Geographic Spread of Seasonal Flu

Clarke Miller MADA Project

Clarke Miller

2024-05-03

# 1. Abstract:

The influenza virus causes respiratory illness in a wide variety of hosts. Further, flu viruses tend to prone to mutation and recombination. Viruses in general have many avenues for mutation. One of the primary mechanisms for this high mutation rate include facots such as an error prone low-fidelity viral RNA dependent RNA polymerase lacking exonulease proofreading capacity, high turnover rate, and host cell specific factors. The influenza virus is also prone to recombination if two, or more, strains infect the same host simultaneously. Given the ability of many flu viruses to cross over from one species to another, the chance of such an occurrence is not zero. [***Domingo (2020)***] As such, seasonal outbreaks of influenza, and especially in swine, poultry, and humans, have the potential to cause outbreaks of flu viruses which could pose a serious threat to public health. [***Taubenberger & Kash (2010)***] Because of these factors, it is advantageous to track the incidence of flu outbreaks across the globe. This data is then used to look for serious outbreaks and species crossover events. The collected flu incidence data is also used to track the spread of seasonal flu as it spreads accross the globe each year. This data helps public health officials to predict patterns of flu and to guess whihc strains should included in vaccine preparation.

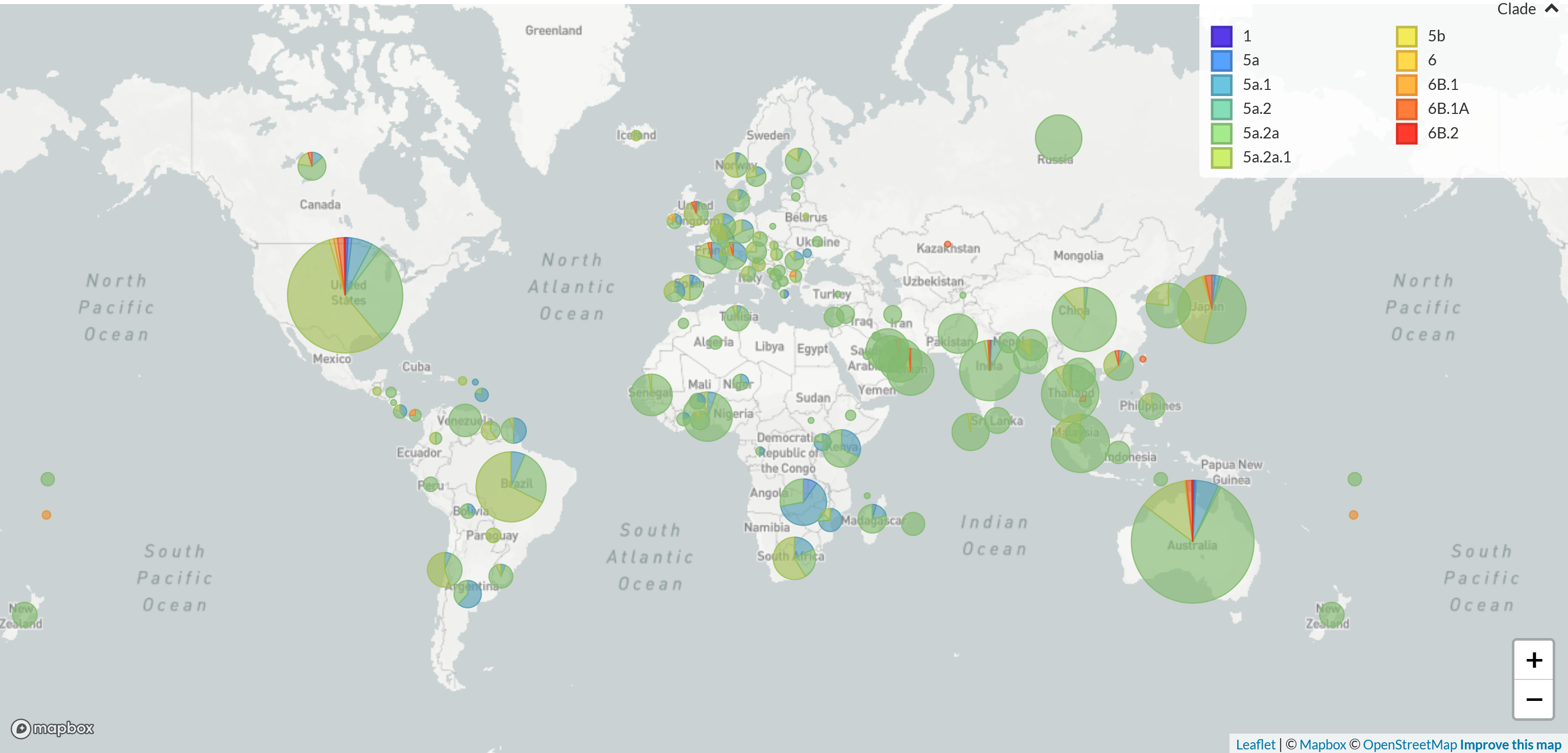
# 2. Introduction:

## 2.1 General Background Information:

Influenza is an acute respiratory virus caused by one of several strains of virus. Seasonal influenza is typically caused by circulating “A” and “B” strains of the disease. Infection is accompanied by typical respiratory symptoms. These may include cough, sore throat, and nasal discharge. Symptoms can also include abrupt onset of high-grade fever, muscle ache, loss of appetite, headache, and malaise. After a typical course, influenza can go on to affect other organs such as the lungs, brain, and heart. Severe infections may require hospitalization. Influenza infection can be much more severe for children, elderly adults, patients with chronic illnesses, and immunocompromised hosts. According to data provided by the World Health organization, seasonal respiratory illnesses are linked to over 650,000 deaths each year worldwide. In the United States the flu is responsible for approximately 42,000 deaths per year and ranks 13th most frequent cause of death. As such, influenza is the subject of extensive public health surveillance and mediation efforts.[***Krammer (2018)***; ***Moghadami (2017)***]

## 2.2 Description of Data and Data Source:

Detailed data on covid and flu strains, outbreaks, and general epidemiological information is available via several sources. Data on these infectious diseases is available from sites such as GISAID (<https://gisaid.org/>) and the Global Health Data Exchange (<https://ghdx.healthdata.org/>). These databases are highly detailed and can provide data that includes information about lineage and geographic data. (See **Figure 1** Below.) Yearly surveillance data is also available from the World Health Organization Global Influenza Programme.



***Figure 1:*** *World Flu Burden by Clade.* Figure generated at GISAID.org.

## 2.3 Questions/Hypotheses to be Addressed:

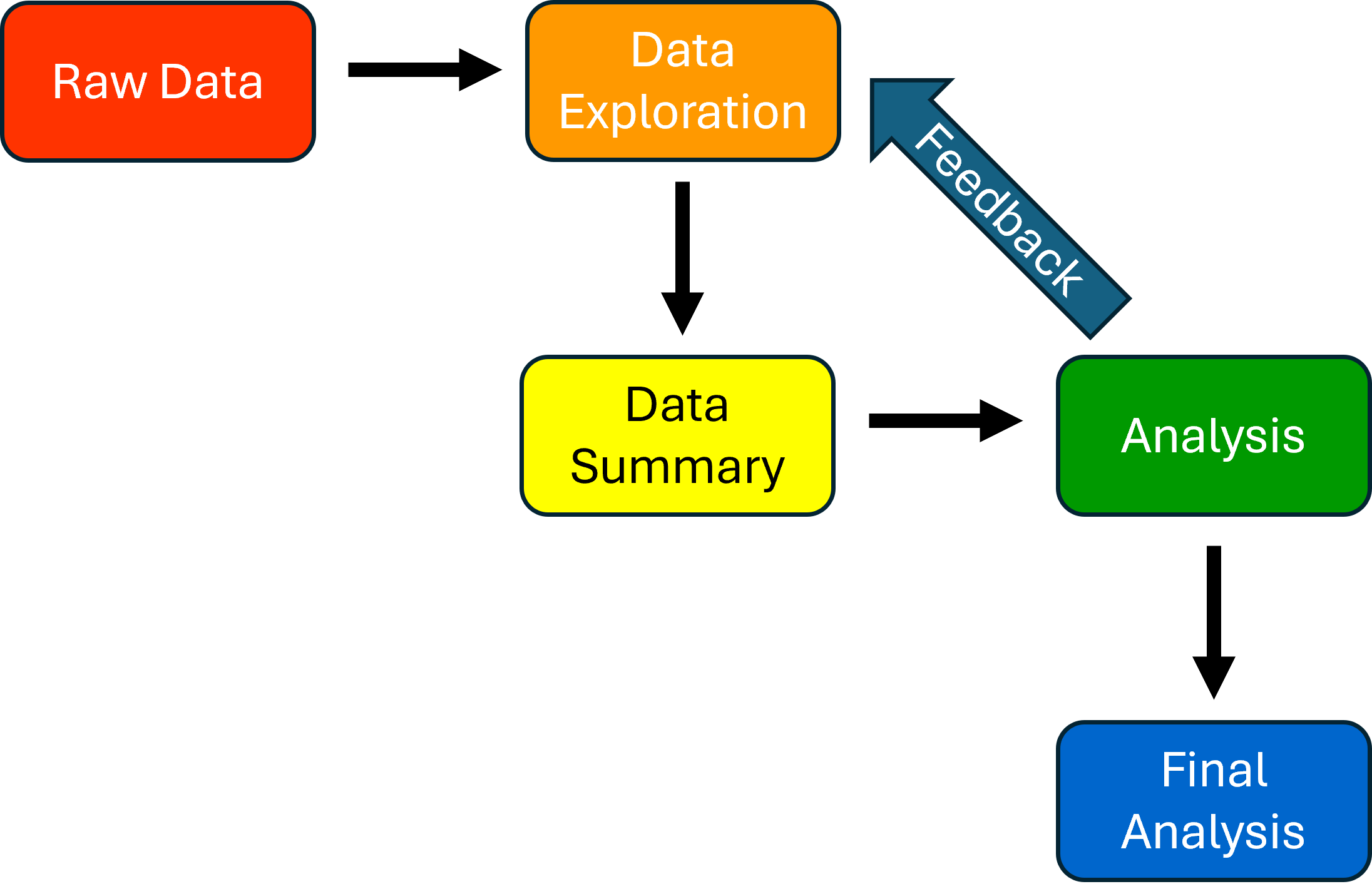
The research questions that I would like to answer are:

1. Can the progression of flu outbreaks be traced geographically or temporally over the course of a year?
2. Is there a specific location from which the annual predominant strain of flu originates?
3. If a pattern can be established, does it repeat on a yearly basis? Does the pattern shift from year to year? Can other factors, such as weather, account for any changes to this pattern?

I am certain that efforts like this already exist. The best of these is software produced by Nextstrain (<https://nextstrain.org/>) which is capable of resolving complex time/geographic coupled with genetic data for reported viral infections. Nextstrain offers real-time tracking of pathogen evolution for viruses such as SARS-CoV-2, seasonal flu, avian flu, mpox, RSV, and dengue. In fact, the graphic used in Figure 1 (above) was made by GISAID using Nextstrain software. My purpose for this project is to use this project as a learning tool and to see if I can replicate, in a small way, those prior efforts. Data to support this project is readily available from multiple sources such as GISAID (<https://gisaid.org/>), the World Health Organization Global Influenza Programme (<https://www.who.int/tools/flunet>).

# 3. Methods:

## 3.1 Schematic of Workflow:



***Schematic 1:*** ***Schematic of Workflow For Data Exploration and Processing.*** *The workflow associated with this project. Raw data is explored graphically and via numerical analysis methods, initial data for analysis is selected and summarized, models are built to examine how well the data matches the proposed hypothesis, data may be redefined in an iterative process, finally a final data analysis is performed and presented.*

## 3.2 Data Aquisition:

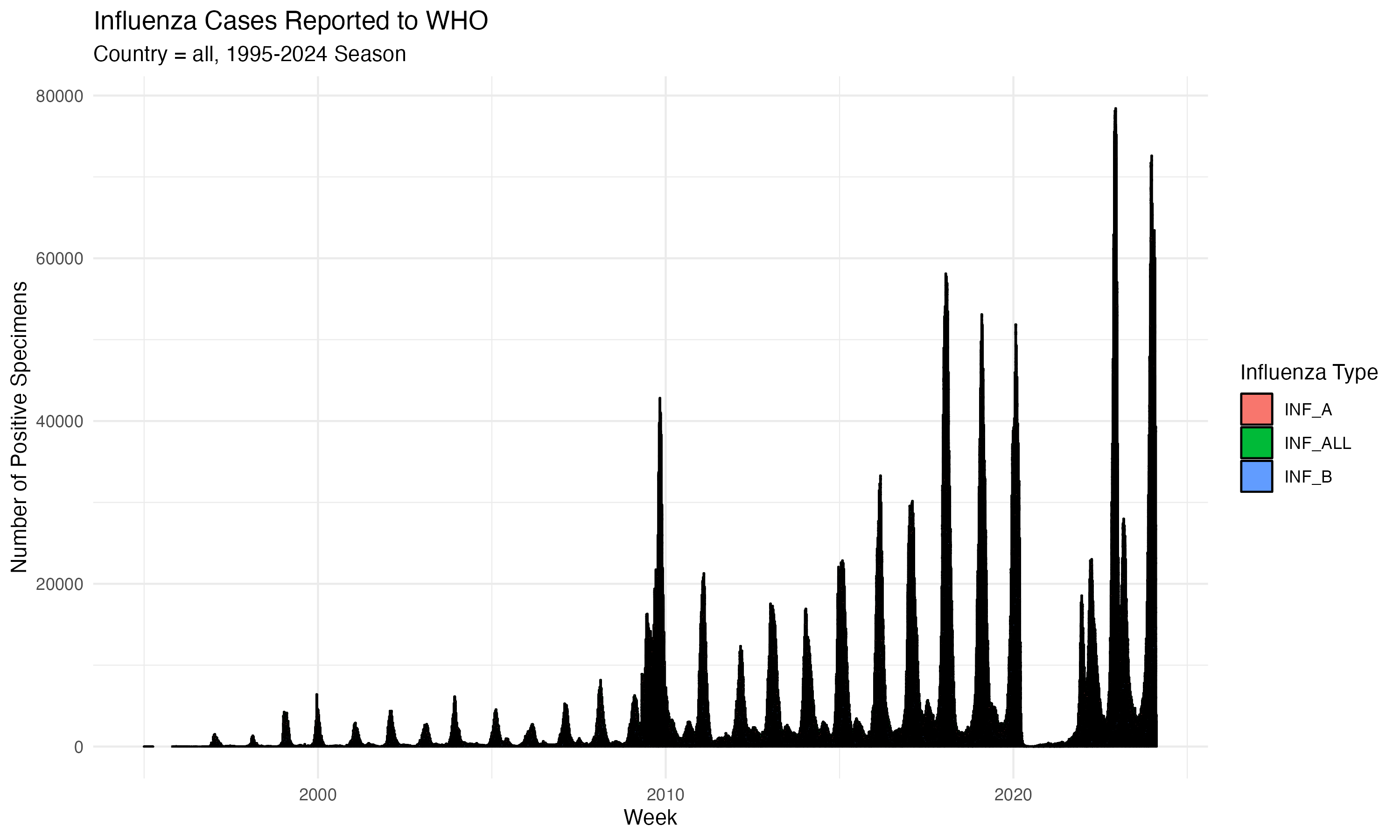
Data was acquired from the World Health Organization Global Influenza Programme. The data includes the excel files containing raw data (VIW\_FNT\_CM.xlsx) and a file containing the legend for the data (VIW\_FLU\_METADATA.csv). This data set is one example of the data available for a project of this sort. This specific data set includes information on 89 different parameters associated with influenza outbreaks as reported world-wide since about 1995. The data includes time, region, type of flu, and many other variables that are important for answering the questions posed above.

## 3.3 Data import and cleaning:

Data was imported directly from the repository as an xlsx file. (VIW\_FNT\_CM.xlsx) The file was cleaned to remove missing data. A .rds version of the data was saved (VIW\_FNT\_CM.rds). Data that was filtered or summarized for graphs, tables, and for eventual statistical analysis were saved as individual .rds or csv files. All data analysis and statistical work was performed and documented in a separate .qmd file (workspace.qmd).

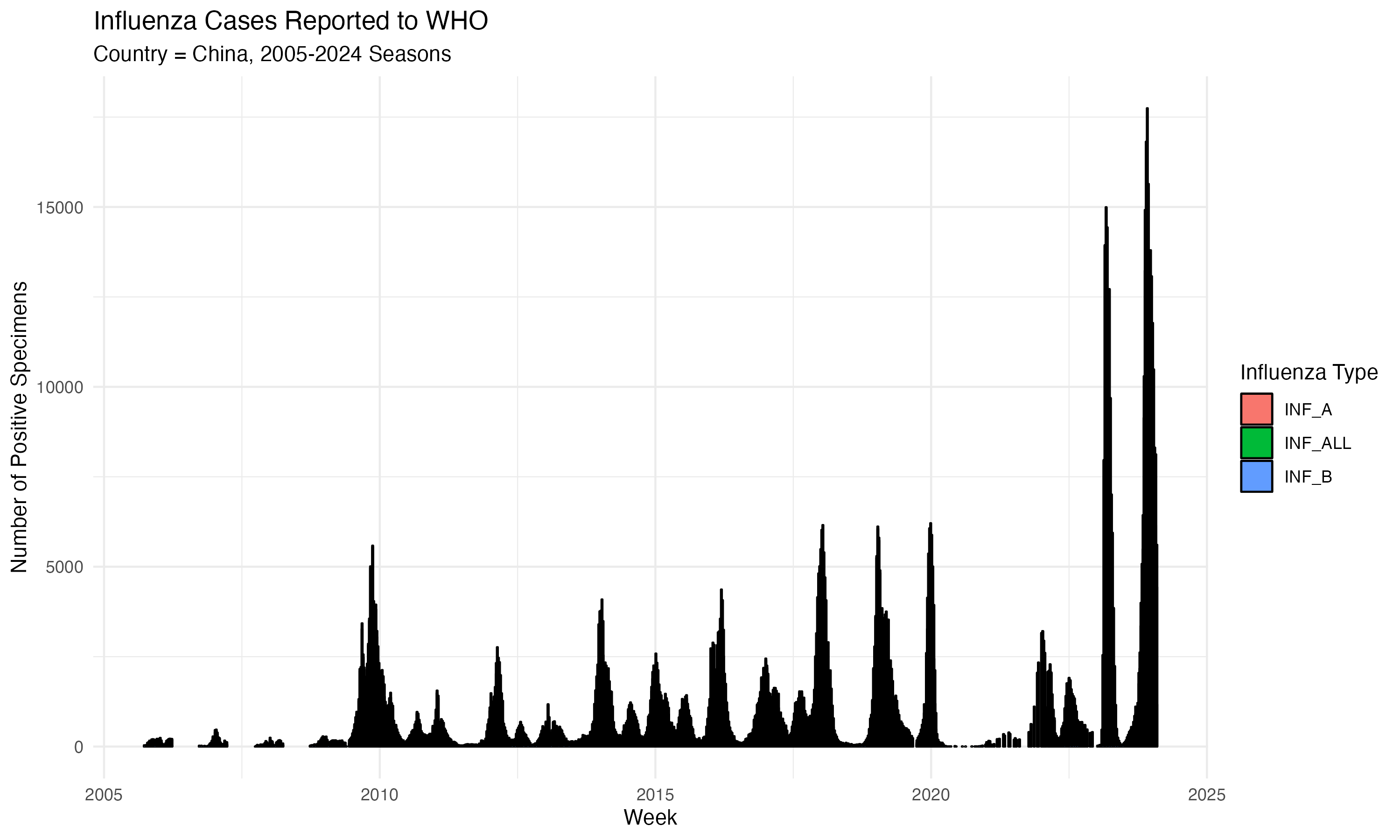
## 3.4 Exploratory/Descriptive analysis:

In order to get a feel for the scale of data in the VIW\_FNT\_CM.xlsx one often needs to summarize or tabulate the data. In this case, the data set is very large and summary tables exceedingly cumbersome. Given the size of the data set, it is best to begin with a graph of the incidence of influenza type A, type B, and the total number of reported cases for the duration of reported. (See **Figure 2.**)



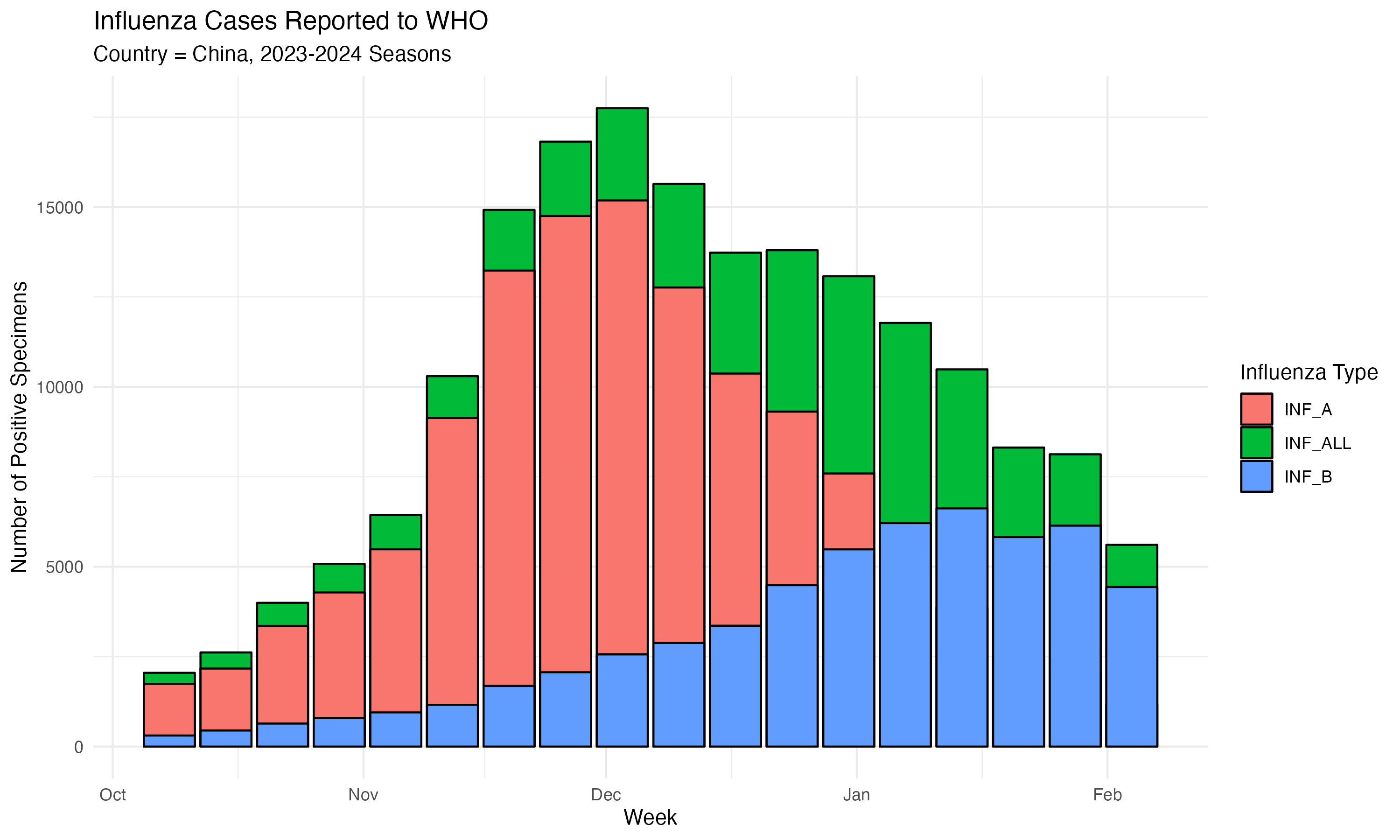
***Figure 2:*** *World Flu Burden by Year.*

The graph reveals the extent of the data available. The graph also reveals interesting features of the data set that may not have been readily apparent from if viewing the data in a tabular format. Since 1995 the number of seasonal flu cases has followed a pattern that peaked during the winter months in the Northern hemisphere. Notable features of this graph include the large jump in reported cases in the year 2010 and the large gap in reported cases in the year 2021. The jump in cases on or about 2010 is due to increased participation in surveillance efforts and improved diagnostic techniques. Hand in hand with these factors was an increasing awareness of the issue and consequent increased rate of reporting. [***Perdue & Nguyen (2012)***] The gap in the 2021 flu season was likely due to the measures taken to curb the SARS-CoV-2 pandemic through the 2020-2021 flu season. Basic measures such as maintaining proper social distance, hand washing, public wearing of masks to prevent aerosol droplet spread, and other efforts were extremely effective at preventing the spread of seasonal flu. [***Rubin (2021)***] Never the less, the amount of data reported worldwide since 1995 is massive. In fact, the data is so dense that none of the categories of flu selected are visible as separate itmes in the bar graph. Further filtering is obviously required. For example, we might want to look at the total flu cases compared to the number of total cases of the A strains and the total number of the B strains in countries such as China, Australia, Canada, and France. It is also a good idea to clean up the data by omitting any missing values. **Figure 3** depicts the data set filtered for information from China.



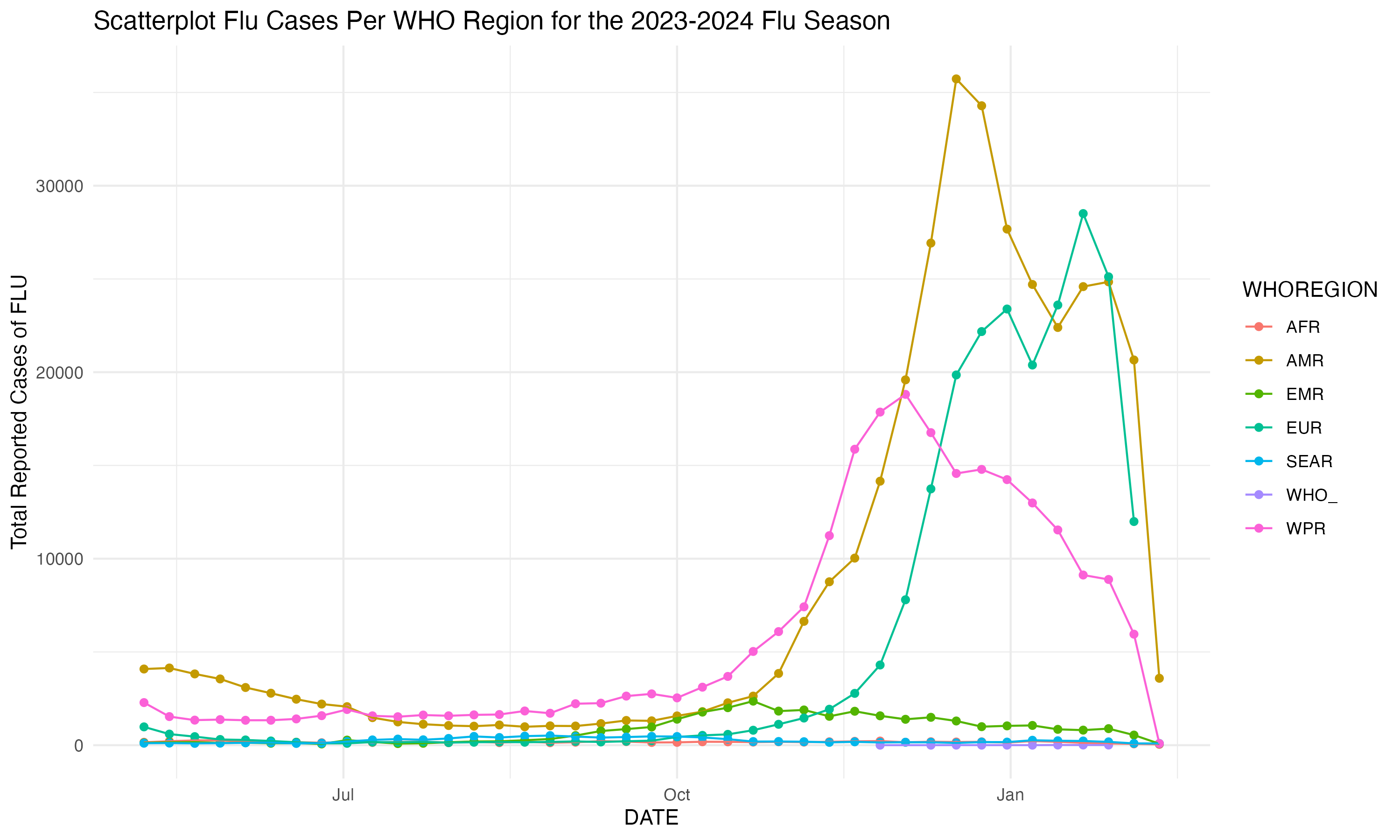
***Figure 3:*** *World Flu Burden by Year for China Since Reporting Began.*

Restricting the data set to a single country doesn’t seen to resolve the data in a useful manner. Overall the data still shows the same basic structure and features as the entire data set. It may be helpful to focus in on a single year. **Figure 4** depicts the influenza surveillance data for the country of China for a single flu season.



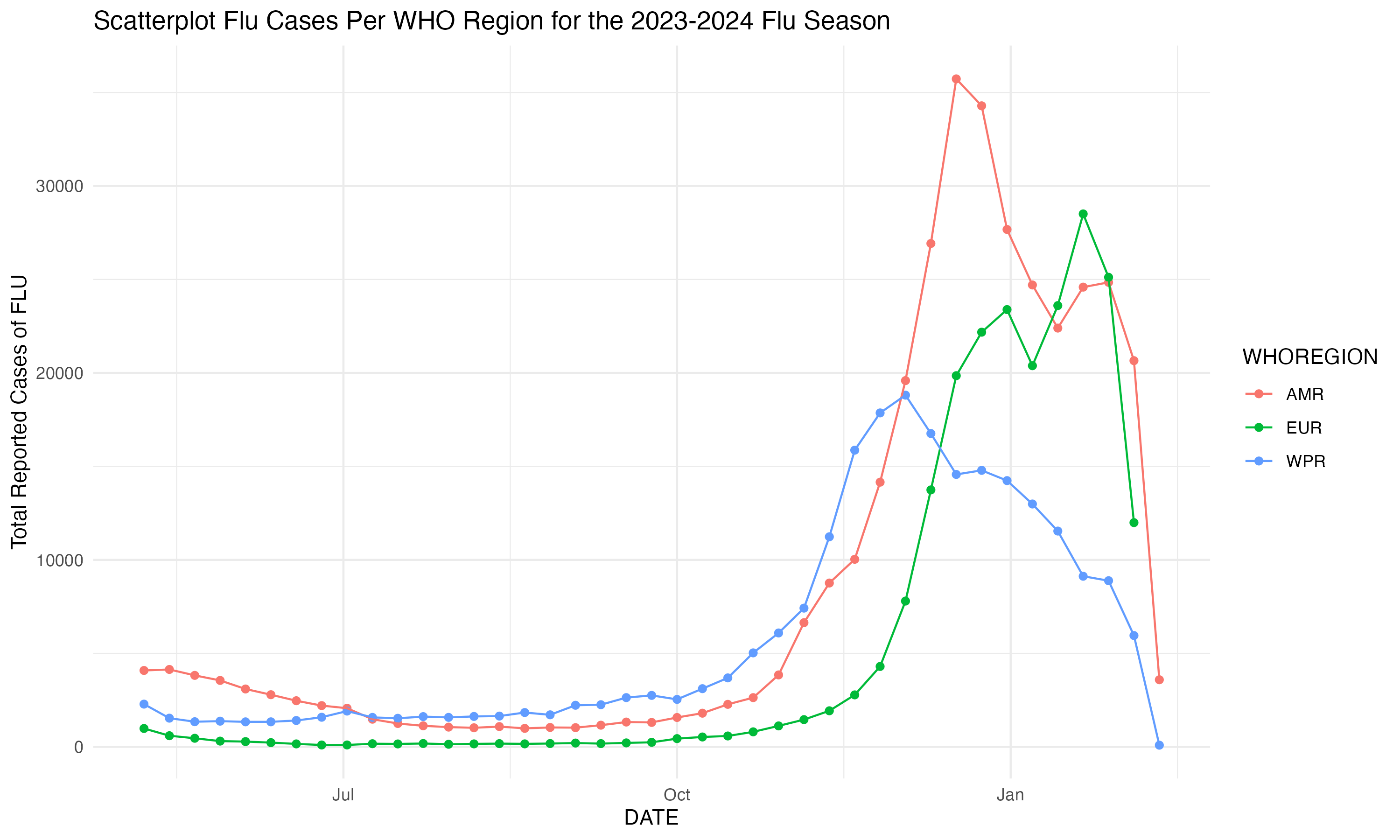
***Figure 4:*** *World Flu Burden by Year: China, 2023-2024 Flu Season.*

These graphs help whittle down the data a bit. For example, we can look at the stacked graph in **Figure 4** to see that over the course of the year the major flu variant shifted from influenza A to influenza B. That being said, this format is not helpful for looking at the incidence over time per country or per region and comparison of those regions. To look at the data in that manner and, perhaps, look for time and geographical patterns a simple scatter plot may be helpful. For example, a graph of total reported cases of flu for each WHO region over a single flu season may be revealing. (See **Figure 5**)

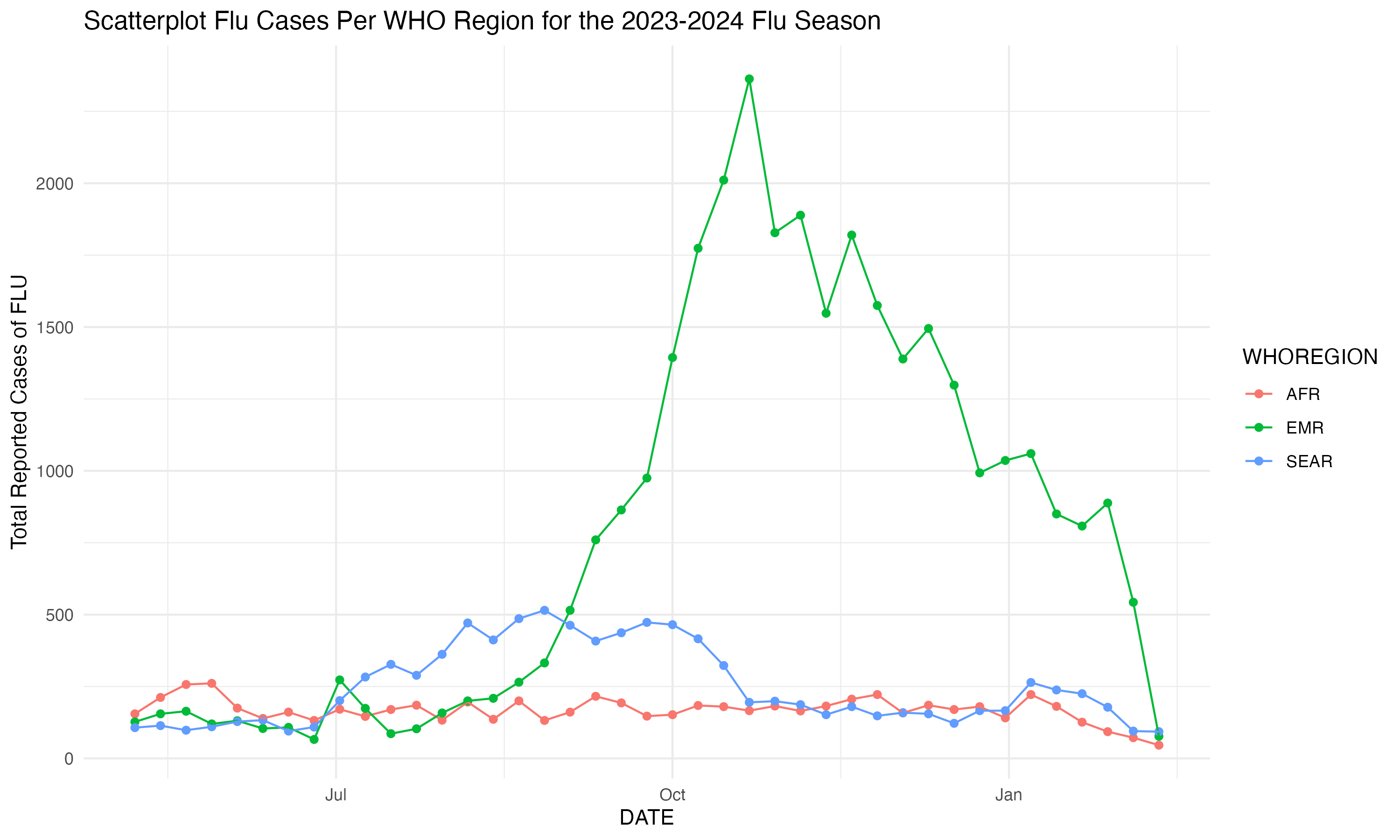


***Figure 5:*** *World Flu Burden by Region Over the 2023-2024 Flu Season.*

The resulting graph shows that the maxima for the three largest WHO flu regions, WPR, AMR, and EUR, do follow each other by approximately 3-4 weeks. The region encompassing Europe shows two distinct maxima. The first maxima is preceded by the maxima in the People’s Republic of China WHO region by 3-4 weeks and corresponds to the largest maxima in the AMR data. The second is preceded by the maxima in the data from the WHO Americas Region by the same 3-4 week time frame. The data from the AMR region also show a second peak in the number of flu cases reported that follows the first EUR region peak by approximately 3-4 weeks. Perhaps this indicates spread from EUR back to the AMR region? As such there certainly seems to be a geographic/time component to the spread of the flu worldwide. (West to east in the northern hemisphere.) Further analysis of the available phylogeographic data may be able to pinpoint the origin of each specific strain. In **Figure 5** it is also apparent that the relative magnitudes of the values of the WPR, AMR, and EUR regions overwhelm the data from the other WHO regions. In this case it may be prudent to separate the graphs into separate graphs for WPR, AMR, EUR and AFR, EMR, SEAR. (See **Figure 6** and **Figure 7.**)



***Figure 6:*** *World Flu Burden by Region Over the 2023-2024 Flu Season.* WHO Regions: WPR, AMR, and EUR



***Figure 7:*** *World Flu Burden by Region Over the 2023-2024 Flu Season.* WHO Regions: AFR, EMR, SEAR

By splitting the graphs a great deal more detail can be discerned. In addition to the putative west to east pattern that appeared for the WPR, AMR, and EUR WHO regions, there is a similar pattern for the SEAR and EMR regions. An easy interpretation of this data would point to the seasonal flu beginning in the southeast Asia WHO region progressing to the eastern Mediterranean WHO region, then progressing to the WPR, AMR, and the EUR WHO regions respectively. Of course, it is impossible to draw any conclusions from this data set and a single season. To make a solid determination the data set would need to track not just time and geographic information, but connect the actual flu strains as they move around the globe. This can be done with genetic data, i.e. DNA or RNA sequence data, from the flu cases reported. In fact this is exactly the approach that the Nextstrain software takes. In addition to presenting strong evidence of the central question of this research. Further, graphing the data as opposed to attempting to generate a summary table is a far more effective way to explore the data set. As can be seen from **Table 1** (below), a tabular format doesn’t highlight the patterns in the data in the same explicit manner as the graphic exploration.

| WHOREGION | 2023-08 | 2023-09 | 2023-10 | 2023-11 | 2023-12 | 2024-01 | 2024-02 |
| --- | --- | --- | --- | --- | --- | --- | --- |
| AFR | 664 | 717 | 864 | 775 | 834 | 622 | 118 |
| AMR | 4128 | 4817 | 12122 | 39591 | 144201 | 96535 | 24244 |
| EMR | 1006 | 3114 | 9370 | 6832 | 6211 | 3606 | 620 |
| EUR | 666 | 829 | 3470 | 10463 | 86959 | 97608 | 11993 |
| SEAR | 1884 | 1781 | 1598 | 667 | 768 | 905 | 188 |
| WPR | 6819 | 9862 | 20454 | 52385 | 79170 | 42541 | 6042 |

***Table 1:*** *World Flu Burden by Region Over the 2023-2024 Flu Season.* The data is restricted to September 2023 through February 2024.

## 3.5 Statistical Analysis:

# 4. Basic statistical analysis:

A cursory look at the data presented in **Figure 6** (above) shows a putative dependence of the AMR region flu season on the flu season peak approximately 3 week prior in the WPR region. Likewise the peaks in reported flu cases in the EUR region seem to correspond to the WPR maxima approximately 3 weeks prior followed by a second maxima at a 3 week delay from the AMR region maxima. An appropriate initial analysis would be to simply perform a linear regression of the data with the AMR and the EUR WHO regions being the dependent variables based on all other variables. (See **Table 2** and **Table 3** below.)

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | 691.5842347 | 2270.8976309 | 0.3045422 | 0.7625183 |
| AFR | 8.0083193 | 12.0021127 | 0.6672425 | 0.5089929 |
| EMR | -1.0685590 | 0.8987467 | -1.1889435 | 0.2424661 |
| EUR | 0.7563276 | 0.0768632 | 9.8399220 | 0.0000000 |
| SEAR | -5.4989117 | 3.6230269 | -1.5177673 | 0.1380548 |
| WPR | 0.8139670 | 0.1468065 | 5.5444873 | 0.0000031 |

***Table 2:*** *Simple Linear Regression of AMR Dependent and All Other Variables as Independent.*

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | 2697.627894 | 2535.9891482 | 1.0637379 | 0.2947317 |
| AFR | -25.971601 | 12.9631547 | -2.0034939 | 0.0529143 |
| AMR | 0.971133 | 0.0986932 | 9.8399220 | 0.0000000 |
| EMR | 0.099732 | 1.0386332 | 0.0960223 | 0.9240505 |
| SEAR | 5.048779 | 4.1515501 | 1.2161189 | 0.2320782 |
| WPR | -0.448861 | 0.2149949 | -2.0877753 | 0.0441629 |

***Table 3:*** *Simple Linear Regression of EUR Dependent and All Other Variables as Independent.*

From the results of the initial analysis, it is clear that there is a large error associated with the fit. Therefore the model mat not be a good model for this data set. The data may create a better model if the data for each flu region were normalized for number of reported cases. The data may create a better model if the individual region data were aligned for peak dates. In each case the peak number of reported cases is apporximately three weeks apart.

A simple multiple linear regression of the resulting data set allowing the EUR data to be dependent data, while using offset data from the WPR and AMR regions as independent data should reveal a strong relationship. Once again, the easiest model to use for this analysis is a simple regression model. The models should explore the dependence of current flu levels in the AMR and EUR WHO regions based on the levels of flu reported in the EMR, and WPR regions and the EMR, WPR, and AMR regions respectively. If we consider that the trend may not be merely west to east, but that there may be feedback from the EUR region to the AMR region then another possible model would be the dependence of flu cases in the AMR region on cases in the later two regions. The remaining regions are also normalized to number of reported cases to make the comparison equitable. (See **Table 4.**, **Table 5.**, and **Table 6.** below.)

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | -2114.6808300 | 633.1932228 | -3.339709 | 0.0021950 |
| EMR | 0.1460598 | 0.1427509 | 1.023179 | 0.3141417 |
| WPR | 1.6968388 | 0.1755900 | 9.663639 | 0.0000000 |

***Table 4:*** *Simple Linear Regression of AMR Dependent with EMR and WPR as Independent Using Date Aligned Data.*

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | -2428.3274404 | 868.7692632 | -2.7951351 | 0.0089590 |
| EMR | -0.5926022 | 0.1707743 | -3.4700903 | 0.0015987 |
| WPR | 2.3027461 | 0.4138434 | 5.5642928 | 0.0000047 |
| AMR | -0.0864317 | 0.2113249 | -0.4089991 | 0.6854456 |

***Table 5:*** *Simple Linear Regression of EUR Dependent with EMR, WPR, and AMR as Independent Using Date Aligned Data.*

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | -2258.7459219 | 732.1697218 | -3.0850032 | 0.0043473 |
| EMR | 0.1072310 | 0.1730699 | 0.6195824 | 0.5402103 |
| WPR | 1.8351641 | 0.3821847 | 4.8017730 | 0.0000408 |
| EUR | -0.0641558 | 0.1568604 | -0.4089991 | 0.6854456 |

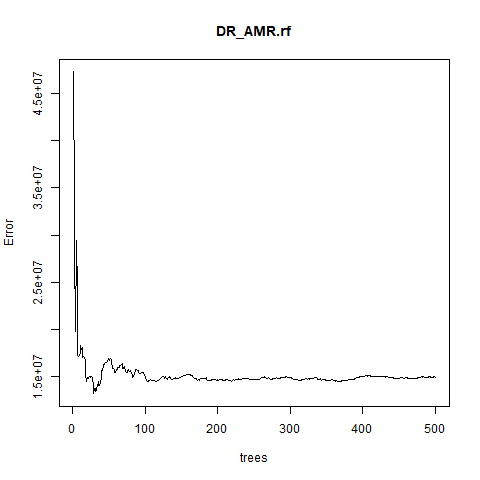
***Table 6:*** *Simple Linear Regression of AMR Dependent with EMR, WPR, and EUR as Independent Using Date Aligned Data*

# 5. Results

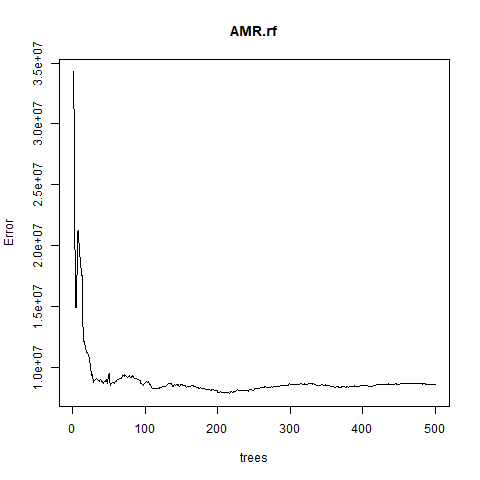
It is obvious from the regression data presented in **Table 4** that the cases of flu reported in the AMR WHO region are likely associated with the prior outbreak in the WPR region. The EMR region is likely not a contributing factor. Likewise, in **Table 5** the same can be said of the EUR WHO region outbreak. It seems to be much more heavily dependent on the WPR region outbreak than the outbreaks in the AMR or EMR regions. The results presented in **Table 4** were upheld by the results from **Table 6** where the linear regression set the AMR data as dependent and with EMR, WPR, and EUR as independent variables. In this case we also see the association with previous outbreaks being potentially linked to the WPR region outbreak.

## 5.1 Full analysis

In an attempt to determine how well a linear regression model fits the data that is being used a random forest model was constructed. Data that was unaligned data (**Figure 8 and Table S1**) and date aligned data (**Figure 9 and Table S2**) was used. In each case the AMR variable was held as dependent and the reported flu cases from the other WHO regions were the independent variables.



***Figure 8:*** *Graph of the Random Forest Output for the Unaligned Data Set.*



***Figure 9:*** *Graph of the Random Forest Output for the Date Aligned Data Set.*

# 6. Discussion

It is obvious from the data that there is a time/location dependence on seasonal flu outbreaks in the AMR and EUR WHO regions. The simple regression models coupled with the visual data exploration support the idea that seasonal flu outbreaks in the WPR WHO region are potentially predictive of outbreaks in Europe and the Americas.

The machine learning model resulted in a very high Mean Squared Error (MSE) with a percentage of explained variance close to one. This suggests that the machine learning model chosen may not be appropriate. If the MSE is very high (13640640 and 9173785, respectively) but the percentage of variance explained (87.64% and 92.72%, respectively) is close to one, this suggests a few possible scenarios:

1. **Over-fitting:** The model may be over-fitting to the training data, capturing noise or specific patterns that do not generalize well to unseen data. This can lead to high MSE as the model’s predictions are inaccurate for new data despite explaining a high percentage of variance in the training data. Since the data set is limited in size this is the most likely case.
2. **Complexity Issues:** The model might be too complex relative to the amount of training data available. Random forests are powerful and can capture complex relationships, but they can also over-fit if not properly tuned or if the data size is limited.
3. **Outliers or Anomalies:** High MSE could also be influenced by outliers or anomalies in the data set, which the model struggles to fit accurately.

For each of these issues there are several possible solutions:

*Overall Model Evaluation:* It is important to evaluate the performance of the model on separate test data set (not used during training) to assess its ability to generalize to new data accurately. In this case, perhaps using more than a single year of data as train/test data may yield better results.

*Hyperparameter Tuning:* Optimizing parameters such as tree depth, number of trees in the forest, minimum samples per leaf/node, etc., may help to find a balance between model complexity and generalization. Since this data set is size limited, this option may not be applicable.

*Feature Engineering and Selection:* Review of the features used in the model and considering if feature engineering (creating new features) or feature selection techniques may improve model performance and reduce over-fitting. Again, for this model it is unlikely that this will help due to the limited data set size.

*Data Quality Check:* It is important to verify the data quality, handle outliers appropriately, and ensure that the data set is representative of the problem domain without introducing biases or noise. In the data that was selected for this project, there was no examination for outliers. Again, including data from multiple years and normalizing that data to population, reported cases per population or some other metric may produce better results.

By addressing these aspects and fine-tuning the model, the model performance may be improved, over-fitting reduced, and its ability to make accurate predictions on new, unseen data enhanced.

Since the machine learning model was not suited for the limited data set size and may not be suited for this type of data other methods must be used to explore the hypothesis. The most conclusive evidence of the geographic and temporal dependence of seasonal flu epidemics comes from the graphic exploration of the data and from the simple linear regressions done in the data exploration and initial analysis steps. These analyses do show a trend for the seasonal influenza outbreak to circulate globally in a west to east direction. This trend is obvious from the graphic exploration alone. The linear regression results confirm this progression with the WPR WHO region being the strongest candidate for the origin of the seasonal flu outbreaks in both the AMR and EUR WHO regions. This analysis confirms other analyses that have been performed using more sophisticated methods and software. [***Bedford (2015)***]

## 6.1 Strengths and Limitations

In this case the data may not be amiable to a machine learning model. Even if prior years of data are included in the analysis the machine learning model may not be able to yield more than is has here. That being said, the relationships that are obvious in the simpler analysis are likely not sufficient to draw any concrete conclusion. The linear regression and visual exploration of the data is certainly compelling and support the idea that seasonal flu outbreaks in the WPR WHO region are potentially predictive of outbreaks in Europe and the Americas. That being said, the data that is available may have the depth to make this prediction.

## 6.2 Conclusions

The conclusion that can be drawn from this analysis is primarily due to the graphic exploration of the data and the simple linear regression work. It is obvious that there is a time/location dependence on the spread of the seasonal flu. However, more data is needed to complete this picture. For a detailed study of time/location dependence of the season flu strains the data needs an added component. The new component should be one that tracks the movement of individual organisms or, in this case, of individual strains of flu. In modern infection tracing the tool that allows this to be accomplished is genetic data. By sequencing DNA and determining the relationships between the circulating flu strains and then adding data on these lineages into the time/location data recorded by the WHO a far more complete model can be created.

# 7. References

Bedford, et. al. (2015). Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature*, *523*(7559), 217–220. <https://doi.org/10.1038/nature14460>

Domingo, E. (2020). Molecular basis of genetic variation of viruses: Error-prone replication. *Virus as Populations*, *2nd Ed.*, 35–71. https://doi.org/<https://doi.org/10.1016/B978-0-12-816331-3.00002-7>

Krammer, S., F. (2018). Influenza. *Nat Rev Dis Primers*, *4*(1), 3. <https://doi.org/10.1038/s41572-018-0002-y>

Moghadami, M. (2017). A narrative review of influenza: A seasonal and pandemic disease. *Iran J Med Sci*, *42*(1), 2–13.

Perdue, M. L., & Nguyen, T. (2012). The WHO research agenda for influenza: Two years later. *Bull World Health Organ*, *90*(4), 246. <https://doi.org/10.2471/blt.11.090175>

Rubin, R. (2021). Influenza’s unprecedented low profile during COVID-19 pandemic leaves experts wondering what this flu season has in store. *JAMA*, *326*(10), 899–900. <https://doi.org/10.1001/jama.2021.14131>

Taubenberger, J. K., & Kash, J. C. (2010). Influenza virus evolution, host adaptation, and pandemic formation. *Cell Host Microbe*, *7*(6), 440–451. <https://doi.org/10.1016/j.chom.2010.05.009>