Christine Gao & Heidi Sipes

SOCI 290   
Scott Duxbury

7 December 2021

Risk Analysis of HIV Prevalence in U.S. Counties

**Introduction**

The AIDS epidemic, caused by the HIV virus, has continued to plague the United States. Without a cure, the disease continues to significantly affect marginalized communities across the nation, most disproportionately racial and ethnic minorities and gay, bisexual, and men who have sex with other men. According to The Centers for Disease Control and Prevention (2021), in 2019 there were approximately 1.2 million identified cases of HIV in the United States. HIV continues to spread at high rates among injecting drug users (IDUs), males who have sex with other males (MSM), African Americans, and the younger age groups (U.S. Dept. of Health, 2021).

Increasing funding for HIV education at a community level is important for lowering HIV cases. The education and prevention strategies are most beneficial when targeted at the high-risk groups within geographic areas. For the COVID-19 virus, a K-means unsupervised training algorithm has proven effective in identifying COVID-19 clusters and the information provided with the algorithm has assisted in decision making regarding policies to help reduce the spread of COVID (Abdullah et al., 2021; Wu & Sha, 2020). A similar strategy might prove beneficial for HIV education and prevention policies. x1

**Background**

Our research question is: Do HIV cases display different clustering across geographic regions (county specific by state) based on gender, race, sexual orientation (MSM), and sexual orientation combined with IDU (MSM IDU)? Additionally, can we create an algorithm to assess the risk level of certain demographic groups and populations according to rate of HIV infection across these groups? Based on existing research, we hypothesize a k-means unsupervised training algorithm can be used to identify counties with HIV high risk groups based on prevalence of HIV cases.

HIV prevention and treatment programs are most beneficial when they are geographically targeted to reach the most at-risk population groups (Aral et al., 2015). A study conducted by Aral et al. (2015) demonstrated that using spatial concentration measurements to identify key geographic locations of HIV infection produced an efficient allocation of prevention targeted resources. The highest at-risk groups within the U.S. population include the MSM (men who have sex with men), youth (age 13-24), IDU (injecting drug users), and African American males (Boutrin et al., 2021; Christopoulos et al., 2011; Koenig et al., 2015; Strathdee & Stockman, 2010). Boutrin and Williams (2021) discuss the high rates of HIV among African American youth and they report that this high-risk demographic receives the least amount of intervention in the form of health policies. In addition, HIV affects the MSM Black population at a higher rate than the MSM White population (Lechuga et al., 2013). Christopoulos et al. (2011) note that one reason there is a high rate of HIV spread among the Black MSM population is because Black MSMs usually have Black MSM partners. Injecting drug use increases the risk to all ethnic and age groups due to the sharing of contaminated equipment used for injection and is responsible for an estimated 10% of HIV cases globally (Strathdee & Stockman, 2010). Phillips et al. (2020) conducted a study that demonstrated the effectiveness of youth directed education and intervention programs to help reduce HIV spread by making youth aware of behaviors like unprotected sex that may lead to HIV infection. In their report, Christopoulos et al. (2011) highlight the importance of targeting care toward the MSM communities and finding ways to better engage this community in HIV prevention and treatment programs.

**Data and Methods**

The dataset used for the analysis comes from the “2018 HIV infection rates” dataset found on Sakai and it measures the HIV infection case count (number of individuals) per county within the United States. The columns within the dataset contain the number of individuals infected with HIV within each county based on grouped categories like age, MSM, and race. Additionally, the dataset includes geographical spatial codes (or “GEO.ID”) of all counties in the U.S. and the rates of HIV infection among various age groups, racial demographic, and characteristics of the populations such as drug-injecting users and sexual behaviors. The variables we will be focusing on for this research project are HIV cases within the groups *men who have sex with men (MSM), injecting drug use (IDU), age 25-34 (Age.25.34), and the black population (Black).*

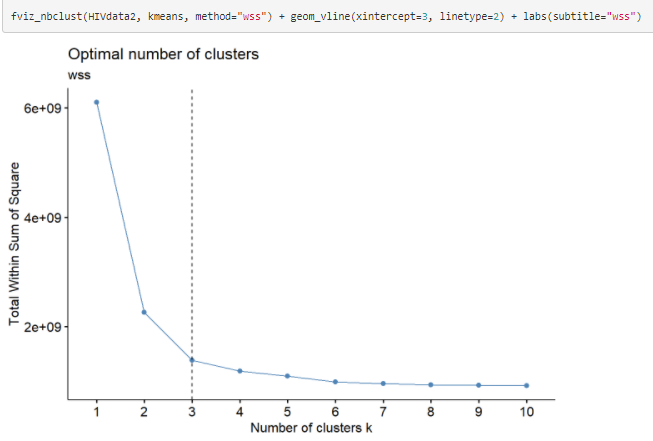
Since the data contains geographical information, we find value in plotting HIV cases on a map of the United States in order to draw conclusions regarding certain locations that are at increased risk or more susceptible to having a high number of infected individuals. Additionally, when taking into account the various factors mentioned above, we can develop some understanding of any underlying correlations between the various demographic groups and certain behaviors, and how they contribute to localized areas of high risk and HIV prevalence.

Our analysis strategy is a k-means clustering algorithm to determine any overlap across groups chosen for the research question. We chose the k-means algorithm based on the research by Wu and Sha (2020) and Abdullah et al. (2021). Both research groups found the k means algorithm beneficial for datasets that are small and also for its usefulness to identify patterns between objects. Our data was similar to the data used by Wu and Sha (2020) and Abdullah et al. (2021) and was easy to develop clusters by dividing the data into groups with similar characteristics and assigning risks for counties with the data given. All statistical analyses were conducted in R Studio.

**Results**

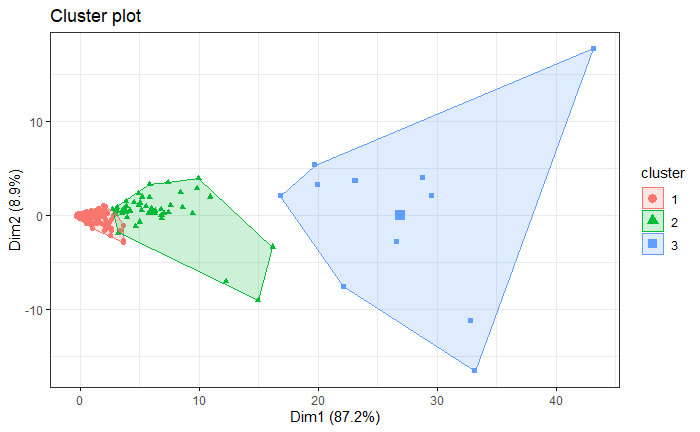
Before conducting our analysis, we utilized the function ‘fviz\_nbclust’ from the package ‘factoextra’ to determine and visualize the optimal number of clusters using various methods. The result of the function using the within cluster sum of squared errors method (wss) proposes 3 as the optimal number of subgroups, which we will use when computing the k-means clusters.

**Figure 1. Optimal number of clusters using fviz\_nbclust() function in R, method=”wss”**



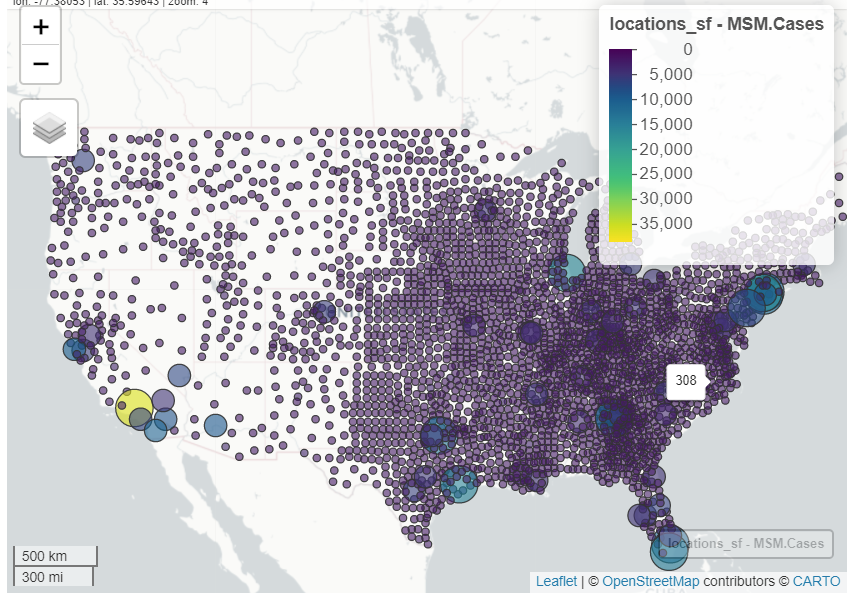
After determining the number of sub-groups, we performed our unsupervised k-means algorithm using the ‘kmeans()’ function from the ‘stats’ package. The graphed results of the clusters are as follows.

**Figure 2. Cluster plot of counties in the U.S. grouped by k-means algorithm, k=3**

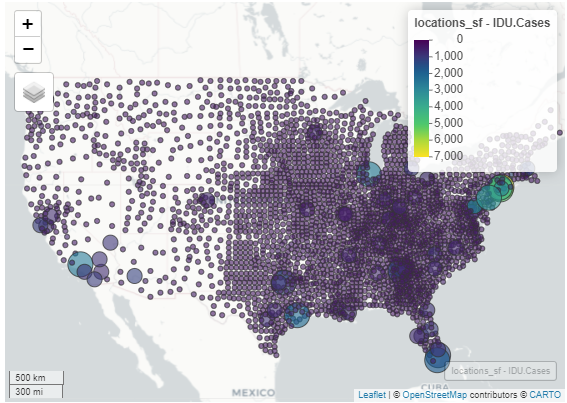


Further investigation of the group assignments reveals that each cluster is correlated with the number of HIV cases. Counties that reported highest overall cases across all variables selected were assigned a label of 3, while counties with the lowest number of cases were assigned a label of 1. To reflect this data geographically, we plotted the coded clusters on an interactive map of the United States reflecting county-level HIV cases. The size of the clusters also reflects the assigned label number. Figures 3-6 reflect the plotted clusters of each unique variable.

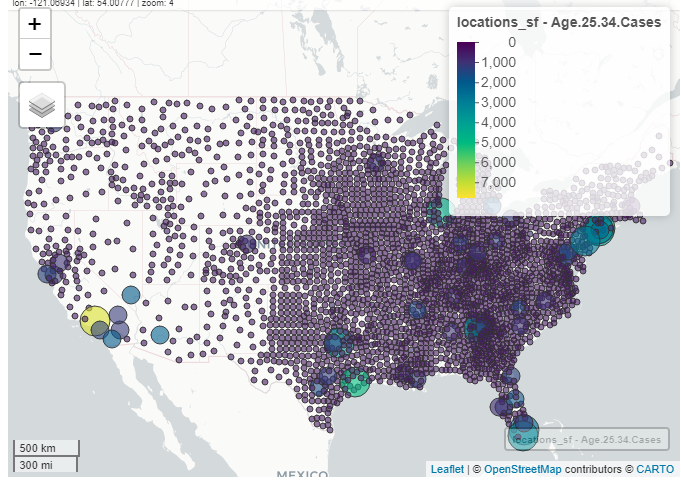
**Figure 3. Interactive spatial plot using mapview() of MSM population**

****

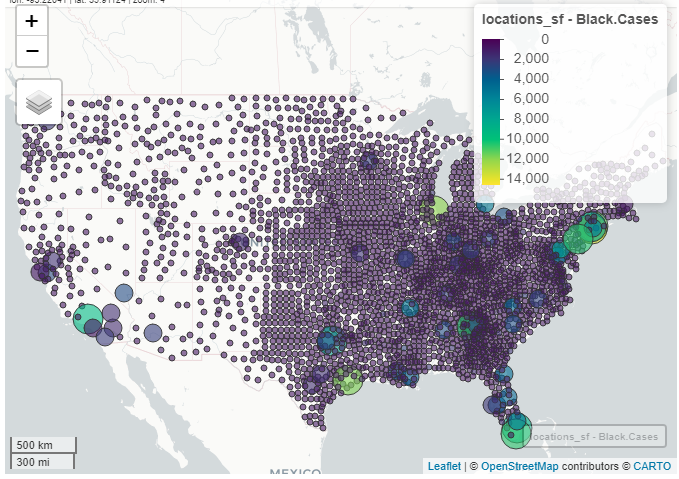
**Figure 4. Interactive spatial plot using mapview() of IDU population**

****

**Figure 5. Interactive spatial plot using mapview() of Age 25-34 population**

****

**Figure 6. Interactive spatial plot using mapview() of black population**

****

To evaluate the accuracy of our unsupervised model and to determine if our results are replicable, we conducted three additional validation tests to ensure the categories assigned are comparable to the dataset. In sample 1, we sampled 1000 random individuals from the original dataset with the same seed. In sample 2, we sampled 1000 different random individuals by changing the set.seed() value. Finally, in sample 3, we sampled 2000 random individuals using the same set.seed() value as sample 1. The performance results are as follows in Table 1.

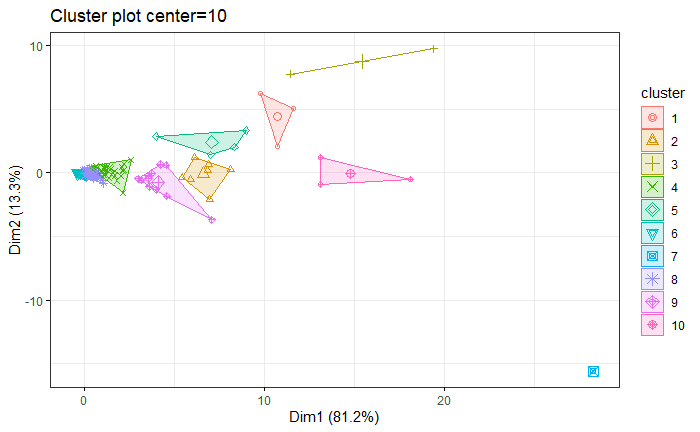
**Table 1. K-means algorithm unsupervised model performance**

| **Performance Measures** |  | **Sample 1**  **n=1000**  **set.seed=21093** | **Sample 2**  **n=1000**  **set.seed=21045** | **Sample 3**  **n=2000**  **set.seed=21093** |
| --- | --- | --- | --- | --- |
| ***Accuracy*** |  | 0.976 | 0.991 | 0.998 |
| ***Recall*** |  |  |  |  |
|  | ***Cluster 1*** | 0.984 | 1 | 1 |
|  | ***Cluster 2*** | 0.529 | 0 | 0.875 |
|  | ***Cluster 3*** | 0.984 | 1 | 1 |
| ***Precision*** |  |  |  |  |
|  | ***Cluster 1*** | 1 | 0.994 | 0.998 |
|  | ***Cluster 2*** | 1 | 0 | 0.966 |
|  | ***Cluster 3*** | 0 | 0.5 | 1 |
| ***F1*** |  |  |  |  |
|  | ***Cluster 1*** | 0.992 | 0.997 | 0.999 |
|  | ***Cluster 2*** | 0.692 | N/A | 0.918 |
|  | ***Cluster 3*** | 0 | 0.667 | 1 |

Our model appears to be good, but not great when assigning label values to our randomly selected samples. Cluster 1, or the group with the lowest risk score, was labeled more consistently than groups originally assigned a risk value of 2 or 3 (medium to high risk). This is likely because there simply exist fewer counties considered to be high risk than there are low risk, and these counties are distinct in terms of prevalence of HIV cases.

We assessed the performance of our model again, but with a greater number of centers, k=10 as shown in Fig. 7.

**Figure 7. Cluster plot of counties in the U.S. grouped by k-means algorithm, k=10**



While there is less overlap between groups, the distinction between clusters is lost, e.g. the county with the highest number of MSM cases received a label of 5, while counties with the lowest number of cases were assigned a label of 7. We do not believe the dataset contains outliers or extreme values given the nature of the variables (cases per county), but we could examine a clustering algorithm that is less susceptible to significant variations in the data, such as a k-medians algorithm. Given this assessment and constraints of our analysis, we will continue our discussion using the recommended number of clusters and without removal of any data points for the final analysis.

**Discussion and Conclusion**

Our findings suggest that when investigating the prevalence of HIV cases across counties in the U.S, the various populations chosen correlate strongly in terms of groups considered to be at high risk for HIV infection. Within the hotspots identified using the interactive map, larger clusters consistently reflected areas of increased infection rate.

Limitations to our study include the use of an unsupervised machine learning algorithm is the lack of existing labels to cross validate the results of the labels. Taking into consideration the possibility of creating more clusters or subgroups from the data, perhaps the granularity of the assignments would be more clear and lead to better distinctions between subgroups. Another possibility for confounding in our results may be the existence of marginalized communities in our regions of interest. Disparities in socioeconomic status have been found to be related to increased HIV risk for males, and higher educational attainment has been proven to be negatively associated with HIV prevalence by increasing self-protection and promoting behaviors that lower vulnerability to HIV risk (Igulot & Magadi, 2018). Thus socioeconomic status is likely a factor influencing assignments made by our k-means algorithm.

The policy implications of our study emphasize geographic hotspots and underlying causes of higher HIV prevalent counties. Further investigation should be conducted on the socioeconomic differences of communities, and thus policies should focus on increasing education, funding for treatment and prevention, and consideration for more regular diagnoses in regions considered to be HIV hotspots.

**Ethics Statement**

Our research adheres to Salganik’s four principles of digital age social research ethics: respect for persons, beneficence, justice, and respect for law and public interest. All data within the dataset is anonymous with no personally identifying information, and poses no harm to any subjects involved in the study.

**References**

Abdullah, D., Susilo, S., Ahmar, A.S. et al. The application of K-means clustering for province

clustering in Indonesia of the risk of the COVID-19 pandemic based on COVID-19 data. *Qual Quant (2021)*. https://doi.org/10.1007/s11135-021-01176-w

Aral, S. O., Torrone, E., & Bernstein, K. (2015). Geographical targeting to improve progression

through the sexually transmitted infection/HIV treatment continua in different populations. *Current opinion in HIV and AIDS*, *10*(6), 477–482. https://doi.org/10.1097/COH.0000000000000195

Boutrin, M. C., & Williams, D. R. (2021). What Racism Has to Do with It: Understanding and

Reducing Sexually Transmitted Diseases in Youth of Color. Healthcare (Basel, Switzerland), 9(6), 673. https://doi.org/10.3390/healthcare9060673

Centers for Disease Control and Prevention. (2021, June 24). *HIV Surveillance Report*. Retrieved

from https://www.cdc.gov/hiv/statistics/overview/index.html

Christopoulos, K. A., Das, M., & Colfax, G. N. (2011). Linkage and retention in HIV care

among men who have sex with men in the United States. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America*, *52 Suppl 2*(Suppl 2), S214–S222. https://doi.org/10.1093/cid/ciq045

Igulot, Patrick, Monica Magadi. (2018) Socioeconomic Status and Vulnerability to HIV Infection

in Uganda: Evidence from Multilevel Modelling of AIDS Indicator Survey Data. *AIDS research and treatment, 2018,* 7812146, https://doi.org/10.1155/2018/7812146

Koenig, L. J., Hoyer, D., Purcell, D. W., Zaza, S., & Mermin, J. (2016). Young People and HIV:

A Call to Action. *American journal of public health*, *106*(3), 402–405.

https://doi.org/10.2105/AJPH.2015.302979

Lechuga, J., Owczarzak, J. T., & Petroll, A. E. (2013). Marketing the HIV test to MSM: ethnic

differences in preferred venues and sources. *Health promotion practice*, *14*(3), 433–440. https://doi.org/10.1177/1524839912460870

Phillips, G., 2nd, McCuskey, D. J., Felt, D., Curry, C. W., Ruprecht, M. M., Wang, X., & Beach,

L. B. (2020). Association of HIV Education with HIV Testing and Sexual Risk Behaviors Among US Youth, 2009-2017: Disparities Between Sexual Minority and Sexual Majority Youth. *Prevention science: the official journal of the Society for Prevention Research*, *21*(7), 898–907. https://doi.org/10.1007/s11121-020-01153-z

Strathdee, S. A., & Stockman, J. K. (2010). Epidemiology of HIV among injecting and non-

injecting drug users: current trends and implications for interventions. Current HIV/AIDS reports, 7(2), 99–106. https://doi.org/10.1007/s11904-010-0043-7

U.S. Department of Health and Human Services. (2021, June 2). *U.S. Statistics*. Retrieved from

https://www.hiv.gov/hiv-basics/overview/data-and-trends/statistics

Wu, J., & Sha, S. (2021). Pattern Recognition of the COVID-19 Pandemic in the United States:

Implications for Disease Mitigation. *International journal of environmental research and public health*, *18*(5), 2493. https://doi.org/10.3390/ijerph18052493