take place during the replication process of a virus, when a copying error of the genetic material results in a slight change. The SARS-CoV-2 virus, unlike other viruses, has the ability to check the RNA of its copies, a function previously unseen in others such as the influenza virus. This significantly decreases the rate of mutation compared to other viruses, but there is still the presence of multiple variants, such as the B.1.1.7 variant present in the United Kingdom or the B.1.351 present in South Africa.

Population density is the measurement of population per unit area, comparing the number of people present to the space they take up. This metric is often used over pure population, as it provides a more accurate representation to the number of interactions people will have with each other. In the case of viral mutations, a higher number of human interactions would also indicate a higher number of viral replication instances, leading to more chances of errors in the process, and hypothetically more mutations.

4. Data

The genome sequences of the SARS-CoV-2 cases were provided by the GISAID dataset in the form of a tsv file, which details 1755981 cases of the virus. The features provided are listed below, with an example:

- *Virus Name* - The name of the virus, including location and date: hCoV-19/Australia/NT12/2020
- *Type* - The type of virus: betacoronavirus
- *Accession ID* - An ID used for identifying the cases: EPI_ISL_426900
- *Collection Date* - The sample collection date: 2020
- *Location* - The continent, country, state/territory, and municipality/city of the case: Oceania / Australia /

Northern Territory

- *Sequence Length* - The length in base pairs of the sequence: 29862
- *Pango lineage* - The variant of the case standardized by PANGO lineages: B.1
- *AA Substitutions* - The mutations encountered in the sample compared to the reference genome: (NSP15_A283V,NSP12_P323L,Spike_D614G)
- Other features regarding coverage, patient data, and more

The data was preprocessed and parsed with regular expressions and relevant data. The Location feature of the data was processed and split into Continent and Country features, while any samples with a Collection Date before 2019 were filtered out. Additionally, any invalid data points were removed in this preprocess. This resulted in the following features, with the same sample shown:

- *Collection Date* - The sample collection date: 2020
- *Location* - The continent, country, state/territory, and municipality/city of the case: Oceania / Australia / Northern Territory
- *AA Substitutions* - The mutations encountered in the sample compared to the reference genome: (NSP15_A283V,NSP12_P323L,Spike_D614G)
- *Continent* - The continent of the sample: Oceania
- *Country* - The country of the sample: Australia

5. Methodology

To examine the mutation rate across several countries, the average number of mutations per case must be calculated. Firstly, the number of cases in each country is calculated from the dataset. For each country, there is an average number of ~9853 cases. If a mutation is found to be common in all the cases of a country, it is not distinct, and therefore has not developed in the country. As such, the mutations that are common across all cases are omitted, and the distinct mutations that were potentially caused by the population density of a country are inserted into the distinct mutations feature. By taking the length of distinct mutations, the mutation count of each country can be calculated and put into the respective feature. Finally, the mutation rate can be calculated by dividing the mutation count by the number of cases. This analysis results in the following features:

- *Country* - The country analyzed: Afghanistan
- *Number of cases* - The number of cases gathered from the country: 16
- *Distinct mutations* - The unique mutations from all samples in the country: [NSP3_T1465I, NSP2_K456N, NSP3_M829I, NSP4_A45…]
- *Mutation count* - The number of distinct mutations: 27
- *Mutation Rate* - The number of mutations/samples in the country rounded to the hundredth: 1.69

In order to clearly evaluate the analyzed data, a visualization helps in clearly grasping trends and patterns. For this reason, the map below was created, displaying the country name and mutation rate over a geographical map. Additionally, there are variants of the map which show this on a city-scale and a continental-scale when the data

is analyzed on these scales respectively. However, for this study, the country-scale map will be analyzed.

To create the map, the data was first processed into a Pandas dataframe structure. The pycountry library was used to format the data into countries. The geopandas library provided a low resolution image of the world used for the map. Finally, the matplotlib.pyplot and mapclassify.classifiers modules were utilized to create the figure by overlaying the information over the image of a map.

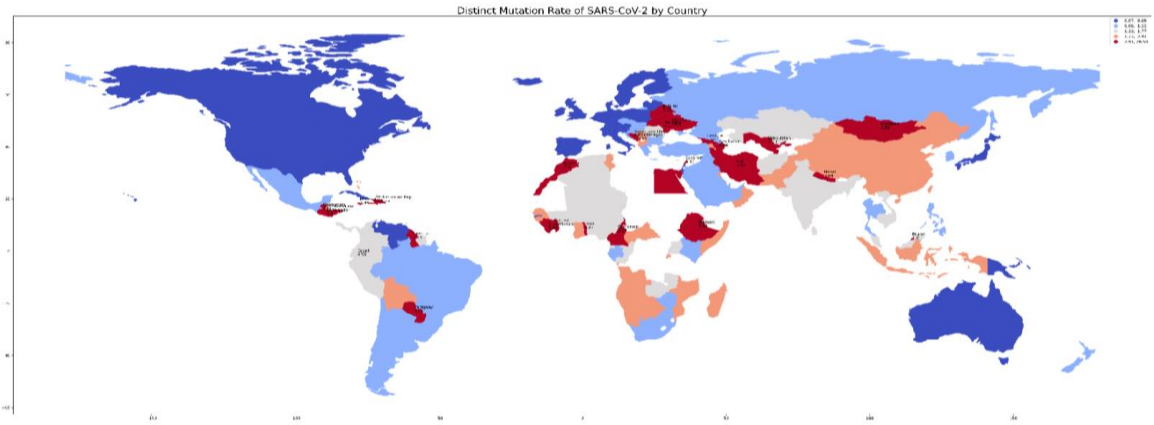

Figure 1: Distinct Mutation Rate of SARS-CoV-2 by Country

While the visualization helps for discussion of maps and the spread of mutation rates geographically, it does not account for the population density. Population density data was taken and preprocessed from the United Nations dataset, and the most recent count was used (2019). For this purpose, a scatterplot was used from the seaborn library, showing the correlation between mutation rate and population density, with each data point representing a country.

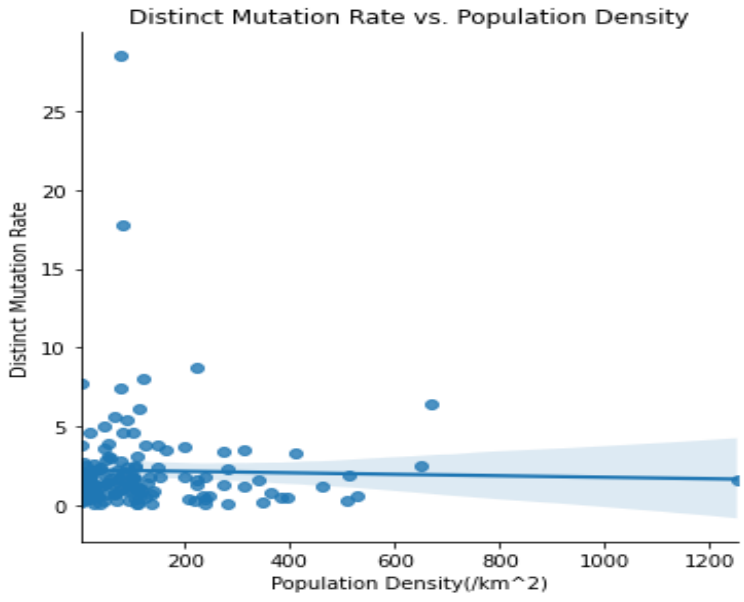

Figure 2: Distinct Mutation Rate vs. Population Density

These visualizations were also applied to the mutation count, as opposed to the mutation rate.

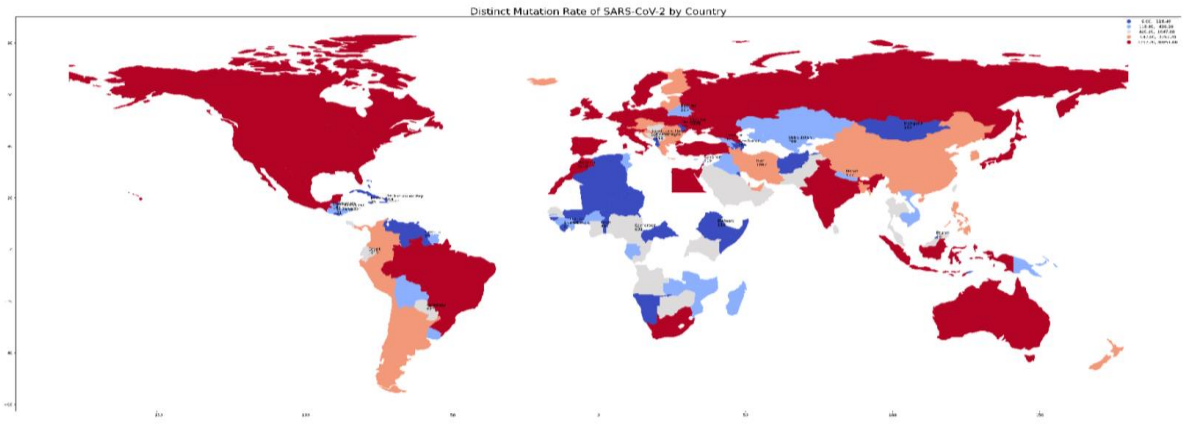

Figure 3: Distinct Mutation Count of SARS-CoV-2 by Country

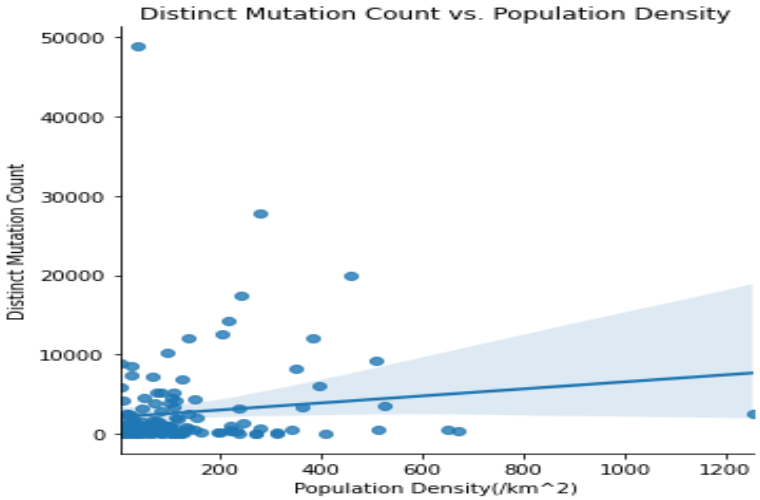

Figure 4: Distinct Mutation Count vs. Population Density

6. Results

Table 1: Table of Countries with highest mutation rate

Country	Number of Cases	Mutation Count	Mutation Rate	Density
Ukraine	181	5158	28.50	75.9401
Morocco	293	5215	17.80	81.7203
Dominican Republic	48	422	8.79	222.2466
Azerbaijan	13	105	8.08	121.5577
Guyana	11	85	7.73	3.9765

As shown in Figure 2, there is no apparent correlation, and shows that mutation rate is neither directly nor indirectly related with population density. Table 1 makes clear that the countries with the highest mutation rate are all over the spectrum on population density, ranging from 3.98/km^2 to 222.25/km^2. This points to another factor, such as the number of social interactions, or temperature being more influential on the mutation rate. However, the visualizations still help narrow the broad scope of causes that could cause higher mutation rates, even with little correlation between the population density and mutation rate. Additionally, it allows for a graphic approach to seeing the spread of the virus and its mutations across countries.

7. Discussion

There are several outliers present in Figure 2 from majority of data points which have a mutation rate of 0-10 mutations per case. The highest outlier for mutation rate is Ukraine, with a mutation rate of 28.50 mutations per case and a population density of 75.9401. One factor that may have made Ukraine an outlier is the classification of Crimea as part of Ukraine. Due to the pycountry library being based on the ISO database, Crimea is considered a region of Ukraine. When Crimea and Ukraine are separated, the data shows that Ukraine has a mutation rate of 29.82 mutations per case and Crimea has a typical mutation rate of 4.67 mutations per case. In the case of Ukraine, the possibility of it being at such an extreme due to having more samples is ruled out by the use of mutation rate as opposed to mutation count, along with the fact that other data points such as United States or Germany have significantly more data points.

The next outlier of mutation rate is Morocco, with a mutation rate of 17.80 with 5158 mutations and 181 cases. With a population density of 75.9401, it is similar to other data points. It is interesting to note that both outliers for mutation rate have 5000 mutations with population densities 80. This can be tested further by gathering more data samples of non-country geographic regions with similar population densities.

There are a handful of limitations to this analysis, as the data used was static, and was not continuously updated over the course of the study. The most significant limitation to this study is likely unreported cases as COVID tests are not mandated for everyone, especially over regular intervals of time. Additionally, there are unavoidable constraints such as inaccurate results of COVID tests due to a lack of a clear international standard, but rather a different situation depending on the country. Finally, the status of the pandemic must be taken into account, as the virus is still mutating and has not fully run its course of becoming a seasonal virus or simply dying out.

8. Conclusion

In this study, the relationship between the population density of cities and the number of mutations of the SARS-CoV-2 virus was examined. The GISAID dataset was downloaded and preprocessed for relevant information. After calculating the number of distinct mutations in a country, a map and a graph were generated. The map shows countries with high mutation rates in warm tones and shows the geographical spread of the virus for centers of mutation. The scatter plot graph indicates that there is no solid correlation or relationship between population density and mutation rate, as the data points had a similar mutation rate regardless of population

density, excluding outliers. Conclusively, there is no observed correlation and improbable causation between population density and the number of mutations. In the future, other features of countries such as rates of social interaction, or temperature may be analyzed and observed trends from these analyses can help in the fight against viruses.

Acknowledgements

Special thanks to mentor Ryan Koo, Ho Joon Lee, and Charlie Ho for technical and conceptual feedback and help.

References

[1] S. C. Manrubia and E. Lázaro. (2006, Jun.). "Viral evolution." *Physics of Life Reviews*. [On-line]. 3(2), pp.65-93. Available: https://www.sciencedirect.com/science/article/pii/S1571064505000436?casa_token=ihWD1TXxrVgAAAAA%3A2GFfs_UDMzrir2Bdm5wkJ4KzuBWboR430CHxuU_4Dp-EPLmAM1-AnVF-YD9-t4xQdgue5QX86LU. [Jun. 17, 2021].

[2] S. E. Luria. (1945, Jan.). "Mutations of Bacterial Viruses Affecting Their Host Range." *Genetics*. [On-line]. 30(1), pp.84-99. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1209277/.

[3] J. W. Drake and J. J. Holland. (1999, Nov.). "Mutation rates among RNA viruses." *PNAS*. [On-line]. 96(24), pp.13910-13913. Available: https://www.pnas.org/content/96/24/13910.short. [Jun. 17, 2021].

[4] S. Lolekha, W. Tanthiphabha, P. Sornchai, P. Kosuwan, S. Sutra, B. Warachit et al. (2001, Mar.). "Effect of climatic factors and population density on varicella zoster virus epidemiology within a tropical country." *American Journal of Tropical Medicine and Hygiene*. [On-line]. 64(3), pp.131-136. Available: https://www.ajtmh.org/view/journals/tpmd/64/3/article-p131.xml. [Jun. 17, 2021].

[5] D. Jiang, Q. Wang, Z. Bai, H. Qi, J. Ma, W. Liu et al. (2020, Apr.). "Could Environment Affect the Mutation of H1N1 Influenza Virus?." *Int. J. Environ. Res. Public Health*. [On-line]. 17(9). pp. 3092. Available: https://www.mdpi.com/1660-4601/17/9/3092. [Jun. 17, 2021].

[6] T. P. Peacock, R. Penrice-Randal, J. A. Hiscox, and W. S. Barclay. (2021, Apr.). "SARS-CoV-2 one year on: evidence for ongoing viral adaptation." *Journal of General Virology*. [Online]. 102(4). Available: https://www.microbiologyresearch.org/content/journal/jgv/10.1099/jgv.0.001584. [Jun. 17, 2021].

[7] A. Bhadra, A. Mukherjee, and K. Sarkar. (2020, Oct.). "Impact of population density on Covid-19 infected and mortality rate in India." *Modeling Earth Systems and Environment*. [Online]. 7. pp.623-629. Available: https://link.springer.com/article/10.1007/s40808-020-00984-7. [Jun. 17, 2021].

[8] Z. Sun, H. Zhang, Y. Yang, H. Wan, and Y. Wang. (2020, Dec.). "Impacts of geographic factors and population density on the COVID-19 spreading under the lockdown policies of China." *Science of The Total Environment*. [Online]. 746. Pp.141347. Available: https://www.sciencedirect.com/science/article/abs/pii/S0048969720348762?casa_token=S3r5cjWCTLYAAAAA%3A-PCyVdA20pc8W-nSHRTCSjXoW9nWc1zhSCgXFR7wsmaXxDXu2-n9MP4T5jNCUcvCTuEEGLCcR7k. [Jun. 17, 2021].

[9] N. Sharif and S. K. Dey. (2021, Jan.). "Impact of population density and weather on COVID-19 pandemic and SARS-CoV-2 mutation frequency in Bangladesh." *Epidemiology & Infection*. [Online]. 149. pp. e16. Available: https://www.cambridge.org/core/journals/epidemiology-and-infection/article/impact-of-population-density-and-weather-on-covid19-pandemic-and-sarscov2-mutation-frequency-in-bangladesh/48A13CDB7ADE4F39C0E8E4BA6E4A2309. [Jun. 17, 2021].