A Brief Visualization and Analysis of the Immigration Network in the U.S.

Immigration has long been both a critical social topic and of my personal interest on labot economy, and network analysis would contribute some interesting findings. The project aims to look at the migration flow between the 51 states (including District of Columbia), and find out potential factors crucial to the immigration flow.

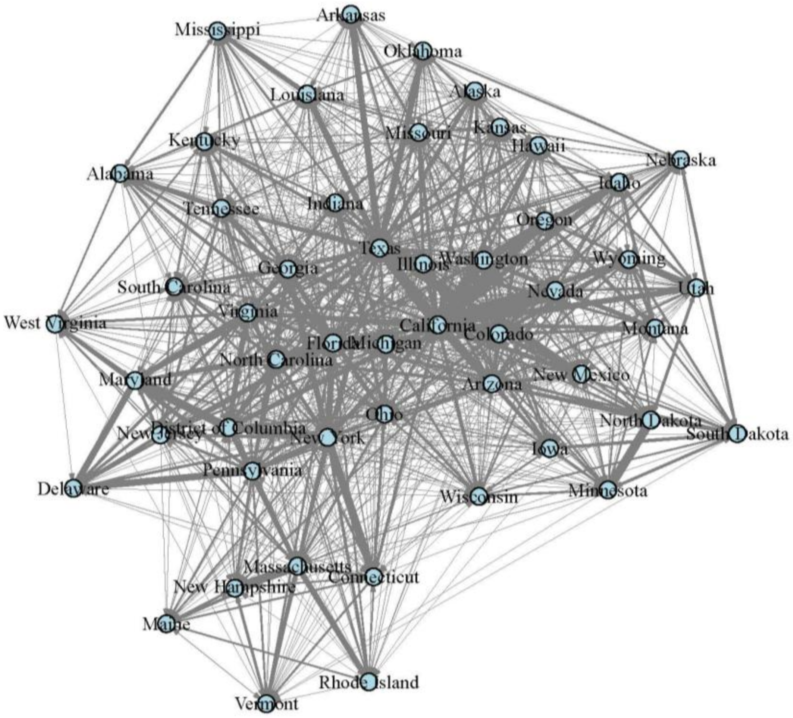
1. The data

The migration data used for this data set came from the 2015 ASEC data, acquired through the IPUMS. The key variables of interest are the following two: One is the state that the person was in currently (in 2015), and the other one is the state that the person was in 5 years ago(in 2010). Only the subset of individuals who were in a different state 5 years ago compared with current location was kept for analysis. Based on these two variables from the original dataset, we would be able to calculate the following 3 key outcome variables in our dataset:

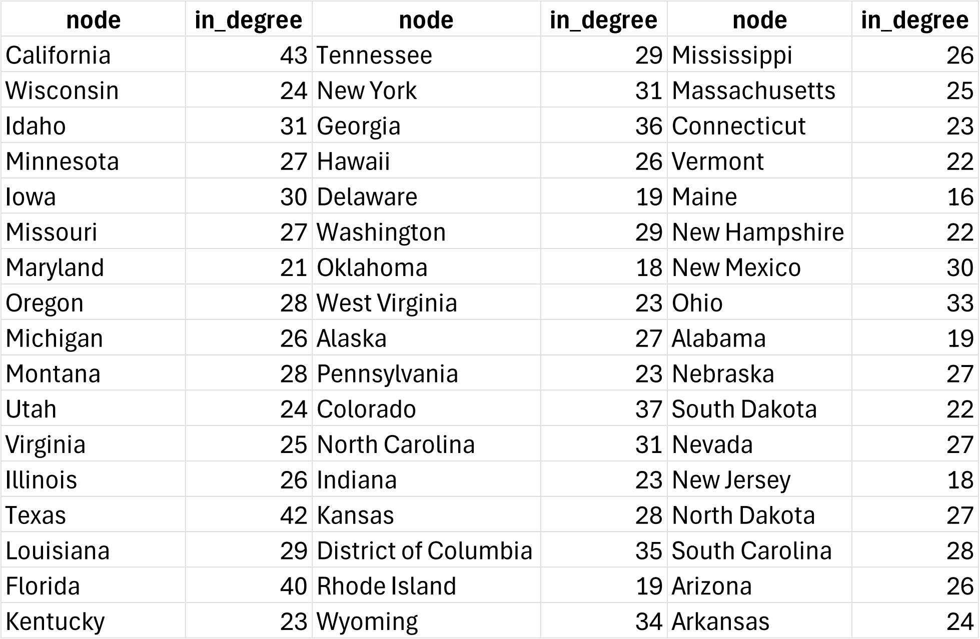
1. State where the individual migrated out: the state that the individual was in 5 years ago;
2. State where the individual migrated in: the state that the individual was now at;
3. Total number of immigrants (from state A to state B): the number of individuals migrating from state A to state B within the last 5 years.

It is important to note that the number of migrants here only considers the status at the starting and ending time point. It is true that there is a situation where a person migrates several times in five years, but in this analysis, for simplicity reasons on the one hand, and because this social network analysis does not overemphasize causality on the other hand, it does not consider complex situations. For the same reason, betweenness and centrality are not crucial in this analysis because there is no "chain" of relationships. If state A is connected to state B, and state B is connected to state C, this is uncorrelated with whether state A is connected to state C, because all the connection only indicated the comparison between the beginning and the end, and ignoring the potential intermediate steps.

2. Visualization of the Network

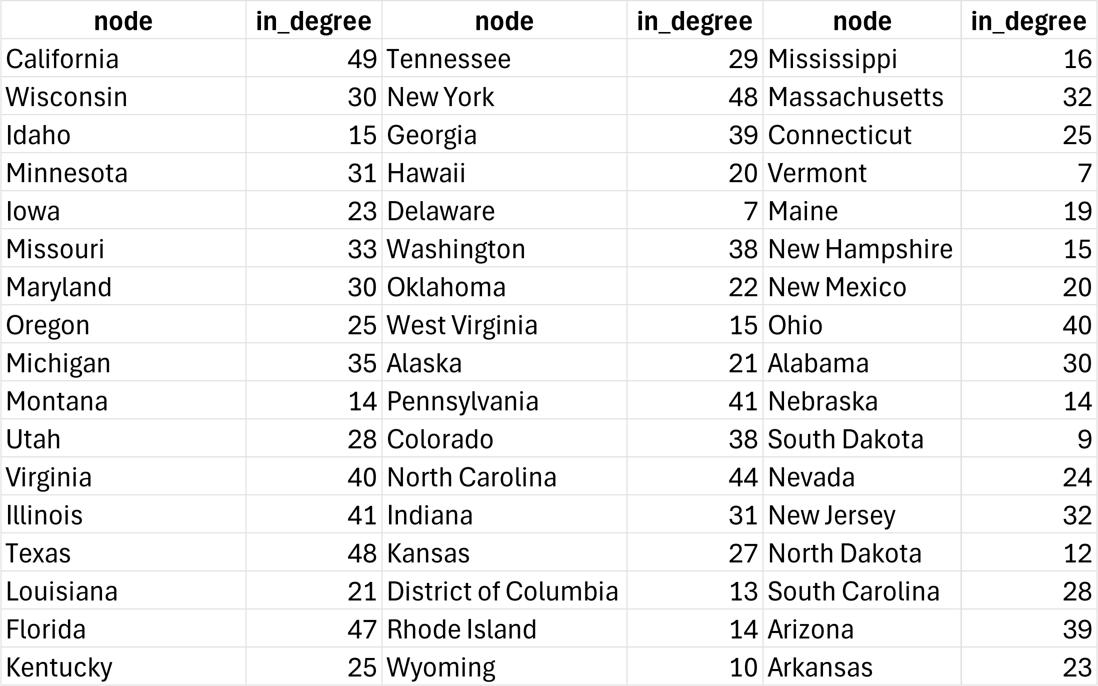


This is a basic network visualization of the immigration flows between 2010 and 2015 in the 51 states (including DC) in the United States. There are a total of 1377 edges (directed edges) in this graph, which makes the visualization too crowded. Therefore, we would look at some other characters of this network, and a migration heatmap generated with the same data.



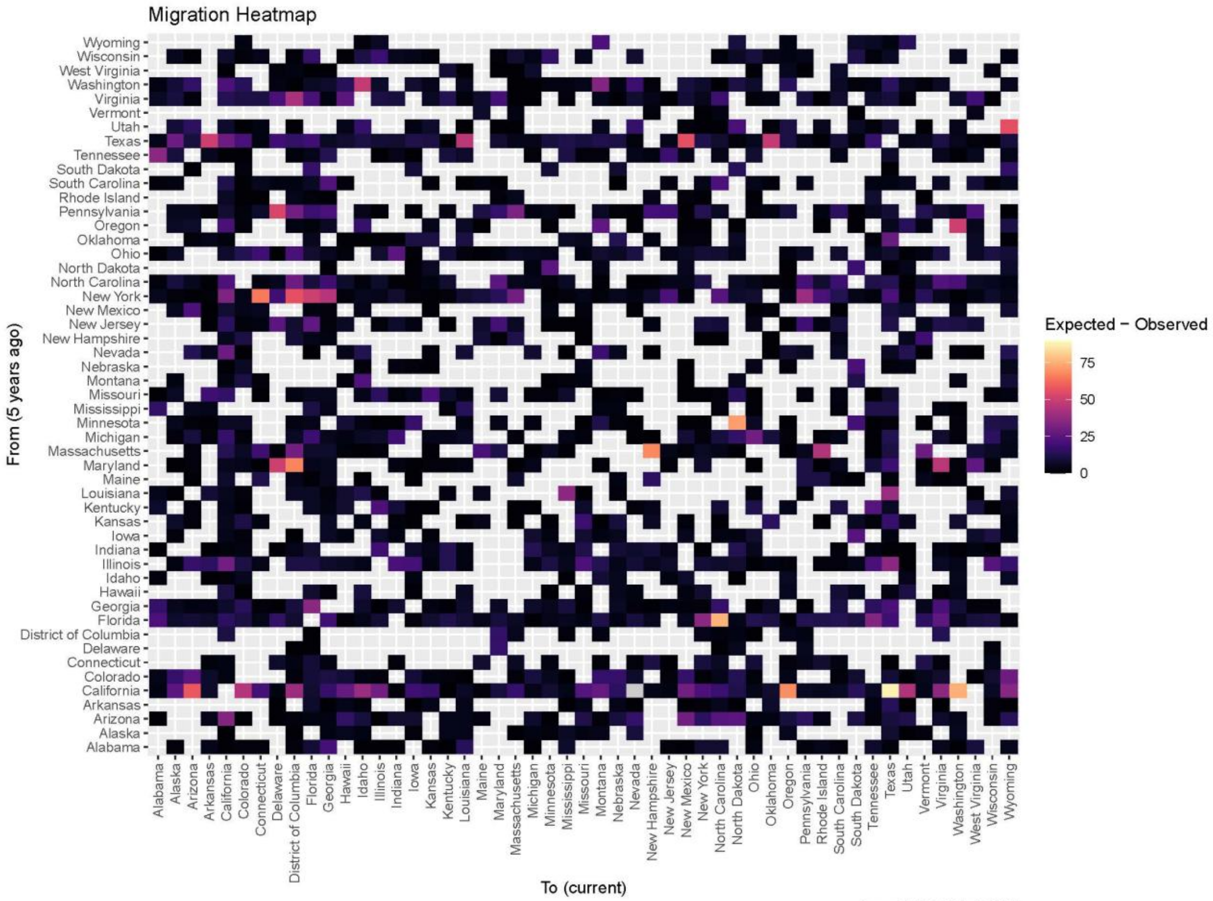
node	in_degree	node	in_degree	node	in_degree
California	43	Tennessee	29	Mississippi	26
Wisconsin	24	New York	31	Massachusetts	25
Idaho	31	Georgia	36	Connecticut	23
Minnesota	27	Hawaii	26	Vermont	22
Iowa	30	Delaware	19	Maine	16
Missouri	27	Washington	29	New Hampshire	22
Maryland	21	Oklahoma	18	New Mexico	30
Oregon	28	West Virginia	23	Ohio	33
Michigan	26	Alaska	27	Alabama	19
Montana	28	Pennsylvania	23	Nebraska	27
Utah	24	Colorado	37	South Dakota	22
Virginia	25	North Carolina	31	Nevada	27
Illinois	26	Indiana	23	New Jersey	18
Texas	42	Kansas	28	North Dakota	27
Louisiana	29	District of Columbia	35	South Carolina	28
Florida	40	Rhode Island	19	Arizona	26
Kentucky	23	Wyoming	34	Arkansas	24

This table contains the edges directed in for each node, which is that for the people migrating into one state, how many states did they come from. We found that generally speaking, each states has no less than 20 in-directed edges from other states, and most of then are between 20 and 30. California has the largest value, 46, which means that in the past 5 years, it faced immigrants from almost every other states. New Jersey, on the other hand, had the smallest value, 18, which means that the people migrating into NJ were limited to certain states.



node	in_degree	node	in_degree	node	in_degree
California	49	Tennessee	29	Mississippi	16
Wisconsin	30	New York	48	Massachusetts	32
Idaho	15	Georgia	39	Connecticut	25
Minnesota	31	Hawaii	20	Vermont	7
Iowa	23	Delaware	7	Maine	19
Missouri	33	Washington	38	New Hampshire	15
Maryland	30	Oklahoma	22	New Mexico	20
Oregon	25	West Virginia	15	Ohio	40
Michigan	35	Alaska	21	Alabama	30
Montana	14	Pennsylvania	41	Nebraska	14
Utah	28	Colorado	38	South Dakota	9
Virginia	40	North Carolina	44	Nevada	24
Illinois	41	Indiana	31	New Jersey	32
Texas	48	Kansas	27	North Dakota	12
Louisiana	21	District of Columbia	13	South Carolina	28
Florida	47	Rhode Island	14	Arizona	39
Kentucky	25	Wyoming	10	Arkansas	23

This table contains the edges directed out for each node, which is that for the people migrating out of one state, how many states became their destination. We found that the number of out-dire ted edges for each states clearly has a larger variance, compared with the in-directed edges. California still has the largest value, 49, which means that in the past 5 years people migrated out of CA came to all expect one state in the United States. Delaware and Vermont have the smallest value, which indicates that the people coming out of these 2 states had a very limited lists of destinations. It is also worth noting that, for the 2 tables for in-directed edges and out-directed edges, all the 51 states (including DC) were covered, which indicated that all the states in the United States faced immigration in both directions in the past.

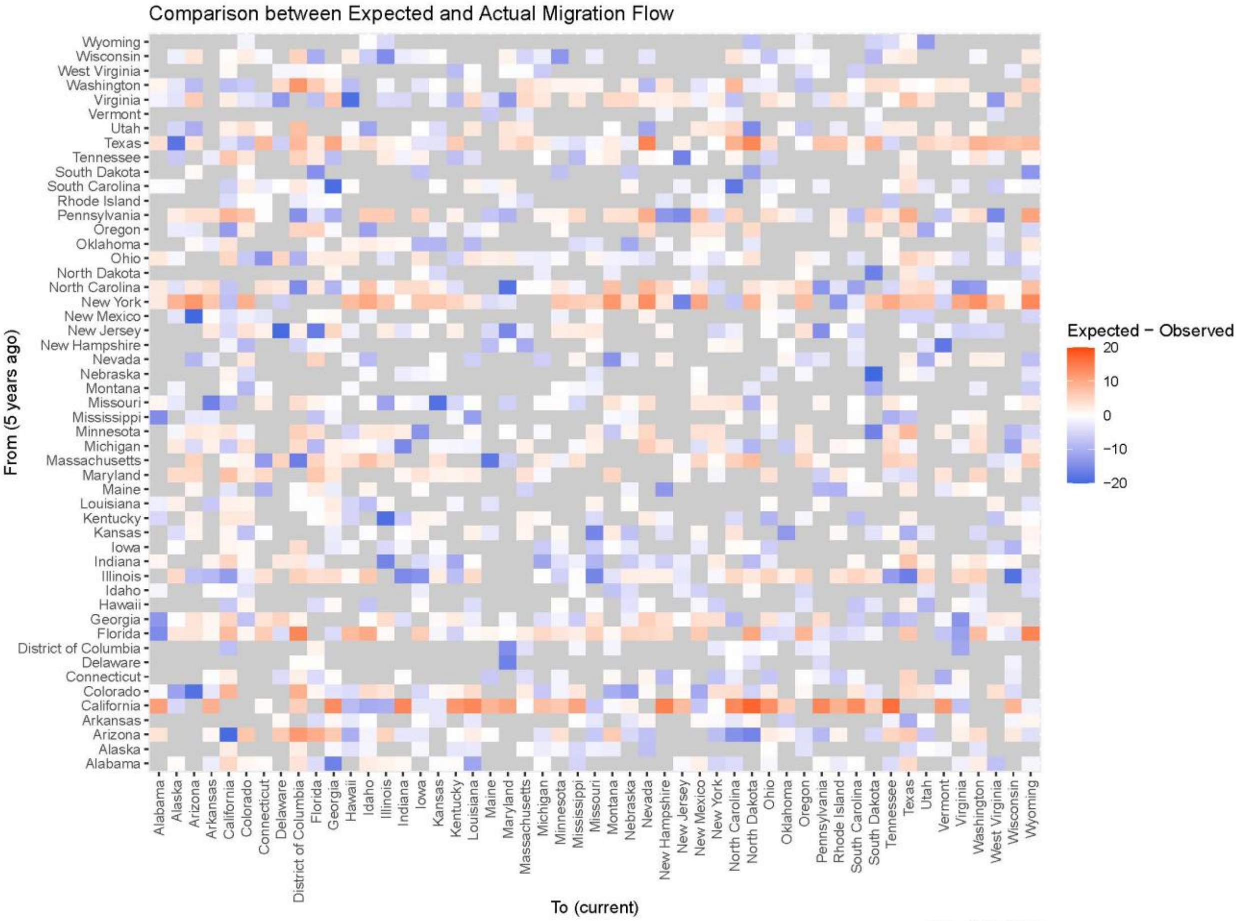


Data: 2015 ASEC, IPUMS
*States Combinations with no immigrants were left blank

The heat map provided us with a cleared view of the migration flow between the 51 states (including DC) from 2010 to 2015. Each block indicates the number of immigrants from the state on Y axis to the state on the X axis. The darker (closer to black) the block is, the smaller the number of immigrants it represents, and all the states combination with no immigration were left white (blank) in their cells.

From the heatmap, we could see that a very large percent of the immigration flow between different states is rather small, with no more than 15 individuals. The noticeable large immigration flow here would be people migrating from California to Texas, and other comparatively large immigration flow includes from Florida to North California, from California to Washington, and from Minnesota to North Dakota. For the out-migration from a state-level, California had a noticeable large amount of out-migration flow, and New York also had many people migrating out. From the macro in-migration level, however, there is no certain state receiving a noticeable large amount of people migrating in.

3. Chi Square Test



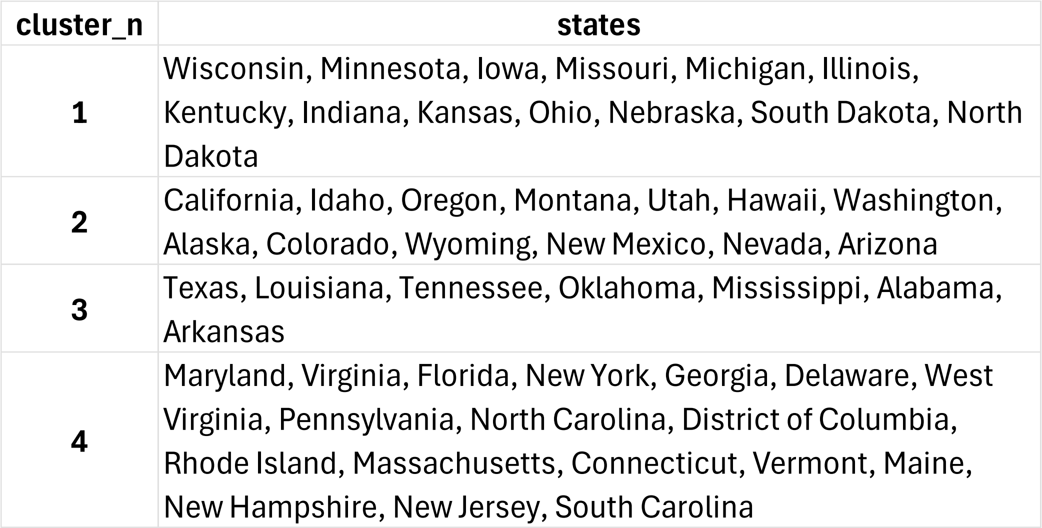
Data: 2015 ASEC
* Grey indicates no observation of immigration between x and y state

We then conducted a chi square test to found out that which combination of starting and destination has the more-than-expected immigration flow. All the combinations with no immigration flow in the original data were left grey, the combinations with more expected immigration than observed actual data were marked orange, and the combinations with less expected immigration than observed were marked blue.

We found that the result quite interesting, comparing with the general migration heatmap. Despite that facing already a large number of out-migrants, California and New York have many orange cells as the starting (out-migration state), which had two possible explanations: One is that the large number of out-migration from these 2 states were contributed by other factors, for instance, the large population, instead of the lack of attractiveness. The second one is that maybe there would be a different distribution of the people flowing into those destinations states compared with the expected one.

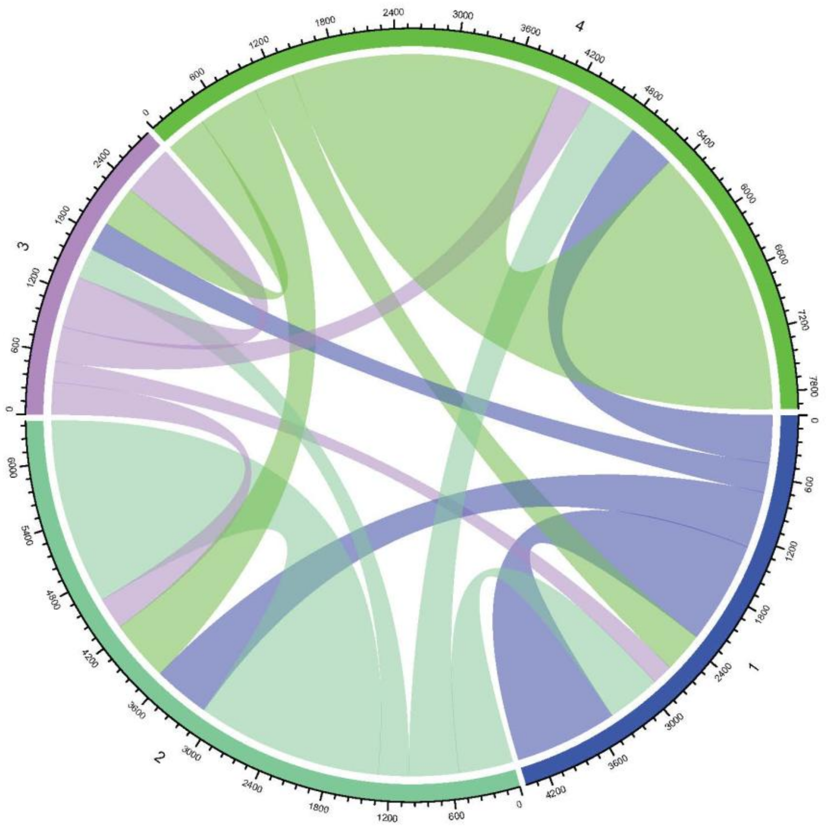
We also found that despite the blue cells are distributed quite widely in this heatmap, there seems to be no specific vertical line with many blue cells. This means that despite a lot of combinations of starting-destination states had more immigrations than the expected value from chi-square test, there is no specific state that received an unexpectedly high flow of immigration from all the other states.

4. Cluster (CNM algorithm)



cluster_n	states
1	Wisconsin, Minnesota, Iowa, Missouri, Michigan, Illinois, Kentucky, Indiana, Kansas, Ohio, Nebraska, South Dakota, North Dakota
2	California, Idaho, Oregon, Montana, Utah, Hawaii, Washington, Alaska, Colorado, Wyoming, New Mexico, Nevada, Arizona
3	Texas, Louisiana, Tennessee, Oklahoma, Mississippi, Alabama, Arkansas
4	Maryland, Virginia, Florida, New York, Georgia, Delaware, West Virginia, Pennsylvania, North Carolina, District of Columbia, Rhode Island, Massachusetts, Connecticut, Vermont, Maine, New Hampshire, New Jersey, South Carolina

We used the Clauset–Newman–Moore algorithm to generate a cluster of states in this migration flow network. The results were listed in the table. There are 4 clusters in our network, which indicated which states are more closely connected with each other.



We also visualize the clustering and the migration flow with a chord diagram. [1]

5. Predictive model

As stated in the previous section, this analysis does not emphasize causality analysis, so we will use a predictive model to detect which variables are important for THIS network of immigration flow. The predicators take account here could be divided into the following 2 categories:

Distance: the distance between states is a practical concern on migration

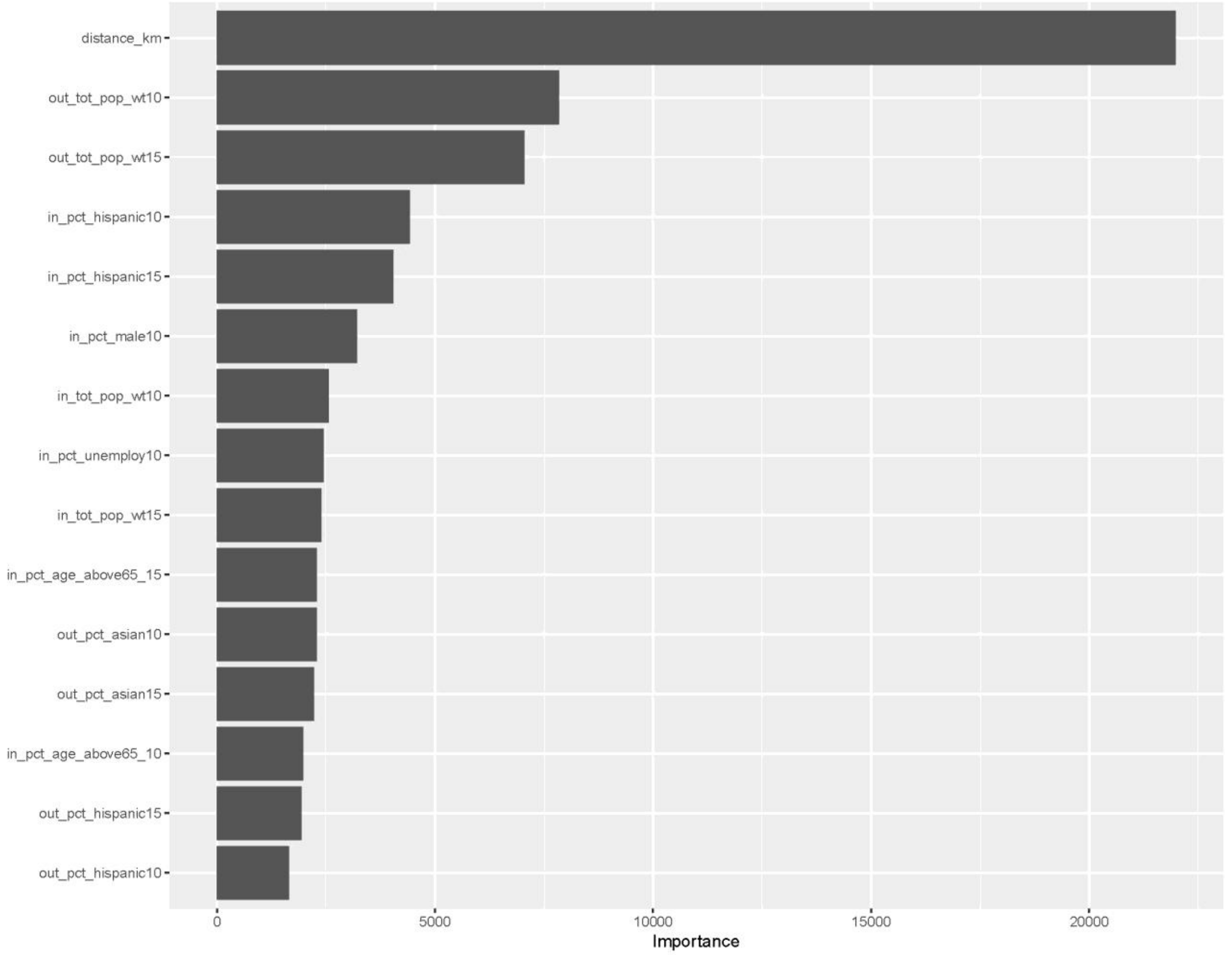
Demographic and Socia-Economic factors: The demographic and socio-economic factors, such as total population, income, races, employment, both at the beginning and ending time point (2010 and 2015), and both for the out-migration state and the in-migration state, would

[1] Blue represents cluster 1, darker green represents cluster 2, purple represents cluster 3, and the lighter green represents cluster 4.

potentially affect people’s choice on migration. Therefore, a same set of demographic and SES factors would be repeated 4 times in our predictive model for the 4 categories respectively.

The data for state distances were generated from the Tigris and sf packages in R. The demographic and SES factors for states were acquired from 2010 and 2015 ACS (1-year estimates). For the predictive model, the outcome variable is the number of migrants for a given combination of the start and destination state. The algorithm used would be random forest and Poisson regression with regularization. We would look at the top 15 predictors with the largest predicting power from the 2 algorithms, and the remaining predicators after the regulation.

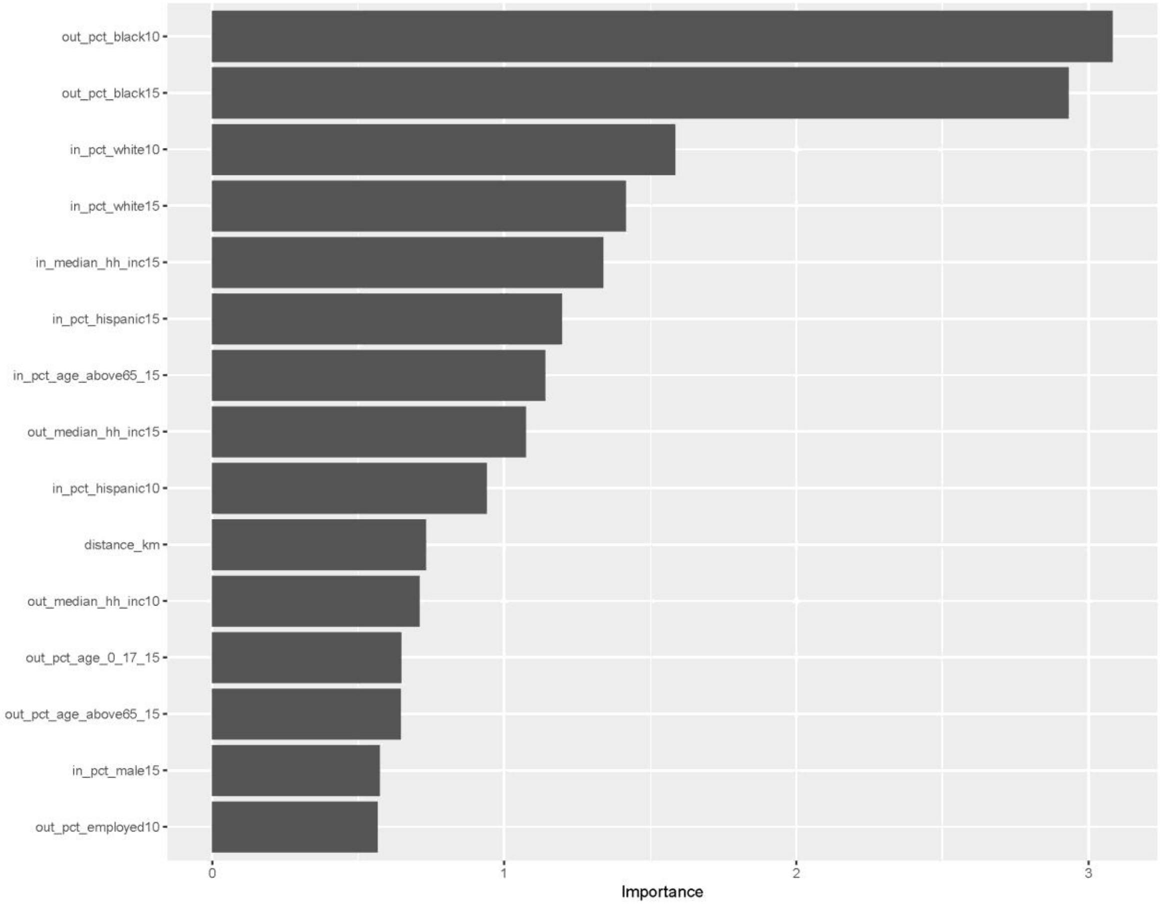
5.1.Result from random forest model



For the random forest model, distance has the largest predicting power, and is higher than all the other predictors. This suggest that distance largely affects people’s immigration selections. Another important factor for the number of immigrants in this immigration network here is the total population, which actually aligns with our previous guess that states like CA and NY faced a large number of out-migration not because the lack of attractiveness at the state level, but because these states had a large population. Race is another important predictor, as 2 different subcategories, the percentage of Asian and percentage of Hispanic, appeared here. And the only SES factors here is the unemployment rate for "in-migrating" state in 2010.

We found that there are more "out" predictors, describing factors of the starting states (out-migration), compared with the "in" predictors. This might indicate that the environment of where people come from would tell more about the migration flow. We also found that for the same demographic and SES variable, it would appear repetitively for the in and out state, start and end time point. This might say the variables important for the immigration network would be constantly important for all the states at all time.

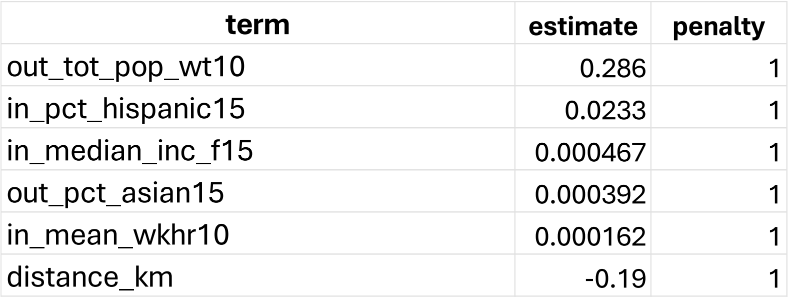
5.2.Result from Poisson regression with regularization



The result from Poisson regression has a slightly different result. Distance is still among the top 15 important predictors here, but not so important as in the random forest. The total population, however, did not even appear here. Race remains to be predicators with large predicting power. There are also more “out” factors, describing the character of the states they migrated out, compared with the “in” factors.

The result from Poisson regression has notably more socio-economic factors, for instance, the employment rate, median household income. Age also seems to be mentioned more here.

5.3.Selected Variables after regularization:



term	estimate	penalty
out_tot_pop_wt10	0.286	1
in_pct_hispanic15	0.0233	1
in_median_inc_f15	0.000467	1
out_pct_asian15	0.000392	1
in_mean_wkhr10	0.000162	1
distance_km	-0.19	1

The table is a list of variables after the selection of Poisson Regression with regularization, which deleted a bunch of variables highly correlated with each other. It is not surprising to find out that total population large contributed to predicting the migration flow. Races seem to be a important predictor, as there are 2 variables related to racial factors at state level appearing here. Income is the only SES factor left here important for the prediction of migration. Considering the value for weekly average working hours, the effect of working hours could somewhat be ignored. All the previous factors have a positive coefficient, indicating that the larger they are, the more immigrants we would expect to see within the state combination. Distance is the only predictor with a negative coefficient, yet the absolute value is rather large, indicating that even a slight increase in distance might lead to notable decline in the expected migration flow.

5.4.Conclusion

In conclusion, we would list the following variables as important factors for the immigration network in the U.S.: (1) The distance between states; (2) The percentage of non-white races; (2) The total population; (4) Factors about employment.