Discovering COVID-19 Waves with NMF

Kevin Quinn, Evimaria Terzi, Mark Crovella

The problem

Context

- COVID-19 has spread throughout the U.S. in waves of differing severity since the beginning of the pandemic [5]
- March - May of 2020 saw an initial spike of cases followed by a period of slight relief in the summer
- Fall and Winter of 2020 brought severe surges of COVID spread soon after [3]
- Other, less obvious waves may have also impacted the progression the disease

Questions to Consider

- How might we transform a collection of case data into distinct patterns of spread, i.e. waves?
- How might one characterize a given wave given that different locations might experience it differently? (at different times or at differing levels of severity)
- Can we find geographical correlation in where the waves happen?

Non-Negative Matrix Factorization

Decomposes

A reliable algorithm for factoring a matrix M into parts X and Y determined by its rank k

Is strictly positive

Works particularly well with cumulative COVID case data because it works on the condition that M, X, and Y contain strictly positive values

Highlights underlying patterns

Decomposition allows for an interesting study of the basis vectors that make up the data. What they tell us often reveal patterns that aren’t at first obvious

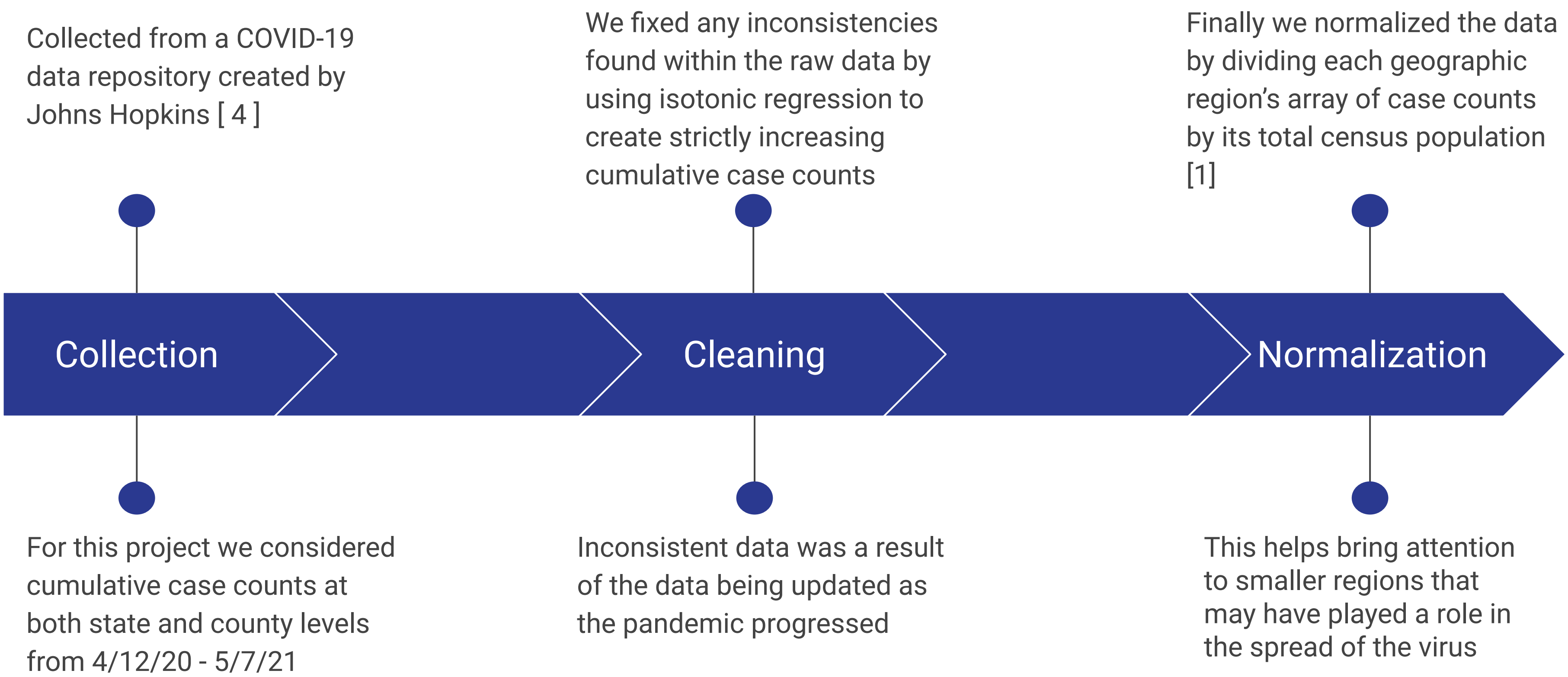
Related Studies

- A previous study [2] performed earlier in the pandemic used many of the same methods presented here
- However, they focused specifically on using Kmeans clustering to group states based on their infection patterns
- Our study has moved away from clustering because we believe it limits the geographical patterns found solely by using NMF
- Being at a later point pandemic our study also has the advantage of having more data to be able to find and characterize distinct wave patterns

Implementation

Studying COVID-19 data using NMF

Data Preparation



Collection

Collected from a COVID-19 data repository created by Johns Hopkins [4]

For this project we considered cumulative case counts at both state and county levels from 4/12/20 - 5/7/21

Cleaning

We fixed any inconsistencies found within the raw data by using isotonic regression to create strictly increasing cumulative case counts

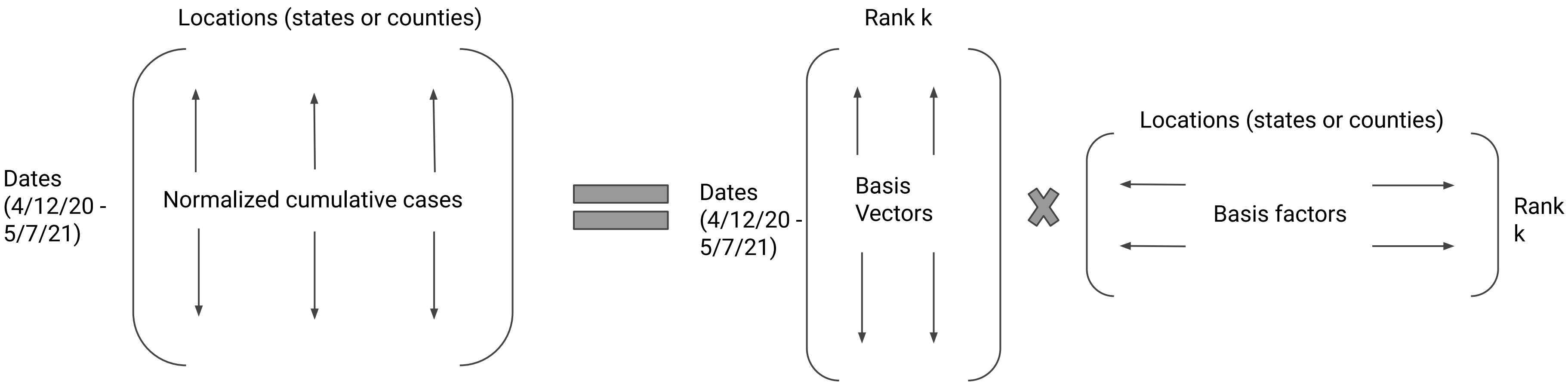
Inconsistent data was a result of the data being updated as the pandemic progressed

Normalization

Finally we normalized the data by dividing each geographic region’s array of case counts by its total census population [1]

This helps bring attention to smaller regions that may have played a role in the spread of the virus

NMF setup


Locations (states or counties)
Dates (4/12/20 - 5/7/21)
Normalized cumulative cases
Rank k
Dates (4/12/20 - 5/7/21)
Basis Vectors
Locations (states or counties)
Basis factors
Rank k

- Performed with sklearn's NMF implementation [6]

Interpretation

- Basis vectors are case curves can be used to reconstruct the original data
- In other words the basis vectors can be interpreted as “waves” and the combination of all these “waves” gives a complete picture of the entire pandemic
- Basis factors are unique numbers assigned to each geographical region that determine the weight that region gives to each “wave”
- For example:
  - Y[0][0] = large factor ---> Location 0 was severely affected by wave 0
  - Y[1][0] = small factor ---> Location 0 was minimally affected by wave 1

Choosing a Rank

- We tested a number of different ranks and compared the Mean Square Error between the original data and the newly reconstructed data using rank k

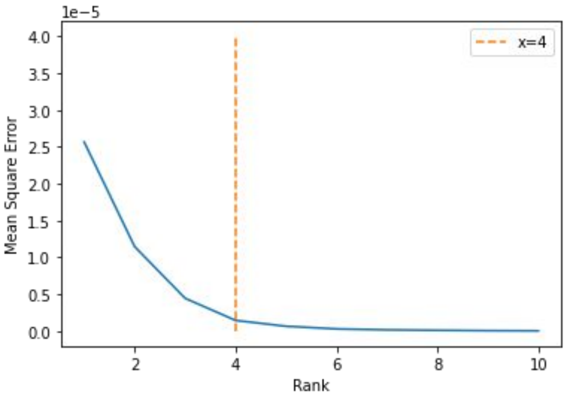


Figure 1: Error by rank for State level data

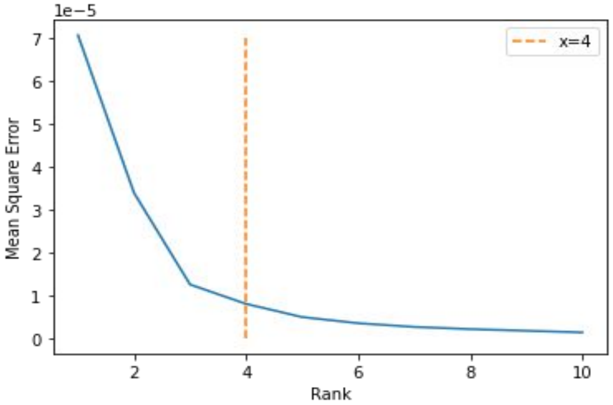
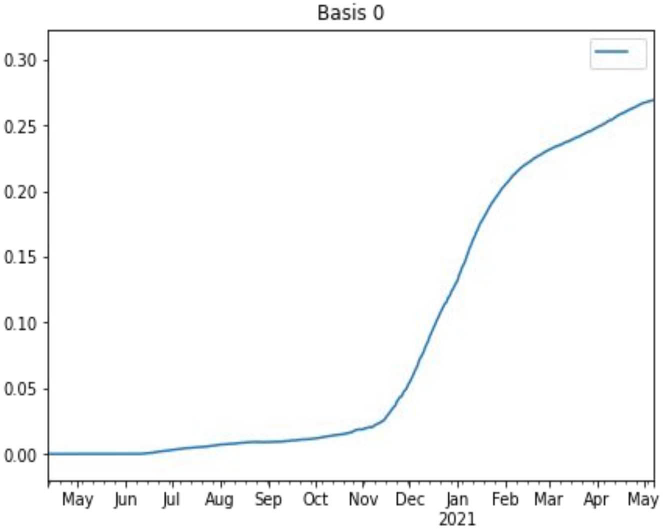


Figure 2: Error by rank for County level data

- To pick from these ranks we simply selected a value where the error began to drop of ("elbow" method) ---> Rank = 4

State Level Results with Rank 4 -- Basis 0



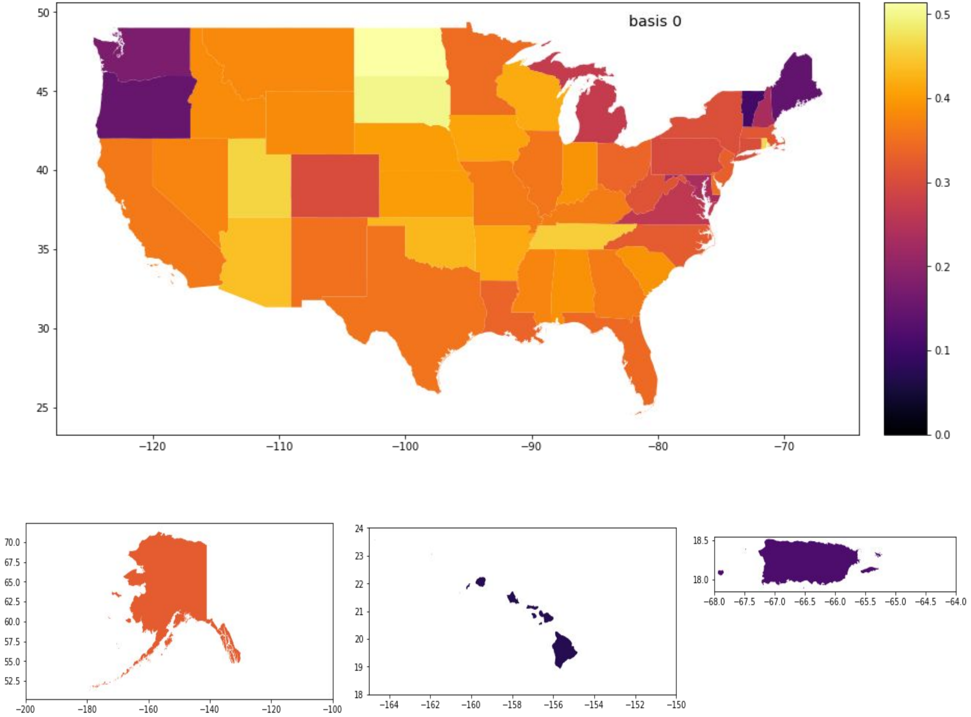
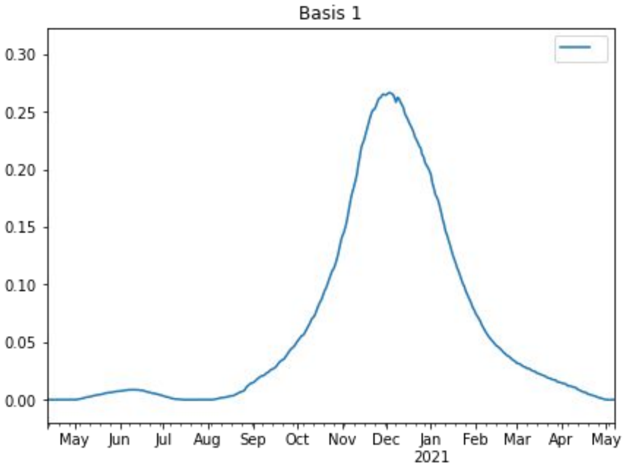


Figure 3: First basis vector of X (Left) -- time period of 4/12/20 - 5/7/21 -- along with a US map (50 states + Puerto Rico) with each state colored by its corresponding factor for Basis 0 in Y

State Level Results with Rank 4 -- Basis 1



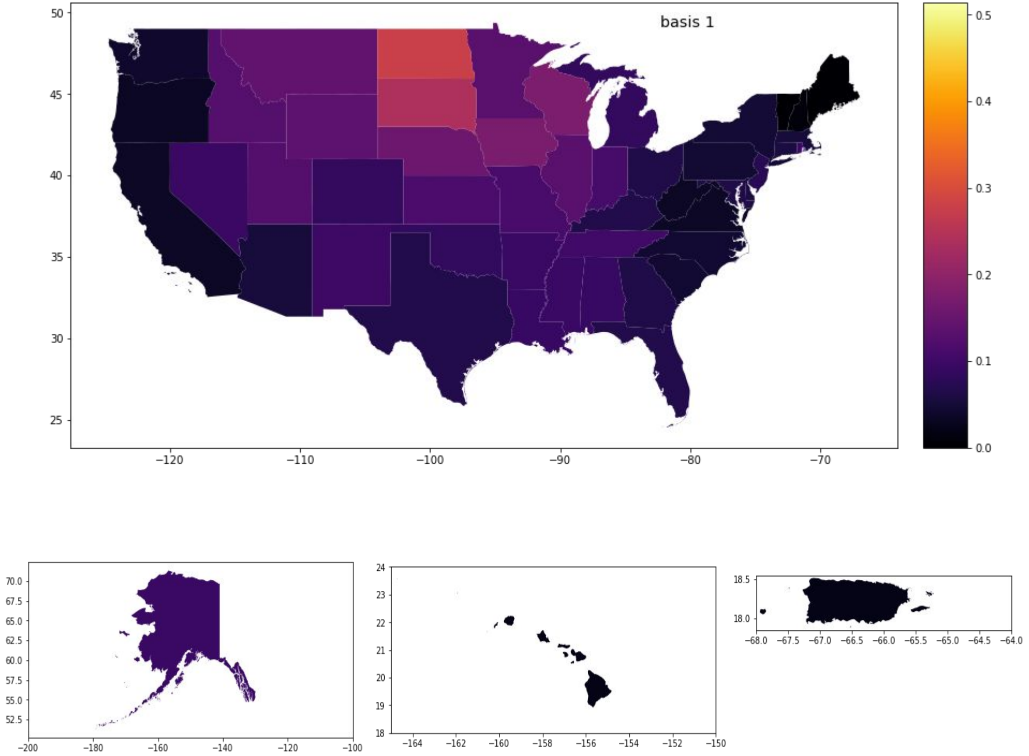
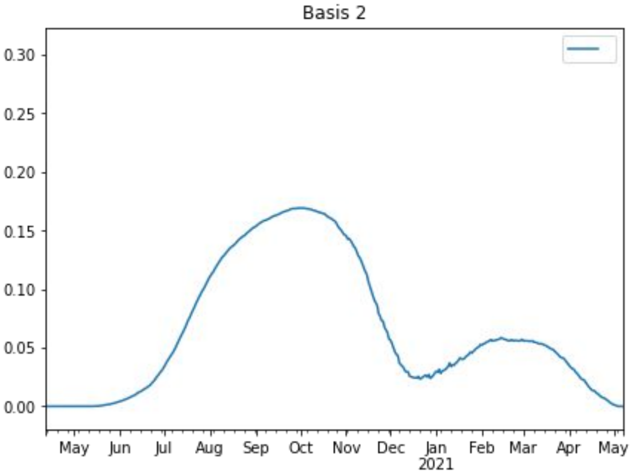


Figure 4: Second basis vector of X (Left) -- time period of 4/12/20 - 5/7/21 -- along with a US map (50 states + Puerto Rico) with each state colored by its corresponding factor for Basis 1 in Y

State Level Results with Rank 4 -- Basis 2



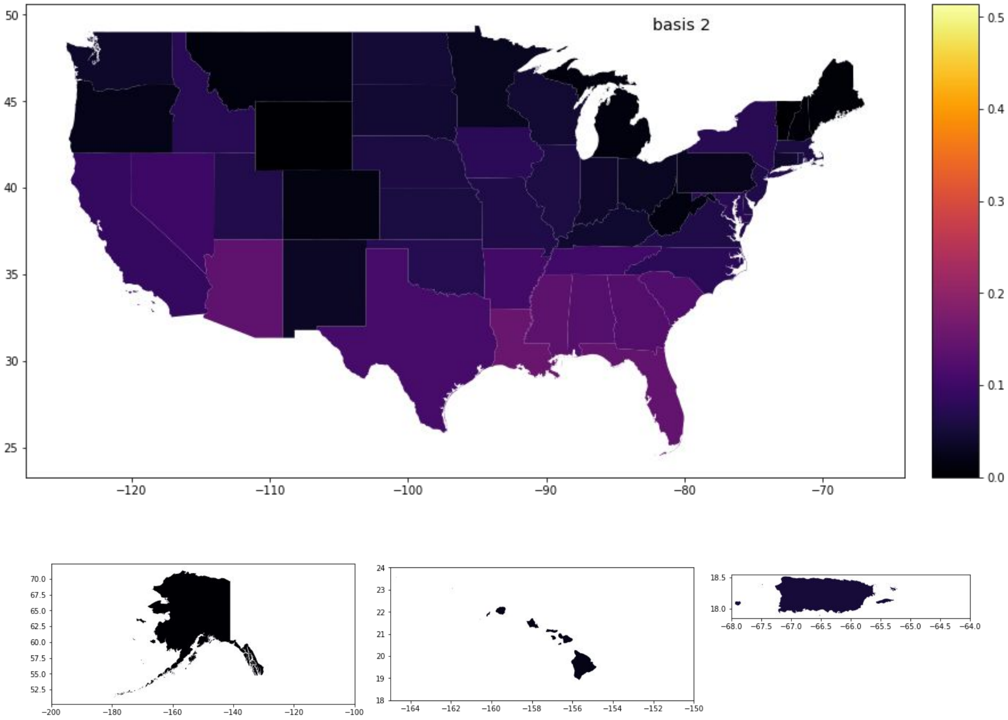
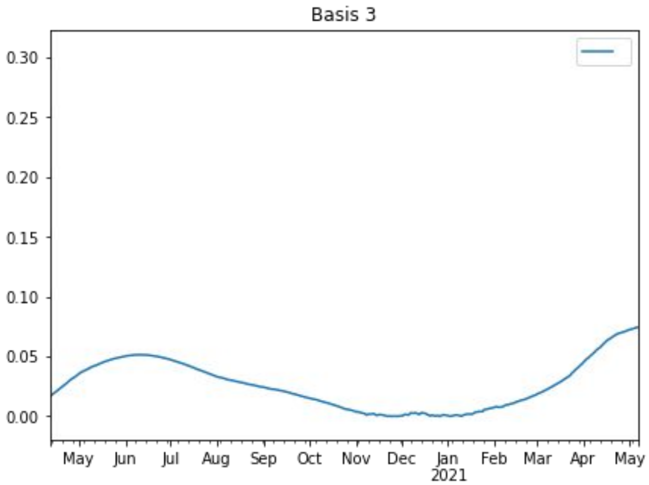


Figure 5: Third basis vector of X (Left) -- time period of 4/12/20 - 5/7/21 -- along with a US map (50 states + Puerto Rico) with each state colored by its corresponding factor for Basis 2 in Y

State Level Results with Rank 4 -- Basis 3



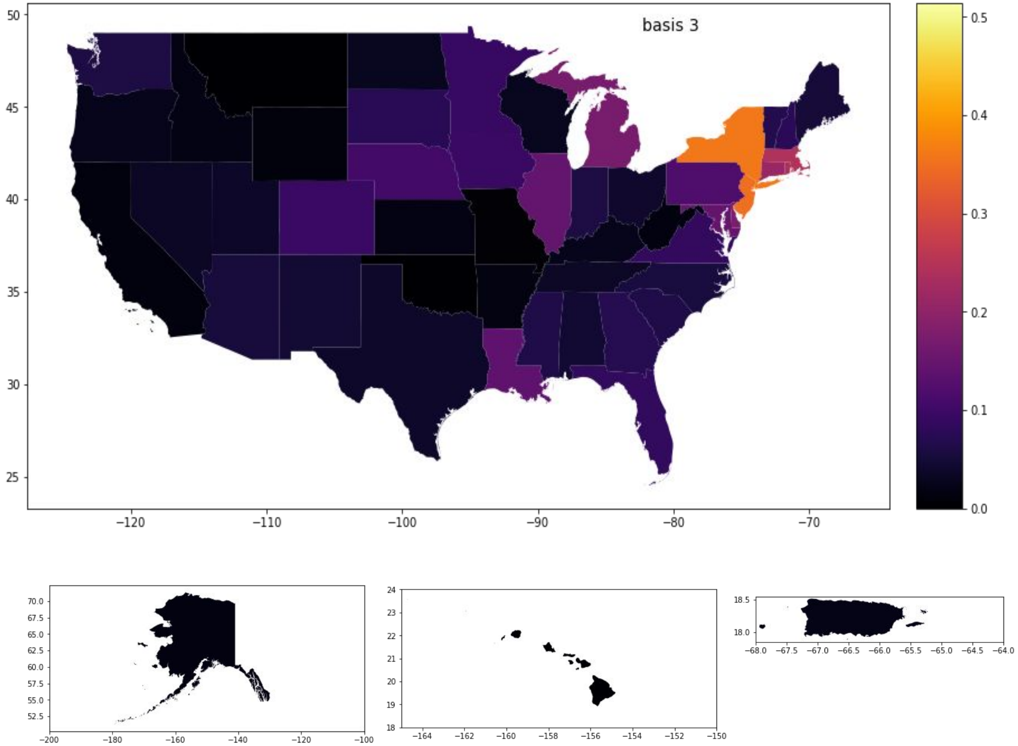


Figure 6: Fourth basis vector of X (Left) -- time period of 4/12/20 - 5/7/21 -- along with a US map (50 states + Puerto Rico) with each state colored by its corresponding factor for Basis 3 in Y

Discussion

- Basis 0 (figure 3) seems to be a general representation of the entire pandemic. Lighter colored states were, per capita, more severely affected during the considered time period
- Basis 0's map of factors is a nearly identical representation of total US cases per capita which is displayed for comparison here:

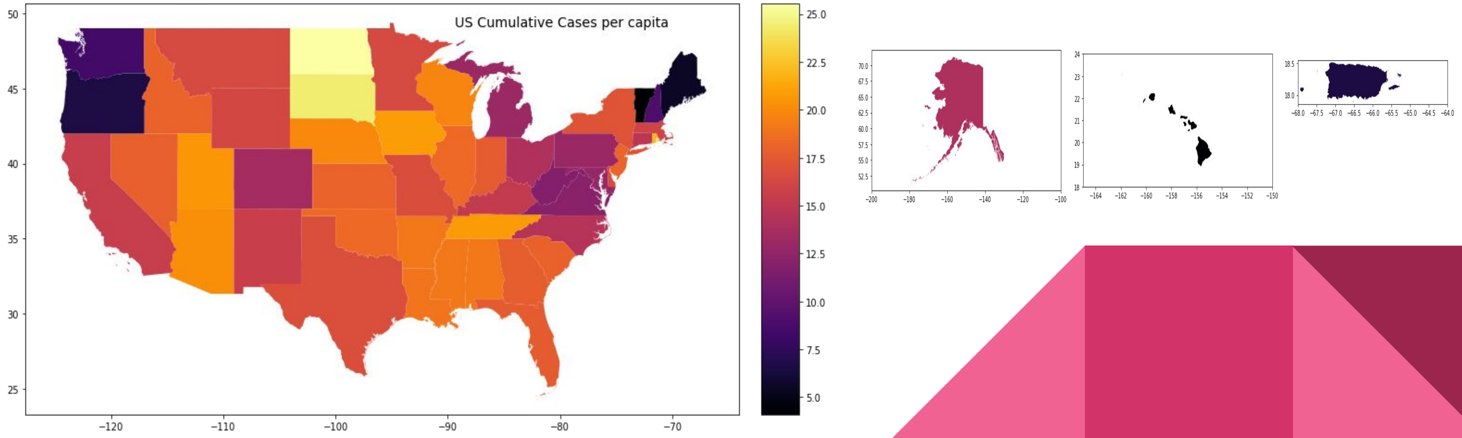


Figure 7: US map with each state shaded by: total # of confirmed cases divided by total population

Time period of 4/12/20 - 5/7/21

Discussion

- If Basis 0 gives a general version of the US cumulative case curve, then the other basis vectors must be accounting for local variation
- Basis 1 (figure 4) shows a clear spike in cases around the time period of late November to early December centered around North and South Dakota
- The timeline + spike in cases agrees with daily case data for North and South Dakota as well as other states in the midwest

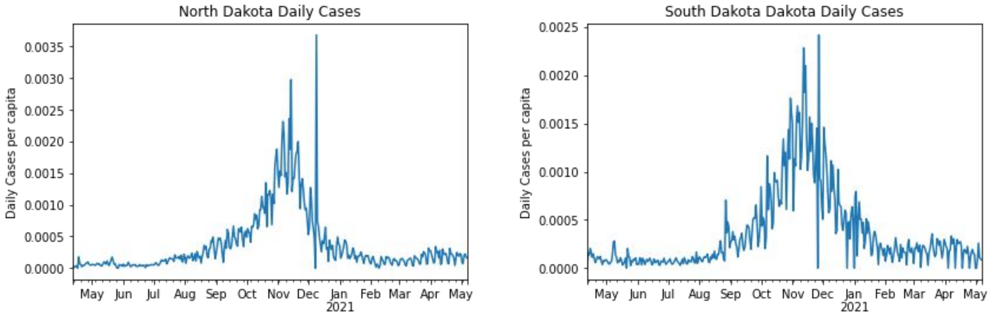


Figure 8: North Dakota (Left) and South Dakota (right) daily confirmed cases per capita. Time period of 4/12/20 - 5/7/21

Discussion

- Basis 2 (figure 5) seems to be representing two surges in southern states from late summer until winter of 2020 as well as in early 2021
- Basis 3 (figure 6) is representative of states like New York and Massachusetts which experienced early surges in March - April 2020 as well as surges from Dec. 2020 - April 2021

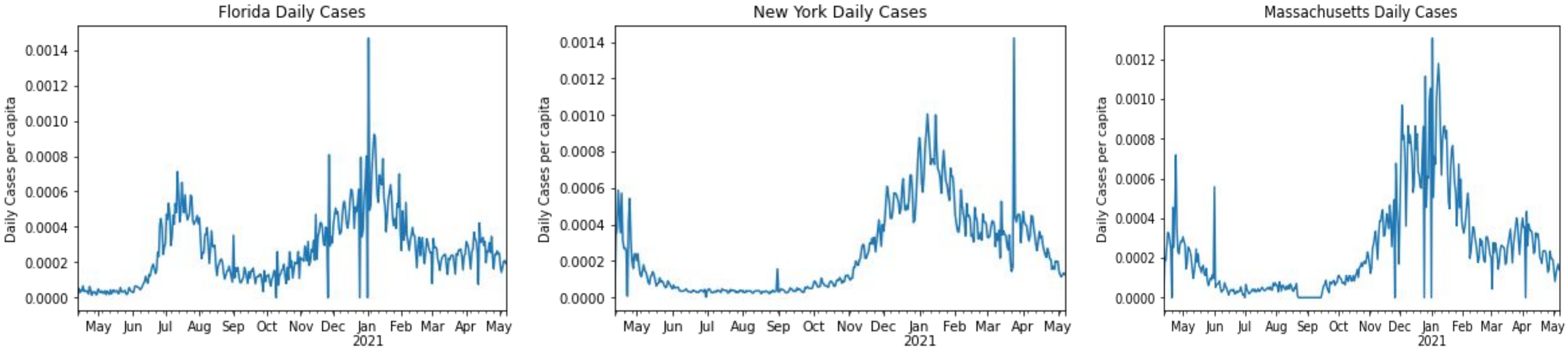
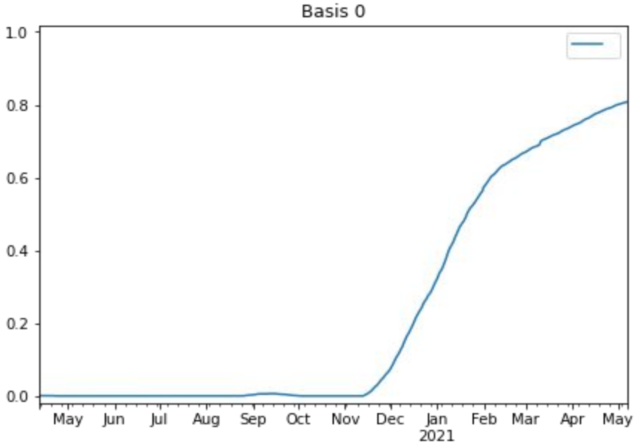


Figure 9: Florida (Left), New York (middle), and Massachusetts (right) daily confirmed cases per capita (divided by population) -- time period of 4/12/20 - 5/7/21

County Level Results Rank = 1 -- Basis 0



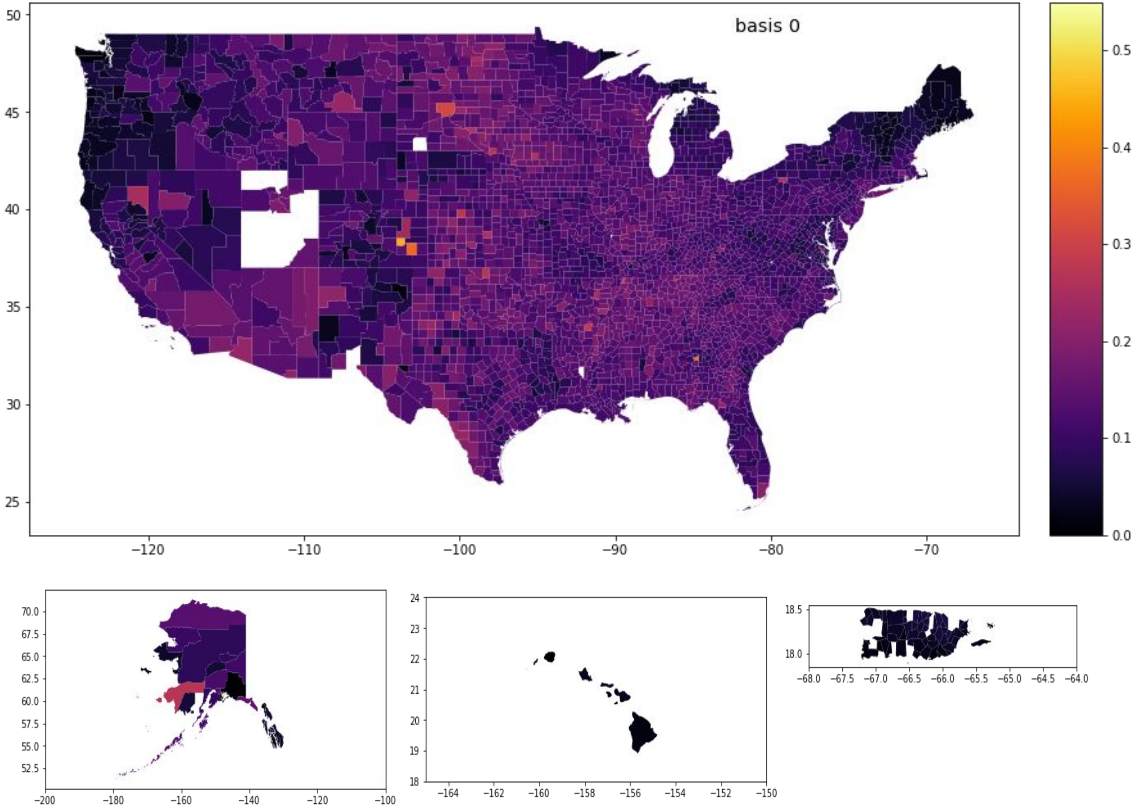
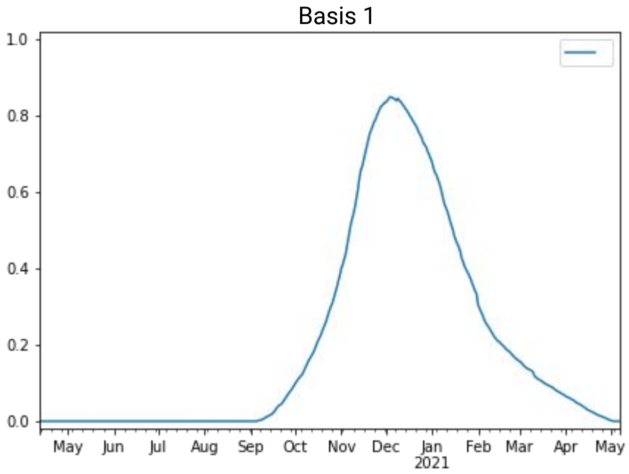


Figure 10: First basis vector of X (Left) -- time period of 4/12/20 - 5/7/21 -- along with a map of 3,194 US counties colored by their corresponding factor for Basis 0 in Y

Note: There are some counties which Johns Hopkins does not report in its data. These counties appear blank and were removed from the dataset before NMF.

County Level Results Rank = 1 -- Basis 1



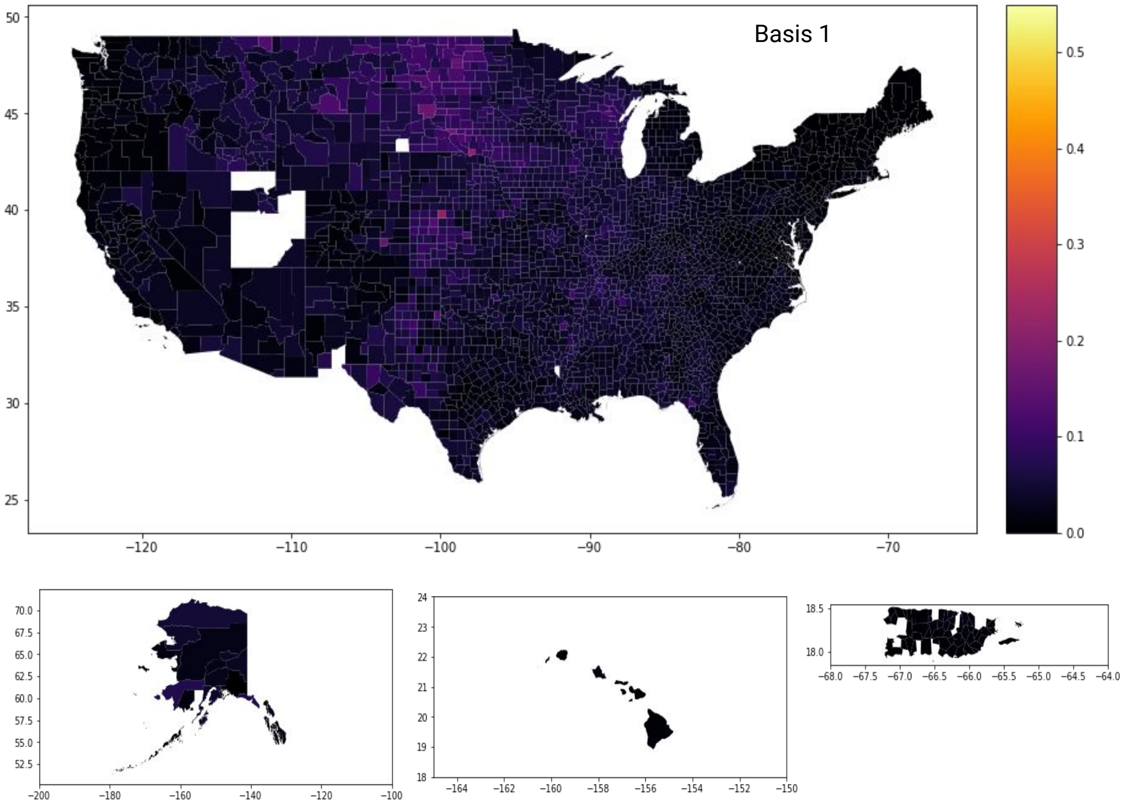
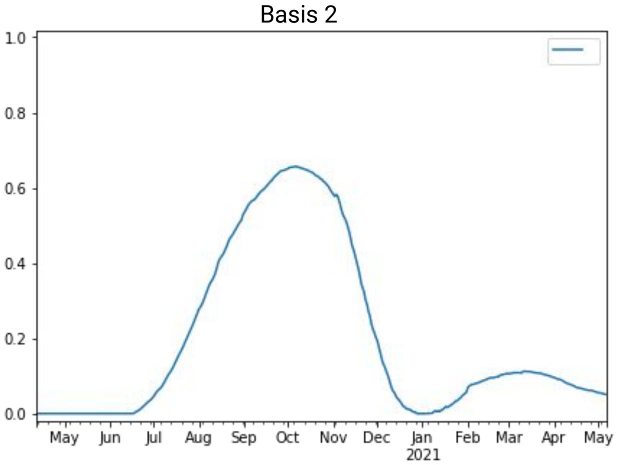


Figure 11: Second basis vector of X (Left) -- time period of 4/12/20 - 5/7/21 -- along with a map of 3,194 US counties colored by their corresponding factor for Basis 1 in Y

Note: There are some counties which Johns Hopkins does not report in its data. These counties appear blank and were removed from the dataset before NMF.

County Level Results Rank = 1 -- Basis 2



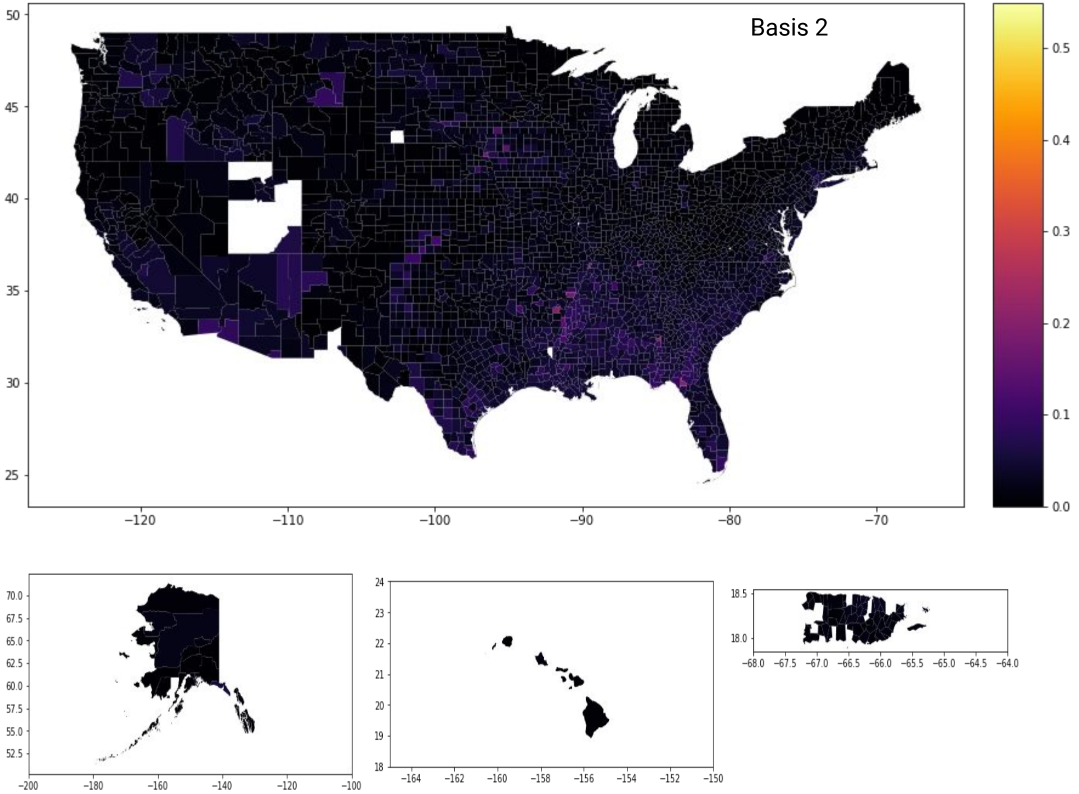
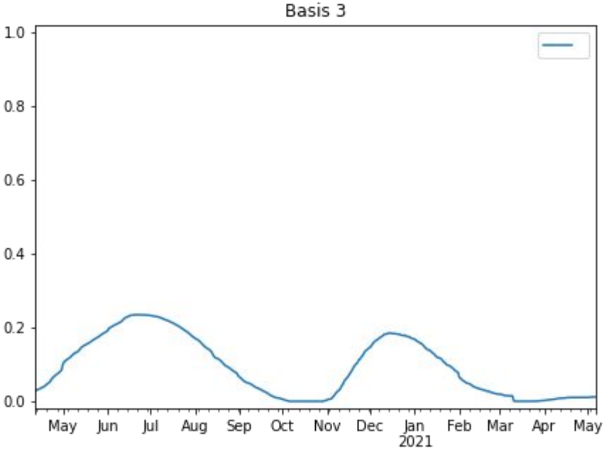


Figure 11: Third basis vector of X (Left) -- time period of 4/12/20 - 5/7/21 -- along with a map of 3,194 US counties colored by their corresponding factor for Basis 1 in Y

Note: There are some counties which Johns Hopkins does not report in its data. These counties appear blank and were removed from the dataset before NMF.

County Level Results Rank = 1 -- Basis 3



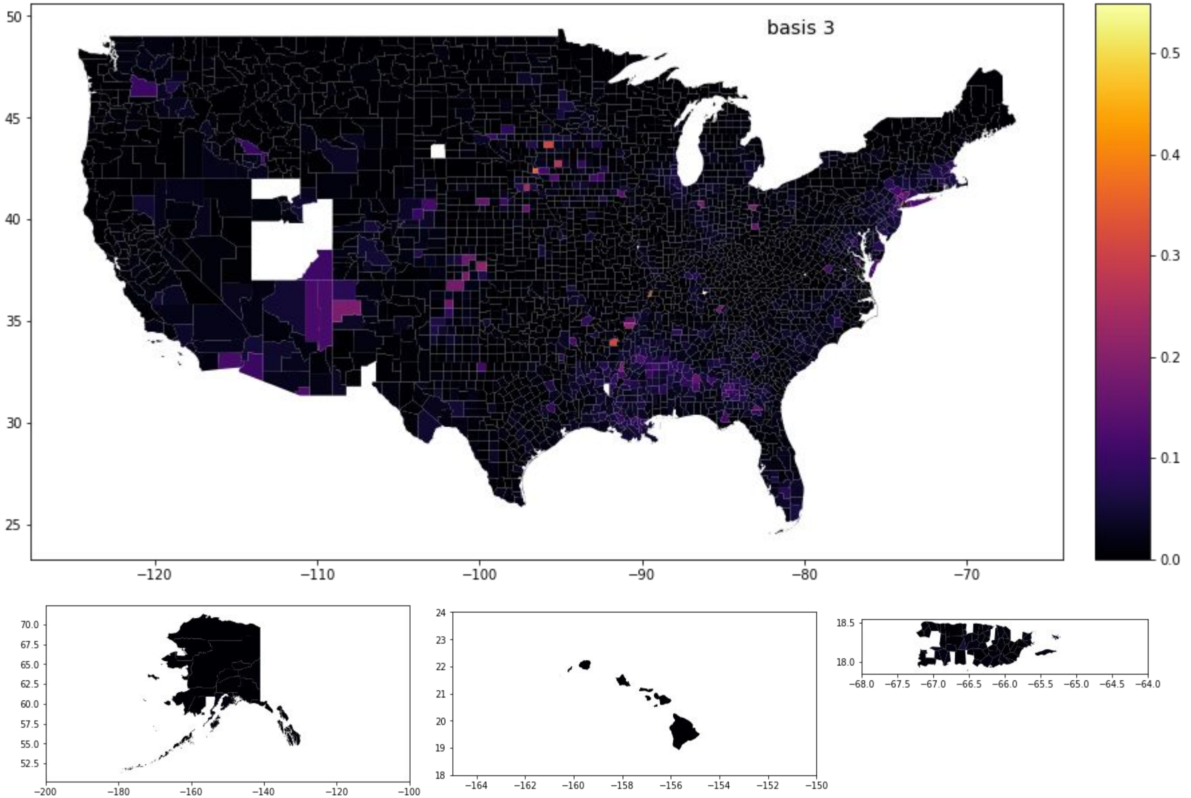


Figure 11: Fourth basis vector of X (Left) -- time period of 4/12/20 - 5/7/21 -- along with a map of 3,194 US counties colored by their corresponding factor for Basis 3 in Y

Note: There are some counties which Johns Hopkins does not report in its data. These counties appear blank and were removed from the dataset before NMF.

Discussion

- The county level results agree with what we saw at the state level
- The basis vectors have changed in size but still show similar patterns
- The geographical trends in the basis factor maps are a harder to distinguish but do agree upon close inspection

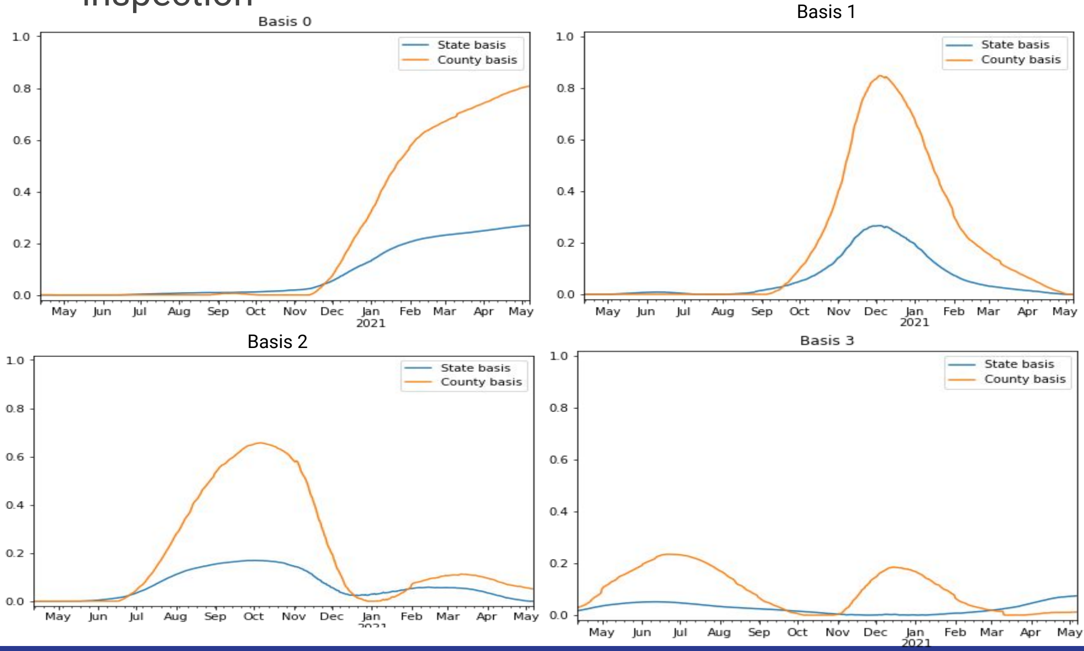


Figure 12: A comparison of state level to county level NMF basis vectors. Basis 0 (Top left), Basis 1 (Top right), Basis 2 (bottom left), Basis 3 (bottom right). State basis vectors are shown in blue and county level are shown in orange.

Conclusion

- At both the state and county levels NMF shows similar patterns of local variation that agree with what is seen in the cumulative/daily case data.
- Propose that NMF is a simple and cost effective way to pinpoint and characterize local variations or “waves” of COVID-19
- Clear geographical patterns in the maps further reinforce our proposal and are a topic of future research within our project

References

1. Bureau, U. S. C. (n.d.). *US Census Data*. Census.gov. https://www.census.gov/data.html.
2. Chen, J., Yan, J., & Zhang, P. (2021, January 15). *Clustering US States by Time Series of COVID-19 New Case Counts with Non-negative Matrix Factorization*. arXiv.org. https://arxiv.org/abs/2011.14412.
3. Gamio, L. (2021, February 10). *Half of U.S. Coronavirus Deaths Have Come Since Nov. 1*. The New York Times. https://www.nytimes.com/interactive/2021/02/10/us/coronavirus-winter-deaths.html.
4. Johns Hopkins University. (n.d.). *CSSEGISandData/COVID-19*. GitHub. https://github.com/CSSEGISandData/COVID-19.
5. Maragakis, L. L. (n.d.). *Coronavirus Second Wave? Why Cases Increase*. Johns Hopkins Medicine. https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/first-and-second-waves-of-coronavirus.
6. *sklearn.decomposition.NMF*¶. scikit. (n.d.). https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.NMF.html.