# Migration Analysis using Gravity Model

Linear, Poisson and Machine Learning Approches

Sudhakar BRAR
Université de Bourgogne

September 20, 2021

# Contents

# 1 Acknowledgement

This paper would not have been possible without the immense support, knowledge and enthusiasm of my supervisor Prof. Katarzyna KOPCZEWSKA, University of Warsaw. I am greatly thankful to her for her help and guidance which steered me towards pursuing this topic. Thank you.

# 2 Host Institution

The University of Warsaw was founded in 1816. It is the largest university in Poland and the best research centre in the country. As of 2017, it has 44,400 undergraduates, 3,000 postgraduates, 3,200 doctoral students and 7,250 administrative staff.

The university consists of 126 buildings and educational complexes with over 24 faculties: biology, chemistry, journalism and political science, philosophy and sociology, physics, geography and regional studies, geology, history, applied linguistics and Slavic philology, economics, philology, pedagogy, Polish language, law and public administration, psychology, applied social sciences, management and mathematics, computer science and mechanics, etc.

When it was established in the early 1800s, the University of Warsaw had 5 faculties. Today, this figure has increased to 24. As of 1st September 2020, five new faculties started their operation: Faculty of Archaeology, Faculty of History, Faculty of Sociology, Faculty of Philosophy, and Faculty of Culture and Arts. The university was also ranked by Perspektywy magazine as best Polish university in 2010, 2011, 2014 and 2016.

In addition to the 21 faculties, the University of Warsaw has over 30 academic units. Their diversity reflects the University's multiple internal aspects and the many functions it performs, while also showing UW's wide range of research and study areas.

The main campus of the University of Warsaw is in the city center, adjacent to the Krakowskie Przedmieście street. It comprises several historic palaces, most of which had been nationalized in the 19th century.

The university has some distinguished alumni one of which is Frédéric Chopin (pianist, composer). The university also has 6 nobel laureates :-

- Menachem Begin, 6th Prime Minister of Israel (1977–1983), Nobel Peace Prize winner (1978)
- Leonid Hurwicz, economist, mathematician, Nobel Prize in Economics winner (2007)
- Józef Rotblat, physicist, Nobel Peace Prize winner (1995)
- Olga Tokarczuk, writer, essayist, psychologist, Nobel Prize in Literature winner (2018)
- Henryk Sienkiewicz, writer, Nobel Prize in Literature (1905)

- Czesław Miłosz, poet, prose writer, Nobel Prize in Literature (1980)

The scholars participate in 1400 projects financed by national or international research programmes, such as EU framework programmes, European Science Foundation, European Cooperation in the Field of Scientific and Technical Research, European Economic Area and Norway Grants, European Molecular Biology Organization.

## 2.1 Faculty of Economic Sciences

The Faculty of Economic Sciences at the University of Warsaw was established in 1953. Originally named Faculty of Political Economy, the Faculty underwent a name change in 1977, and ever since it has kept the name of Faculty of Economic Sciences. Currently, the Faculty has 12 Departments.

The teaching of economics at the University has been present from the beginning of its existence, i.e. since 1816 years. At that moment when the Main School was founded, economics was taught on the faculty of law. After the November Uprising, in 1831 the University of Warsaw was closed. When the Main School was re-established in 1862, and then, in the Imperial University of Warsaw (1869) teaching of economics also took place on the faculty of law. Similar situation occurred in 1915 in reactivated by the Germans University of Warsaw, during independent Poland in the years 1918-1939, and after World War II. In 1950, in its place a Faculty of Political Economy was appointed.

The Faculty hosts five research centers: the Center for Economic Analyses of Public Sector; the Centre of Labour Market Research; the Quantitative Finance Research Group; the Warsaw Ecological Economics Centre; and the Experimental Economics Lab. The Faculty has a hundred members, including seven Presidential Professors, and seven University of Warsaw Professors.

The Faculty runs more than sixty national and international research projects, funded by, among others, the European Union, the National Centre for Research and Development, the National Science Centre, and the Ministry of Science and Higher Education. The projects range from behavioral economics to macroeconomic modeling. Faculty members consistently produce quality research output, much of it published in prestigious international journals and leading Polish journals.

A recent evaluation of scientific institutes in Poland by the Ministry of Science and Higher Education has once again attested to the Faculty's excellence: it is one of only two among 88 research institutes in the field of economics in Poland that was awarded the top evaluation of A+.

During the existence of the Faculty, the master's degree in economics was awarded to more than 5 000 people, 347 doctors and 103 associate professors have been promoted. Today, the Faculty is a leading academic centre in the field of economic sciences in Poland. The faculty currently has 121 researchers, 54 Ph.D. candidates and around 80 administrative staff.

# 3 Introduction

Spatial Interaction is a dynamic flow process from one place to another. It is general concept for any movement over space. It can be migration, journey from home to work and back, information and commodity flows, etc. To capture these spatial interactions, often times **Gravity Models** are used. These models theoretically suggest the interaction between two places decreases with the distance between them.

These models are used to predict the trade flows or to predict the degree of migration between two places. The main focus of this paper is to study empirically the inter state migration of US states, to understand the interactions between the states and also to ascertain which kinds of models will fit better for these kinds of these models.

In this paper, we will first use linear regression of logarithmically transformed variables. However, as we will see further, this leads to certain problems in the models, especially where migration is concerned.

These problems can be corrected by assuming a Poisson distribution for the independent variable, and hence, we will use the poisson regression. Since, the data is panel, hence, we will be using the panel data econometrics to estimate the linear and poisson regression estimates.

We will also use some machine learning methods to see if the degree of predictability can be increased using these methods. With the help of these mehods, we will also see the which dependent variables are important in explaining the data well and the effect or the nature of realtionship of a dependent variable with migration.

In the end we will talk about the problems of using machine learning methods for panel data and how some of these problems can be solved in the future.

# 4 Literature Review

For some decades, the gravity model has become the principle model understanding the gross migration flows between regions. This is mainly due to its intuitive consistency with migration theories, ease of estimation and a good goodness of fit in most applications (Poot (2016)).

It has demonstrated an excellent empirical robustness despite its lack of theoretical background (Anderson and Wincoop (2003)). The use of these models as a tool to analyse international migration flows has increased substantially. But so has the use of these models to analyse internal migration increased since early 2000. John Hicks states, "*differences in net economic advantages, chiefly differences in wages, are the main causes of migration*" (1932).

Keeping this statement in mind several authors have tried to analyse internal migration on the basis of some economic variables. Borjas (2000) considered age (since younger people tend to migrate more), education (as highly educated people

are able to assess employment opportunities in a better way, thus, reducing their migration costs), distance (longer the distance, the lower the incentive to migrate), unemployment (unemployed are more likely to migrate), wage differentials (higher wages do attract more people to migrate).

Other authors like Ivan Etzo (2008), studied the determinants of interregional migration in Italy. They used population size, distance between main cities, GDP per capita, unemployment rate, infrastructure index and crime rates as indepent regions to explain migration. Ignazio (1990) also studied the internal migration in Italy (1970-2005). His findings confirmed that macroeconomic variables like Per Capita GDP, Unemployment Rate, Population were the main drivers of migration flow. He also studied the impact of human capital on migration. He found that although human capital has no role at the destination, at origin it woks as a restraining factor.

Anjomani (2002) carried out an earlier analysis of US interstate migration and used variables like regional income, employment rate, population density, temperature, welfare benefit, ciminality rates, population size, mean educational level, median population age as regressors. Bunea (2012), did a panel data analysis of Romanaia and found that population size, real GDP, amenity index, road density and crime rate had significant impact.

Pietrzak, Wilk and Matusik (2013) studied the internal migration of Poland (2004-2010). They also used key economic factors such as GDP, capital expenditures, unemployment, wages and salaries as factors which can explain internal migration. Their main finding was that the regions which had relatively good economic situation are the centres of population inflow.

Gökhan (2008) presnts an empirical study of determinants of internal migration in Turkey (1990-2000). They show that both economic factors (income differentials, unemployment rates, etc), social factors (presence of social netwoks) and personal characteristics (age, education level) have a significant impact on migration.

# 5 Data

The datasets used in this study comes from two different sources. First, from **American Community Survey**[1] (a demographics survey program conducted by US Census Brureau[2]), we have migration flows from one state to another. These states-to-state migration flows are from the year 2011 till 2019 [3]. Migration flows from and to **Alaska**, **Hawaii** and any **foreign country** have not been taken into account. Hence, we have a panel data of migration of 49 states to 48 states for 9 years. Also, people, who have migrated from one place in a state to another place in the same state, have also not been taken into account. A `shp` file from US Census

[1] `https://www.census.gov/programs-surveys/acs/`

[2] `https://www.census.gov/`

[3] `https://www.census.gov/data/tables/time-series/demo/geographic-mobility/state-to-state-migration.html`

Bureau was also used to plot the states of US[4].

The second source of data is **Bureau of Economic Analysis**[5] of US Department of Commerce[6]. There are two different datasets taken from this data source. First is **Personal Income by State**[7] and second is **Employment by State**[8].

In the Table 1, we can see that the state of Vermont has the lowest *In-Migration* (in most of the years) (Figure 3, VT) & *Out-Migration* (Figure 4, VT) and New York has the minimum *Net-Migration* (Figure 5, NY). The Table 2 we can see that Florida has the maximum *In-Migration* (Figure 3, FL) , California has the maximum *Out-Migration* (Figure 4, CA) and Florida has the maximum *Net-Migration* (Figure 5, FL).

Table 1: Minimum Migration States (2011-2019)

| Year | In-Migration | Out-Migration | Net-Migration |
|---|---|---|---|
| 2011 | Vermont (19740) | Vermont (18104) | California ( -97866) |
| 2012 | Vermont (24332) | Vermont (19977) | New York (-138293) |
| 2013 | Vermont (23079) | Vermont (21767) | New York (-133517) |
| 2014 | Vermont (22410) | Vermont (20309) | New York (-167817) |
| 2015 | Vermont (21215) | Vermont (22082) | New York (-189527) |
| 2016 | Vermont (21221) | Vermont (23524) | New York (-188804) |
| 2017 | Wyoming (18359) | Vermont (20872) | New York (-168556) |
| 2018 | Wyoming (25679) | Vermont (20827) | New York (-203367) |
| 2019 | Vermont (20808) | Vermont (22605) | New York (-184530) |

In the Figure 3, we can see that California (CA), Florida (FL) and Texas (TX) have very high In-Migration. In Figure 4, we can see that California (CA), Florida (FL) New York (NY) and Texas (TX) have very high Out-Migration. In Figure 6, we can see the In-Migration, Out-Migration and Net-Migration for all the states of US for the years 2011-2019. But in this figure all graphs have their own scale. Hence, we won't be able to look at them in a relative way. Hence, when the scale for all the graphs is the same, we get Figure 7. In this Figure we can see that for most states there is not a lot of difference between In-Migration and Out-Migration. This can also be seen in the Figure 5, since, most of the bars are quite small.

Table 2: Maximum Migration States (2011-2019)

[4] https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html

[5] https://www.bea.gov/

[6] https://www.commerce.gov/

[7] https://www.bea.gov/data/income-saving/personal-income-by-state

[8] https://www.bea.gov/data/employment/employment-by-state

| Year | In-Migration | Out-Migration | Net-Migration |
|---|---|---|---|
| 2011 | Texas (501950) | California (549072) | Texas (102610) |
| 2012 | Florida (523267) | California (553736) | Florida ( 99713) |
| 2013 | Texas (535088) | California (565904) | Texas (134973) |
| 2014 | Florida (536438) | California (577703) | Florida (104838) |
| 2015 | Florida (580407) | California (629696) | Florida (138573) |
| 2016 | Florida (601840) | California (641050) | Florida (171601) |
| 2017 | Florida (562984) | California (648627) | Florida (118764) |
| 2018 | Florida (582865) | California (677234) | Florida (116145) |
| 2019 | Florida (592602) | California (640050) | Florida (139632) |

In the Figure 8, we can see that the most states have the Employment-to-Population ratio around 0.5 to 0.7. However, District of Columbia seems to have it around 1.3 for most of the years. Most states also seem to be having a rising Employment-to-Population ratio as can be seen in Figure 9. However, for District of Columbia and some more states, it seems to be falling.

In the Table 3, we can see the states which have the highest number of people employed, regardless the size of the population of those states. California is the state which employs maximum number of people from 2011 till 2019 even though from Figure 5 we can see that it is a state with a lot of negative net migration. But from Figure 9, we can see that the Employment to Population ration seems to keep increasing in California.

Table 3: Maximum Employment States (2011-2019)

| Year | State | People Employed |
|---|---|---|
| 2011 | California | 19986021 |
| 2012 | California | 20666908 |
| 2013 | California | 21319995 |
| 2014 | California | 21997098 |
| 2015 | California | 22687196 |
| 2016 | California | 23164907 |
| 2017 | California | 23538609 |
| 2018 | California | 24086110 |
| 2019 | California | 24602061 |

Similarly, from Table 4, we can see that Wyoming is the state with least amout of people employed from 2011 till 2019. We can see that it has very less In-Migration, Out-Migration and Net-Migration from Figures 3, 4 and 5.

Table 4: Minimum Employment States (2011-2019)

| Year | State | People Employed |
|---|---|---|
| 2011 | Wyoming | 390303 |
| 2012 | Wyoming | 396704 |
| 2013 | Wyoming | 400402 |
| 2014 | Wyoming | 406343 |
| 2015 | Wyoming | 406033 |
| 2016 | Wyoming | 398271 |
| 2017 | Wyoming | 398915 |
| 2018 | Wyoming | 404465 |
| 2019 | Wyoming | 412765 |

# 6 Gravity Model

The formulation of **Gravity Model** is attributer to Henry Carey (1858). However, this model was popularized by Stewart (1948). According to him, the force of interaction ($F_{ij}$) exerted between two places, $i$ and $j$, is proportional to their masses and inversely proportional to the square of distance between them, where population is the mass of a place.

$$M_{ij} = G\frac{P_i P_j}{d_{ij}^2}$$

where,
$P_i$ is the population in place $i$
$d_{ij}$ is the distance between the places $i$ and $j$.

Several authors have already established that movement of person is inversely related to distance and not to distance squared (see Israd (1960), Zipf (1946)). Also, the relationship of interaction of population might not be linear, hence, a more general model was created. Since, our data is a panel data, we will also have the subscript $t$ which signifies the year. Hence, we have the following model:-

$$M_{ijt} = G\frac{P_{it}^{\alpha} P_{jt}^{\beta}}{d_{ij}^{\gamma}} Q_{ijt}^{\delta}$$

where,
$Q_{ijt}$ is the vector of other independent variables.

This model is quite popular, since, the parameters of this model are quite easy to estimate by ordinary least squares if a lograthmic transformation is applied on the model equation. Doing so gives us the following equation:-

$$ln(M_{ijt}) = ln(G) + \alpha ln(P_{it}) + \beta ln(P_{jt}) - \gamma ln(d_{ij}) + \delta ln(Q_{ijt}) + \varepsilon_{ijt} \qquad (1)$$

where $\varepsilon_{ijt}$s are the normally distributed zero-mean errors with variance equal to 0.

Although the estimates of this model are quite easy to compute, however, this model has some problems of itself. First, (see Haworth, J. M., & Vincent, P. J. (1979))

the regression estimates the logarithms of $M_{ijt}$s. But the antilogarithms of these estimates are biased estimates of $M_{ijt}$s.

A second difficulty arises due to logarithmic transformation is when the migration flow is equal to 0 for a specific $i$, $j$ and $t$. Since, logarithm of 0 cannot be computed, we won't either use these observations or we can replace these values with a small positive number (1 in our case).

Another difficulty is that $\varepsilon_{ijt}$s are assumed to be normally distributed which implies that the values taken by the dependent variable are also normally distributed. Considering our case, we know this is a bad assumption, since, the dependent variable must be nonnegative whole numbers. This is because only a nonnegative whole number of people can represent migrats.

Development in statistics have lead to development of generalised linear modelling, where the dependent variable can take a variety of different probability distribution (Nelder, Wedderburn (1972)). One special case of this kind of modelling technique is Poisson regression, which has a dependent variable with Poisson distribution, linked logarithmically to a linear combination of independent variables.

Hence, the above problems can be solved by assuming that the number of people going from place $j$ to $i$ must be nonnegative whole numbers, and hence, each $M_{ijt}$ should have its discrete probability distribution.

According to (Flowerdew, R., Aitkin, M. (1982)), if there is a small constant probability that any individual from place $j$ moves to place $i$, and if the movements of individuals are independent, then, if the population of $j$ is large, the number of individuals recorded as moving from $j$ to $i$ will have a Poisson distribution with mean $\lambda_{ijt}$. Hence, the probability that $k$ people will move is:-

$$Pr(M_{ijt} = k) = \frac{e^{\lambda_{ijt}} \lambda_{ijt}^{k}}{k!}$$

And the relationship between the $\lambda_{ijt}$ and model variables is as following:-

$$\lambda_{ijt} = exp(\beta_0 + \beta_1 P_{it} + \beta_1 P_{jt} + \beta_3 d_{ij} + \beta_4 Q_{ijt}) \tag{2}$$

The distances between the states ($d_{ij}$) have been calculated by first obtaining the centroids of each state. And then, the euclidean distances between these centroids have been calculated. One of the main problem of this kind of distances is that for some people these distances will be over estimated. And for some these will be under estimated. Also, there are many problems with euclidean distances.

## 6.1 Machine Learning Models

In this subsection we will look at two different machine learning techniques which can help us in better underrstand the model. These models will help us in finding which variables have the most effect on the generalising power of the model.

Regression Trees are statistical modelling techniques which generate a tree of decisions whose leaves are a prediction for some numeric class. These are non linear predictive structure where the dependent variable is numeric in nature. They recursively divide data on the columns whose ranges have the lowest standard deviation (or some other performance method). This technique decides "*which input features (and numeric threshold values in the case of numeric features) to make splits on, usually in a top-down and recursive way. So, first, the split at the root node is decided. Then, for each node in the next level of the tree, the split is decided in a similar way to the root node, and so on.* (Sharing Data and Models in Software Engineering (2015))". They are able to make splits also on the basis of categorical inputs. These trees can also fit the data more generally by using resampling techniques like cross-validation, leave one out cross validation, bootstraping, subsampling, etc.

On the other hand, Random forests are a learning method for regression, classification and any other task that uses multiple decision trees. This method can generally outperform Regression Trees (S. Madeh Piryonesi, Ph.D.; and Tamer E. El-Diraby, Ph.D. (2020)). This is a way of averaging many decision trees which are trained on different parts of the same training data. Their main goal is to reduce the variance (Hastie, Tibshirani, Friedman(2008)). However, this causes a small increase in bias and loss in interpretability, but it boosts the performance of the model greatly.

# 7 Results

First, we estimate the model in the equation (1) with more independent variables such as employment in the origin and in destination and per capita personal income in the origin as well as in destination. The model is estimated using **random effects**, since, if we were to use fixed effects, the effect of distance on migration would be unknown. Since, fixed effects removes all the variables which are constant over time. Although it should be noted that performing a Durbin-Hausman test to compare the fixed and the random effects model, we find that fixed effect models does have consistent estimates, and hence, should be preferred.

Table 5: Random Effects Gravity Model Estimates

| | Model 1 | Model 2 | Model 3 |
|---|---|---|---|
| (Intercept) | -10.166000*** | -8.720987*** | -10.305958*** |
| Population: Destination | -0.966104*** | -1.117406*** | -0.964518*** |
| Population: Origin | 0.041249 | 0.060633 | 0.042640 |
| Per Cap Per Inc: Destination | -1.145765*** | -1.400723*** | -1.138968*** |
| Per Cap Per Inc: Origin | -0.365832* | -0.249031** | -0.359464* |
| Distance | -1.066184*** | -1.065674*** | -1.066184*** |
| Employment: Destination | 2.207276*** | 2.365787*** | 2.205577*** |
| Employment: Origin | 1.180381*** | 1.160963*** | 1.178886*** |
| Effects | Individual | Time | Individual-Time |
| Adj. $R^2$ | 0.26159 | 0.59689 | 0.26149 |

The Table 5 has the estimates for the equation (1). All the models have the same variables. The reason why they are three different because of the effects they have. In all the models the coefficient of *Distance* is negative and is significant, hence, this tells us that the relationship between log of migration and log of distance is actually inverse.

The $R^2$ of the Model 2 which has *Time* effects is quite high in comparison to the other two models. The *Per Cap Per Inc: Destination* coefficient is negative, which is quite counter intuitive, since, if per capita personal income in the destination region increases, then, the migration towards that place will decrease. Hence, they have negative relationship.

The coefficient for *Population: Origin* is insignificant in all the three models. But the coefficient of *Population: Destination* is negative and significant, i.e., there is a negative relationship between migration to a place and the population at the same place. This negative relationship is not in congruent with the theoretical Gravity model which says that the relationship between the migration flows and population of destination and origin.

Since, we already do know that these estimates are biased, we will now estimate the results of the poisson regression as stated in the equation (2). The table 6 has these estimates. In the Poisson panel models, the coefficient of *Distance* is negative. So there is still a negative relationship between migration and distance.

Also, (in Table 6) there is no difference in the model coefficients whether they have individual, time or individual-time fixed effects. Incorporating for any of these fixed effects leads to the same estimates. Therefore, there are no rows regarding the effects.

Here (in Table 6) the coefficient of *Population: Destination* is positive and significant. The coefficient for *Population: Origin* is positive and significant only in

Model 1, which is a pure Gravity model. In the other two, they are significant but negative. But the coefficient of *Per Cap Per Inc: Origin* in Model 2 has become insignificant. Model 3 which has all the coefficients highly significant. Also, the coefficient of *Employment: Destination* has become negative.

Also, according to the Akaike Information Criterion, the Model 2 tends to perfom well since, its AIC statistic is smaller than that in Model 1. Model 3 tends to perform better than Model 2 but the change is quite insignificant in their AIC's.

In Table 7, we can see the model performance results for different Regression Tree models. The Normal column consists of the results for just a normal random forests model. The Holdout column is for a model with *holdout* resampling with a ratio of 0.8, i.e., 80% of the data was used to train the model and 20% of the data was used to test model. The Cross Validation column is for the model with *cross-validated* resampling, with 5 folds.

Table 6: Poisson Random Effects Gravity Model

| | Model 1 | Model 2 | Model 3 |
|---|---|---|---|
| (Intercept) | 7.80416928*** | 7.90088122*** | 7.90132036*** |
| Population: Destination | 0.00413270*** | 0.00752588*** | 0.00753887*** |
| Population: Origin | 0.00456502*** | -0.00141526*** | -0.00143237*** |
| Per Cap Per Inc: Destination | | 0.00331459*** | 0.00995546*** |
| Per Cap Per Inc: Origin | | 0.00004710 | |
| Distance | -0.04538806*** | -0.04119177*** | -0.4118385*** |
| Employment: Destination | | -0.00475223*** | -0.00476358*** |
| Employment: Origin | | 0.00406148*** | 0.00408024*** |
| Log-Likelihood | -2587110 | -2577647 | -2577647 |
| AIC | 5174231 | 5155313 | 5155311 |

From this table (Table 7), we can see that the Regression tree models perform quite well according to their *R-Squared*. However, while choosing the model, one should not choose the Normal model, since, it has been tested on the same data that it was trained on. So, one should choose either *Holdout* or *Cross Validation* model. In our case the model with *Holdout* resampling techniques seems to perform better.

Table 7: Regression Tree Models

| | Normal | Holdout | CV |
|---|---|---|---|
| R-Squared | 0.6639373 | 0.66467 | 0.6406088 |
| Mean Squared Error | 11607474 | 11064365 | 12415298 |
| Mean Squared Log Error | 5.265185 | 4.623989 | 4.890827 |

In Table 8, we can see the results for the Random Forest model. From this table (Table 8), we can see that all of the models tend to perform pretty, especially the Normal Model.

In this table 9 we also have *ST CV* (Spatio-Temporal Cross Validation) for both the models. We use this kind of cross validation because observation in spatio-temporal data *inherit a natural grouping, either in space (spatial autocorrelation) or time (temporal autocorrelation) or in both space and time (spatiotemporal autocorrelation)* (Legendre (1993)). To account for the spatiotemporal autocorrelation, we use the *Repeated Leave-Location-and-Time-Out* Cross Validation. This gives us a decrease in our model performance for Regression Trees and a very large decrease in performance of Random Forests. In normal cross validation, train and test observations would be located side-by-side across the full study area which leads to high similarity between train and test sets, resulting in *better* but **biased** performance estimates in every fold of a CV. These good results are due to presence of Spatiotemporal autocorrelation and not because of the power of the model.

In Figure 1, we can see the Feature Importance of each feature in the Regression Trees Models. The importance of a feature is calculated by the measuring the increase in the error of the model when the feature is permuted. A feature is **important** if permuting it increases the model error, since, in this case, the model relies on the feature. A feature is **unimportant** if permuting it leaves the model error unchanged. From the figure (Figure 1), we can see that the distance, Population of destination and the total employment in origin are the three main features which explain a lot variation in the data.

In Figure 2, we can see the Feature Importance of each feature in the Random Forests Model. We can see that the three main variables of the Gravity Equation (Euclidean Distance, Population: Destination, Population: Origin) are the variables with the highest importance. Employment in the Origin region as well as the Destination region are also quite important variables.

Table 8: Random Forest Models

| | Normal | Holdout | CV | ST CV |
|---|---|---|---|---|
| R-Squared | 0.9892 | 0.9556 | 0.9481 | 0.6564 |
| Mean Squared Error | 371192 | 1562207 | 1798210 | 11985580 |
| Mean Squared Log Error | 1.39 | 2.23 | 2.1989 | 3.3640 |

Figure 10 is a *Partial Dependence Plot* for Regression Tree model and signifies the effect of a feature (independent variable) on the dependent variable. It shows whether the relationship between them is linear, monotonic, or more complex. In Figure 10, we can see that the relationship between *Migration* and *Euclidean Distance* is negative but it more stepwise. It is not linearly negative. Similarly, the relationship with *Population: Destination* and *Total Employment: Origin* is positive, but not linearly positive. With the other variables, the realtionship is quite static.

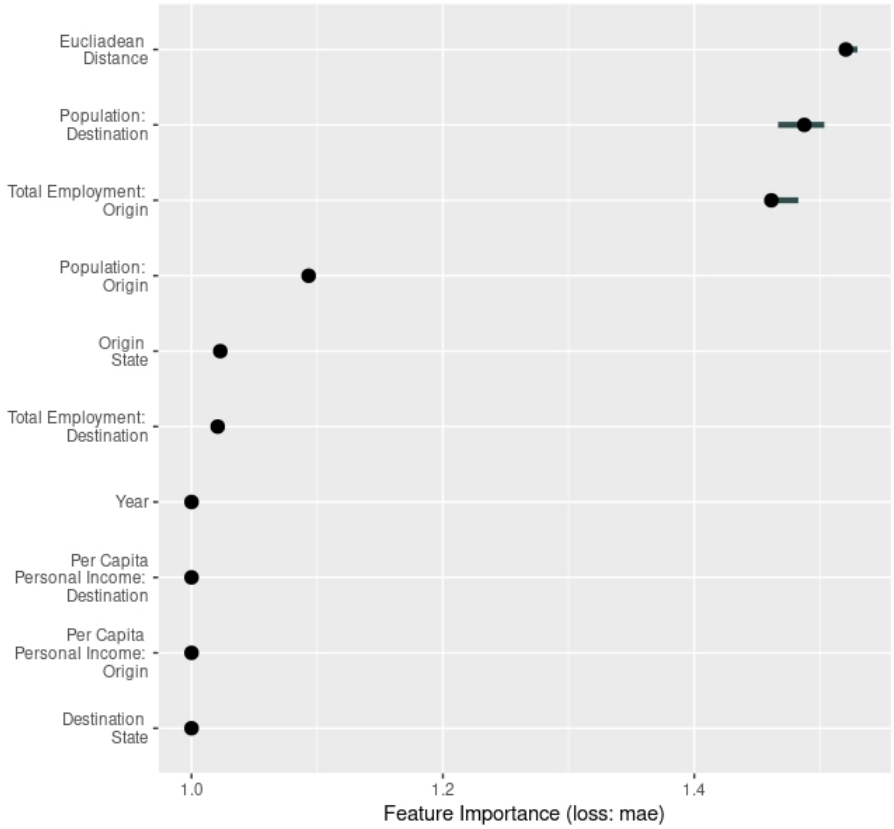


Figure 1: Feature Importance: Regression Tree Model

This can also be because as we saw in *Feature Importance* plot for Regression Trees, these variables did not have a lot of importance.

Table 9: Spatiotemporal Cross Validated Models

| | Regression Trees | Random Forests |
|---|---|---|
| R-Squared | 0.5974 | 0.6564 |
| Mean Squared Error | 14076871 | 11985580 |
| Mean Squared Log Error | 3.48542 | 3.3640 |

Figure 11 is the *Partial Dependence Plot* for the Random Forests model. It shows a much more smoother relationships between migration and independent variables than in Figure 10. In this we can see that there is a positive relationship of migration with total population of the destination and origin state. And there is a negative relationship with distance. These results were expected and are inline with the theoretical relationship of migration with these variables. However, the realtionship with the distance is quite interesting. It is not a negatively linear. It flattens out after 20 units of distance between the centroids of two regions.

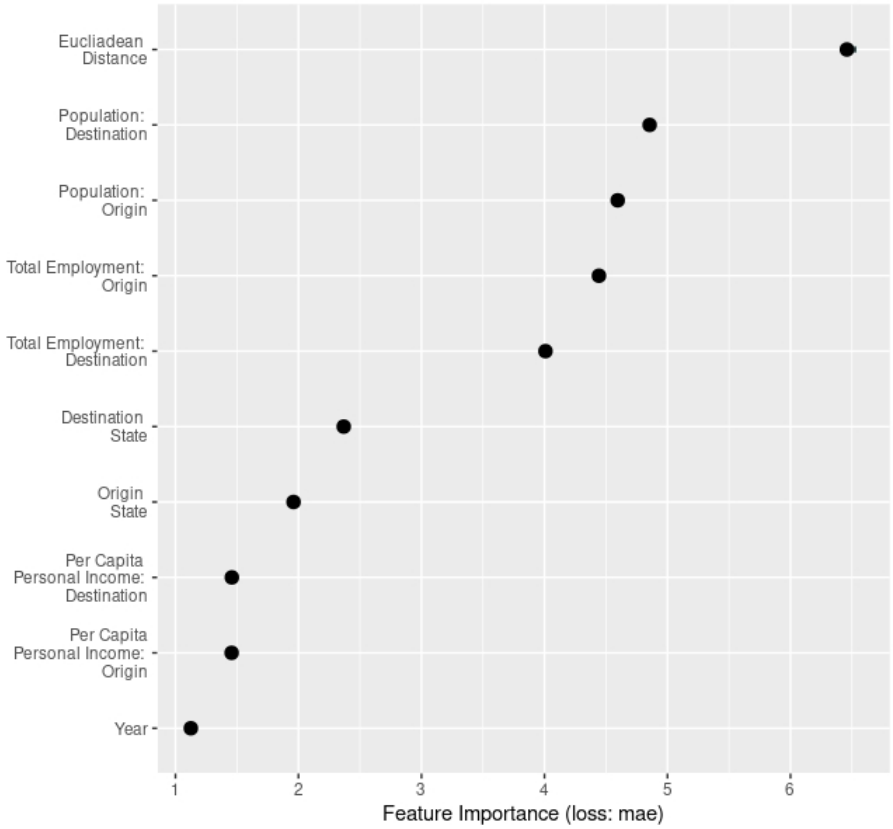


Figure 2: Feature Importance: Random Forests Model

# 8 Conclusion

From this analysis, we were able to see that although it is easy to calculate and interpret the model in equation 1, it shouldn't be done, as it produces biased coefficient estimates and other problems. Hence, we should use the Poisson regression. We also saw that the regression tree models tend to be a good fit since they have the ability to explain the variation in the data well. And that Random Forest models tend to do even better than the regression tree models with $R^2$ of more than 0.9 in the three models (Non-spatiotemporal CV). But in the Spatiotemporal CV model, our $R^2$ decreases a lot (0.65) for Random Forests and (0.59) for Regression Trees, but at least we deal with spatiotemporal autocorrelation and don't get biased estimates at every fold of a CV. We also saw that the relationship of migration with population in destination and origin states was in tune with the economic theory of Gravity models in both Regression Trees and Random Forests models.

In the normal regression model (Table 5), the coefficient of population in destination state had a sign which is not in line with the theory. However, this was corrected, at least in, Model 1 of Table 6. We also saw that distance and population in destination state is the most important variable in both Regression Trees and Random Forests models. Population in the origin state is the third important variable in Random Forests and fourth important feature in Regression Trees.

Although Regression Trees and Random Forests might have looked like good mod-

els when compared to the normal regression or to the poisson regression model. However, they do come with their own drawbacks. Machine learning methods in general do not have the ability to tackle common panel data, let alone panel data with three or more indices since they cannot remove individual heterogeneity. However, in our Regression Tree and Random Forests model, the way we have taken this into account is by putting dummy variables for destination (Destination: State), the origin (Origin: State) and year Year. Hence, we do not eliminiate like we do in normal panel models. However, there is new some research going for panel data which is coming out recently.

Droesch (2021) explains why it is difficult for machine learning models to take care of individual heterogeneity, at least in the case of *within-individual* estimator. He explains that the relationship between individual heterogeneity and the dependant variable is linear, and hence, due to Frisch–Waugh–Lovell theorem we will still get the same estimates we would estimated if we did not perform the within-individual transformation. However, in machine learning, the relationship between the individual heterogeneity is not known. Most of the time it is going to be non linear. Due to this non linear relationship the machine learning models tend to fail. The author therefore has developed extensions of neural networks which can be used on a panel data and has demonstrated "*unbiasedness of parametric estimates, good properties of estimated confidence intervals and effiiciency both in a simulated dataset and in an application to yield prediction from weather data*". These methods are also available in a R-package called **panelNNet** which is in an experimental stage right now.

# 9 References


- Anderson, James, E., and Eric van Wincoop. 2003. "Gravity with Gravitas: A Solution to the Border Puzzle ." American Economic Review, 93 (1): 170-192.
- Anjomani, Ardeshir, 2002. "Regional growth and interstate migration," Socio-Economic Planning Sciences, Elsevier, vol. 36(4), pages 239-265, December.
- Bivand, Roger S. and Wong, David W. S. (2018) Comparing implementations of global and local indicators of spatial association TEST, 27(3), 716-748. URL https://doi.org/10.1007/s11749-018-0599-x
- Borjas G. 2000. "Issues in the Economics of Immigration". National Bureau of Economic Research. University of Chicago Press. Volume URL: `http://www.nber.org/books/borj00-1`
- Breiman, Leo & Friedman, Jerome & Olshen, Richard & Stone, Charles. (2017). Classification And Regression Trees. 10.1201/9781315139470.
- BUNEA, Daniela. (2010). Modern Gravity Models of Internal Migration. The Case of Romania. Theoretical and Applied Economics. 4(569).
- Croissant Y, Millo G (2008). "Panel Data Econometrics in R: The plm Package." Journal of Statistical Software, 27(2), 1-43. doi: 10.18637/jss.v027.i02 (URL: https://doi.org/10.18637/jss.v027.i02).
- Etzo, Ivan. (2008). Internal migration: A review of the literature.
- Flowerdew R, Aitkin M. A method of fitting the gravity model based on the Poisson distribution. J Reg Sci. 1982 May;22(2):191-202. doi: 10.1111/j.1467-9787.1982.tb00744.x. PMID: 12265103.
- Greene, William H. (2012). Econometric Analysis (7th ed.). Pearson. pp. 379–380, 420.
- H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
- Hicks, J.R. (1932) The Theory of Wages. Macmillan, London.
- Isard, W. (1960), THE SCOPE AND NATURE OF REGIONAL SCIENCE. Papers in Regional Science, 6: 9-34. https://doi.org/10.1111/j.1435-5597.1960.tb01698.x
- J M Haworth & P J Vincent, 1979. "The Stochastic Disturbance Specification and its Implications for Log-Linear Regression," Environment and Planning A, , vol. 11(7), pages 781-790, July.
- Lang M, Binder M, Richter J, Schratz P, Pfisterer F, Coors S, Au Q, Casalicchio G, Kotthoff L, Bischl B (2019). "mlr3: A modern object-oriented machine learning framework in R." Journal of Open Source Software. doi: 10.21105/joss.01903 (URL:https://doi.org/10.21105/joss.01903)
- Matt Dowle and Arun Srinivasan (2020). data.table: Extension of 'data.frame'. R package version 1.13.6. https://CRAN.R-project.org/package=data.table
- Molnar C, Bischl B, Casalicchio G (2018). "iml: An R package for Interpretable Machine Learning." JOSS, 3 (26), 786. doi: 10.21105/joss.00786 (URL:https://doi.org/10.21105/joss.00786)

- Nelder, J., & Wedderburn, R. (1972). Generalized Linear Models. Journal of the Royal Statistical Society. Series A (General), 135(3), 370-384. doi:10.2307/2344614
- Odlyzko, Andrew, The Forgotten Discovery of Gravity Models and the Inefficiency of Early Railway Networks (September 1, 2014). OEconomial, vol. 5, no. 1, 2015, pp. 157-192.
- Poot, Jacques and Alimi, Omoniyi and Cameron, Michael P. and Maré, David C., The Gravity Model of Migration: The Successful Comeback of an Ageing Superstar in Regional Science. IZA Discussion Paper No. 10329, Available at SSRN: `https://ssrn.com/abstract=2864830`
- R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
- Roger S. Bivand, Edzer Pebesma, Virgilio Gomez-Rubio, 2013. Applied spatial data analysis with R, Second edition. Springer, NY. http://www.asdar-book.org/
- Ryan Hafen (2020). geofacet: 'ggplot2' Faceting Utilities for Geographical Data. R package version 0.2.0. https://CRAN.R-project.org/package=geofacet
- S. Madeh Piryonesi, Ph.D.; and Tamer E. El-Diraby, Ph.D. Role of Data Analytics in Infrastructure Asset Management: Overcoming Data Size and Quality Problems.
- Stewart, John Q. "Demographic Gravitation: Evidence and Applications." Sociometry 11, no. 1/2 (1948): 31-58. doi:10.2307/2785468.
- Tim Menzies, Ekrem Kocagüneli, Leandro Minku, Fayola Peters, Burak Turhan. Chapter 10 - Data Mining (Under The Hood), Sharing Data and Models in Software Engineering. 2015.
- Wooldridge, Jeffrey. (2002). Econometric Analysis of Cross Section and Panel Data.
- Yves Croissant (2020). pglm: Panel Generalized Linear Models. R package version 0.2-2. https://CRAN.R-project.org/package=pglm
- Zipf, G. (1946). The P1 P2/D Hypothesis: On the Intercity Movement of Persons. American Sociological Review, 11(6), 677-686. doi:10.2307/2087063

# 10 Appendix

## Variable Definitions

- **Destination**: The state to which people are migrating to.
- **Origin**: The state to which people are migrating from.
- **Population: Destination**: The total population of the *destination* state.
- **Population: Origin**: The total population of the *origin* state.
- **Per Cap Per Inc: Destination**: Per capita personal income of the *destination* state;
- **Per Cap Per Inc: Origin**: Per capita personal income of the *origin* state.
- **Total Employment: Destination**: The total number of people employed in the *destination* state.
- **Total Employment: Origin**: The total number of people employed in the *origin* state.
- **Destination State**: A dummy variable for the *destination* states.
- **Origin State**: A dummy variable for the *origin* states.

## Data Descriptive Statistics

Table 10: Mean Values by Year

| Year | Migration | Per Capita Personal Income | Population | Total Employment |
|---|---|---|---|---|
| 2011 | 137677.8 | 42285.73 | 6315950 | 3567476 |
| 2012 | 139571.4 | 44070.59 | 6362279 | 3625945 |
| 2013 | 141967.2 | 44213.82 | 6406398 | 3693793 |
| 2014 | 144624.2 | 46257.80 | 6453753 | 3773260 |
| 2015 | 149828.1 | 47931.37 | 6501583 | 3856388 |
| 2016 | 150184.6 | 48668.16 | 6548986 | 3918537 |
| 2017 | 148696.2 | 50466.24 | 6590926 | 3978822 |
| 2018 | 150634.9 | 52832.53 | 6626091 | 4059222 |
| 2019 | 146881.0 | 54553.63 | 6656750 | 4130969 |

Table 11: First Quantile by Year

| Year | Migration | Per Capita Personal Income | Population | Total Employment |
|---|---|---|---|---|
| 2011 | 55572 | 37506 | 1856606 | 1064267 |
| 2012 | 54205 | 39054 | 1857446 | 1067211 |
| 2013 | 55891 | 39271 | 1865813 | 1075465 |
| 2014 | 60851 | 40778 | 1879955 | 1083772 |
| 2015 | 59618 | 42599 | 1892059 | 1092255 |
| 2016 | 63624 | 42787 | 1906483 | 1092500 |
| 2017 | 56202 | 44376 | 1916998 | 1095372 |
| 2018 | 59558 | 46921 | 1925512 | 1110785 |
| 2019 | 58541 | 48188 | 1932571 | 1130618 |

Table 12: Second Quantile by Year

| Year | Migration | Per Capita Personal Income | Population | Total Employment |
|---|---|---|---|---|
| 2011 | 107899 | 40942 | 4576244 | 2427409 |
| 2012 | 107504 | 42896 | 4602067 | 2453282 |
| 2013 | 102474 | 43176 | 4626040 | 2500792 |
| 2014 | 109843 | 44930 | 4645938 | 2551872 |
| 2015 | 109921 | 46274 | 4666998 | 2587451 |
| 2016 | 112774 | 46520 | 4681346 | 2619796 |
| 2017 | 111851 | 48473 | 4673673 | 2649260 |
| 2018 | 109939 | 51144 | 4664450 | 2691497 |
| 2019 | 105650 | 52829 | 4658285 | 2735740 |

Table 13: Third Quantile by Year

| Year | Migration | Per Capita Personal Income | Population | Total Employment |
|---|---|---|---|---|
| 2011 | 189097 | 45345 | 6827479 | 4175311 |
| 2012 | 192986 | 47728 | 6898599 | 4253327 |
| 2013 | 185063 | 47932 | 6966252 | 4353568 |
| 2014 | 194932 | 50687 | 7057531 | 4447626 |
| 2015 | 201684 | 52336 | 7167287 | 4631417 |
| 2016 | 194456 | 52431 | 7299961 | 4709906 |
| 2017 | 194362 | 54930 | 7427951 | 4778377 |
| 2018 | 204761 | 57346 | 7526793 | 4854274 |
| 2019 | 194737 | 58830 | 7614024 | 4936751 |

Table 14: Fourth Quantile by Year

| Year | Migration | Per Capita Personal Income | Population | Total Employment |
|---|---|---|---|---|
| 2011 | 501950 | 67366 | 37636311 | 19986021 |
| 2012 | 523267 | 68310 | 37944551 | 20666908 |
| 2013 | 535088 | 67821 | 38253768 | 21319995 |
| 2014 | 536438 | 71217 | 38586706 | 21997098 |
| 2015 | 580407 | 75277 | 38904296 | 22687196 |
| 2016 | 601840 | 77629 | 39149186 | 23164907 |
| 2017 | 562984 | 78974 | 39337785 | 23538609 |
| 2018 | 582865 | 80943 | 39437463 | 24086110 |
| 2019 | 592602 | 83111 | 39437610 | 24602061 |

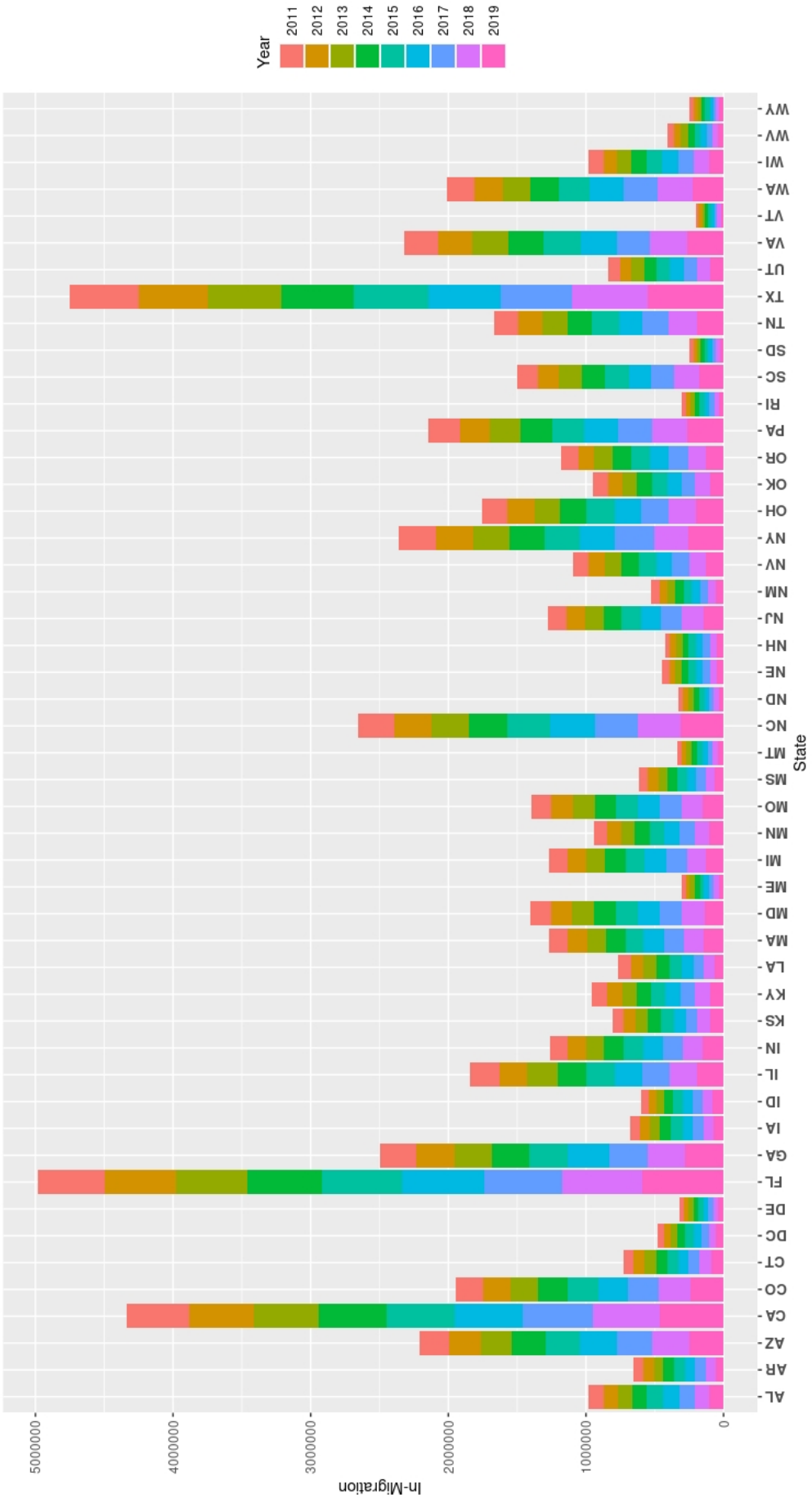


Figure 3: In-Migration by State (2011-2019)

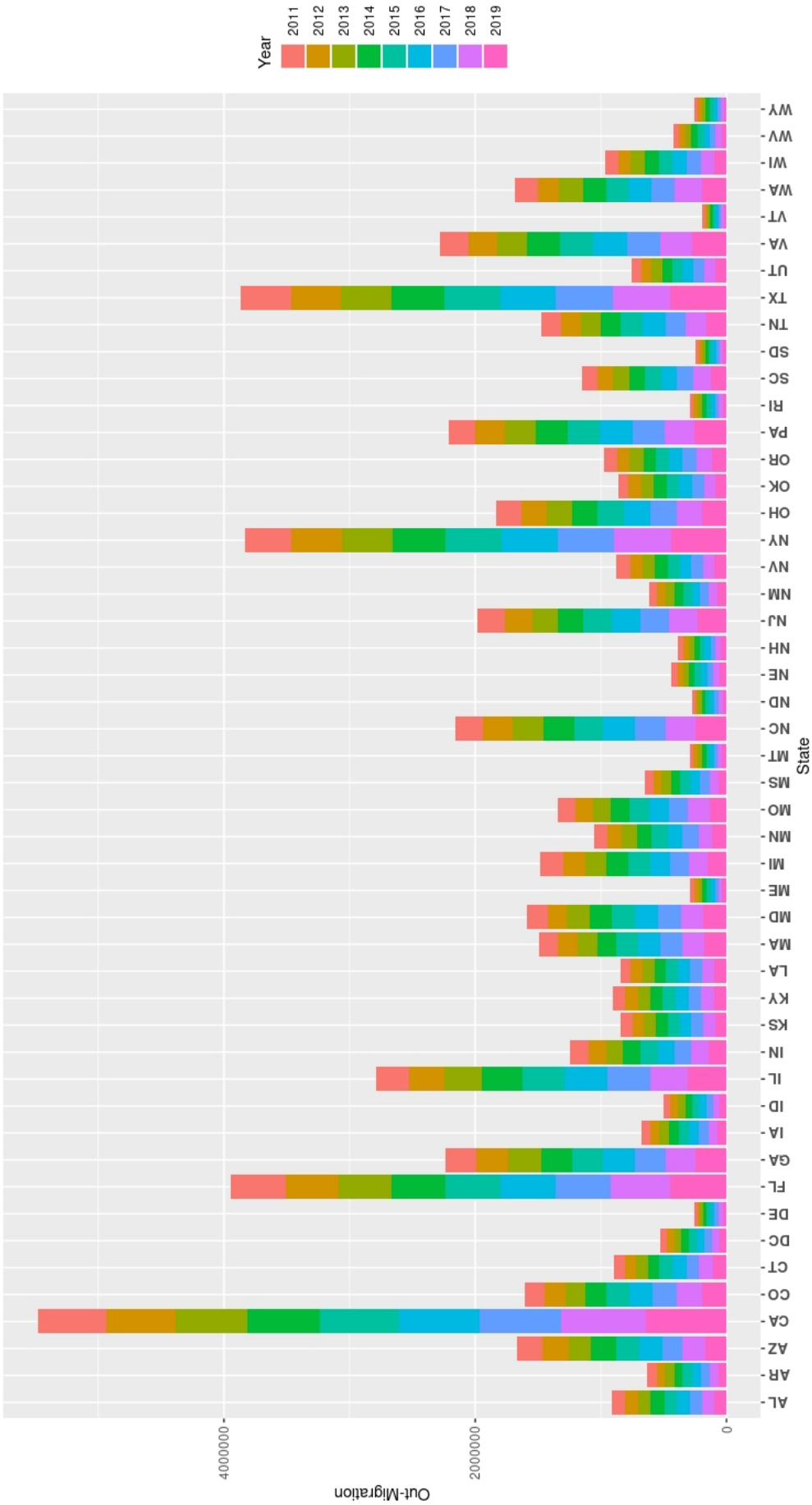


Figure 4: Out-Migration by State (2011-2019)

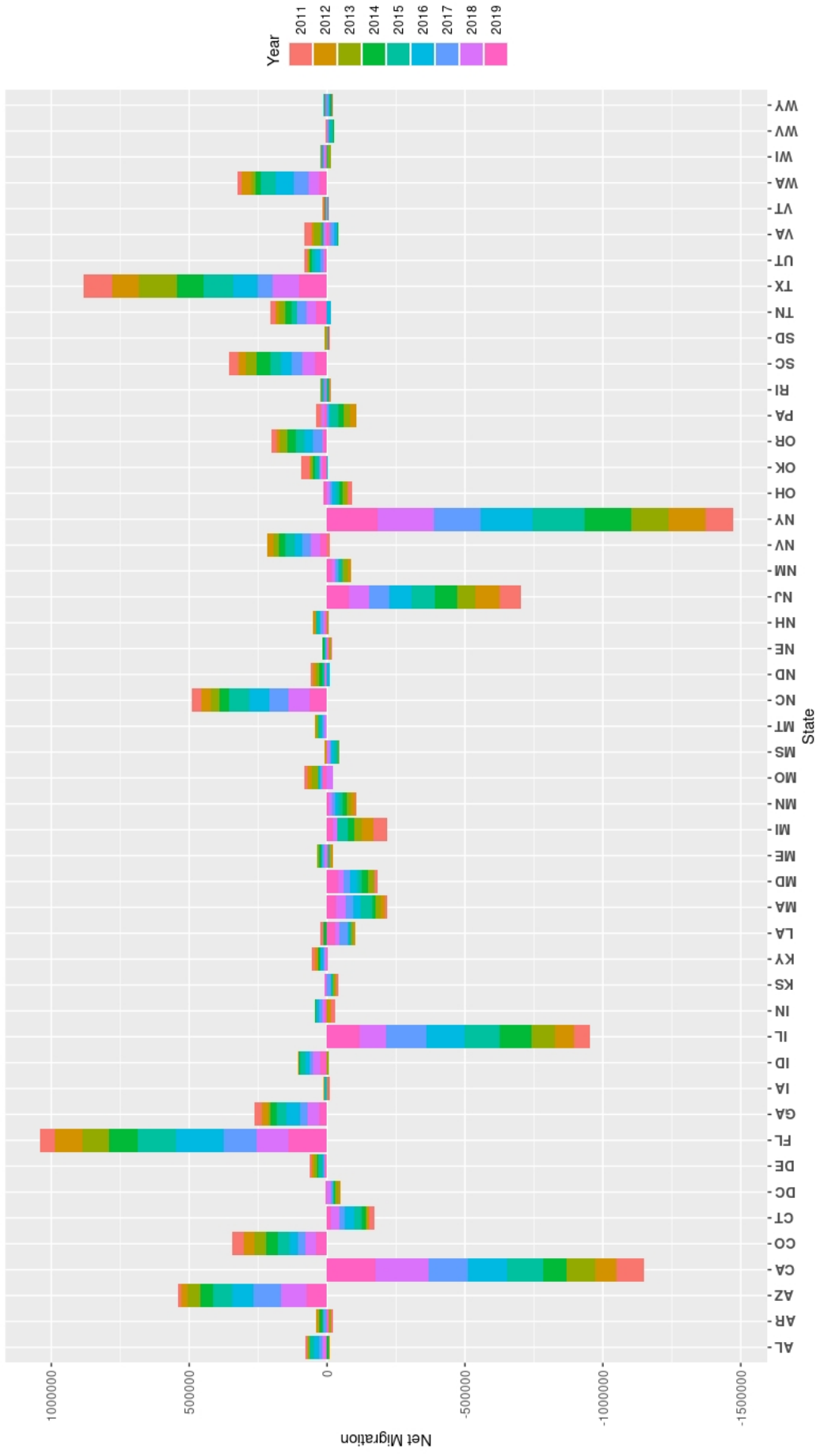


Figure 5: Net-Migration by State (2011-2019)

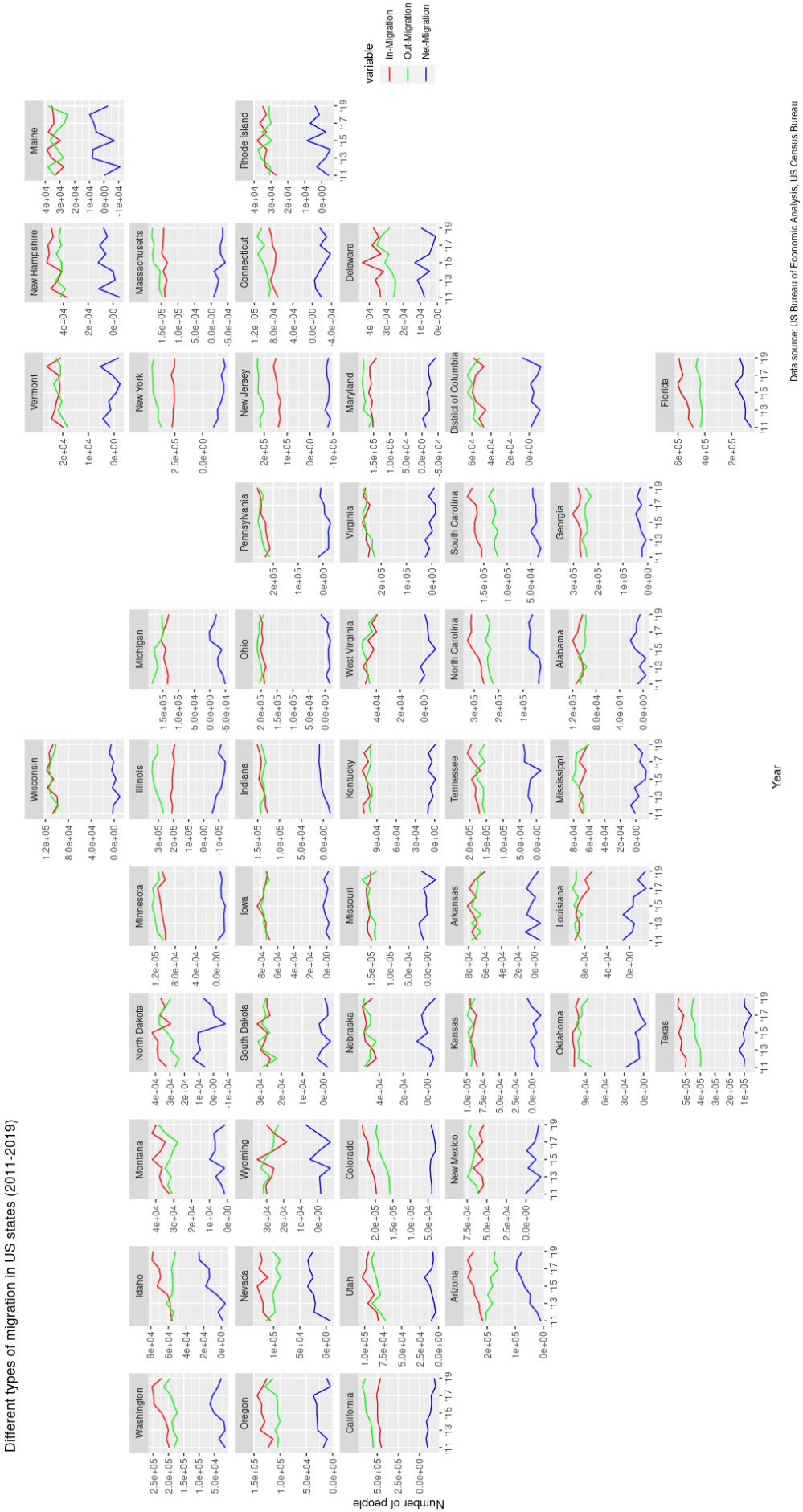


Figure 6: Migration in States (2011-2019)

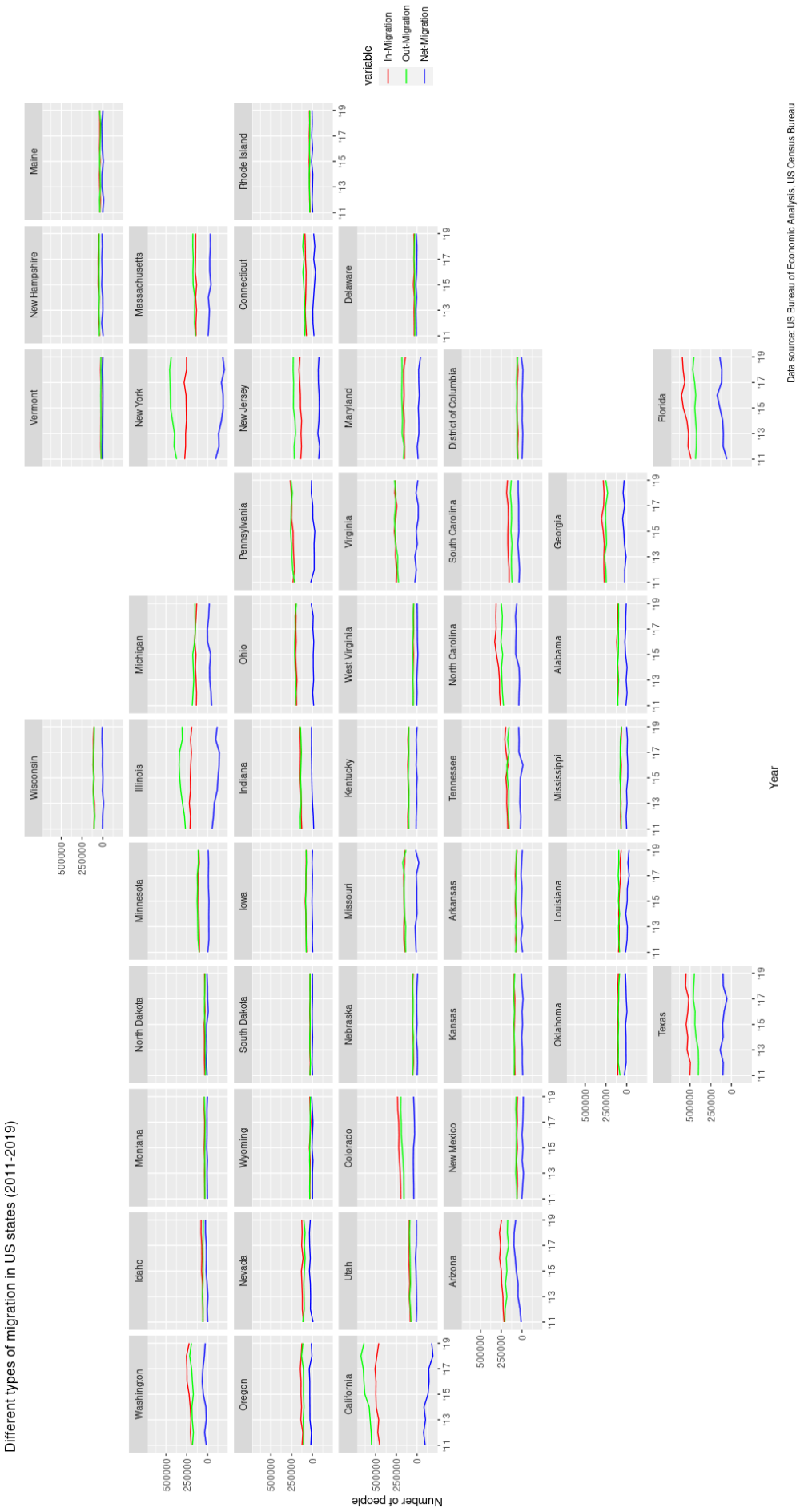


Figure 7: Migration in States (2011-2019)

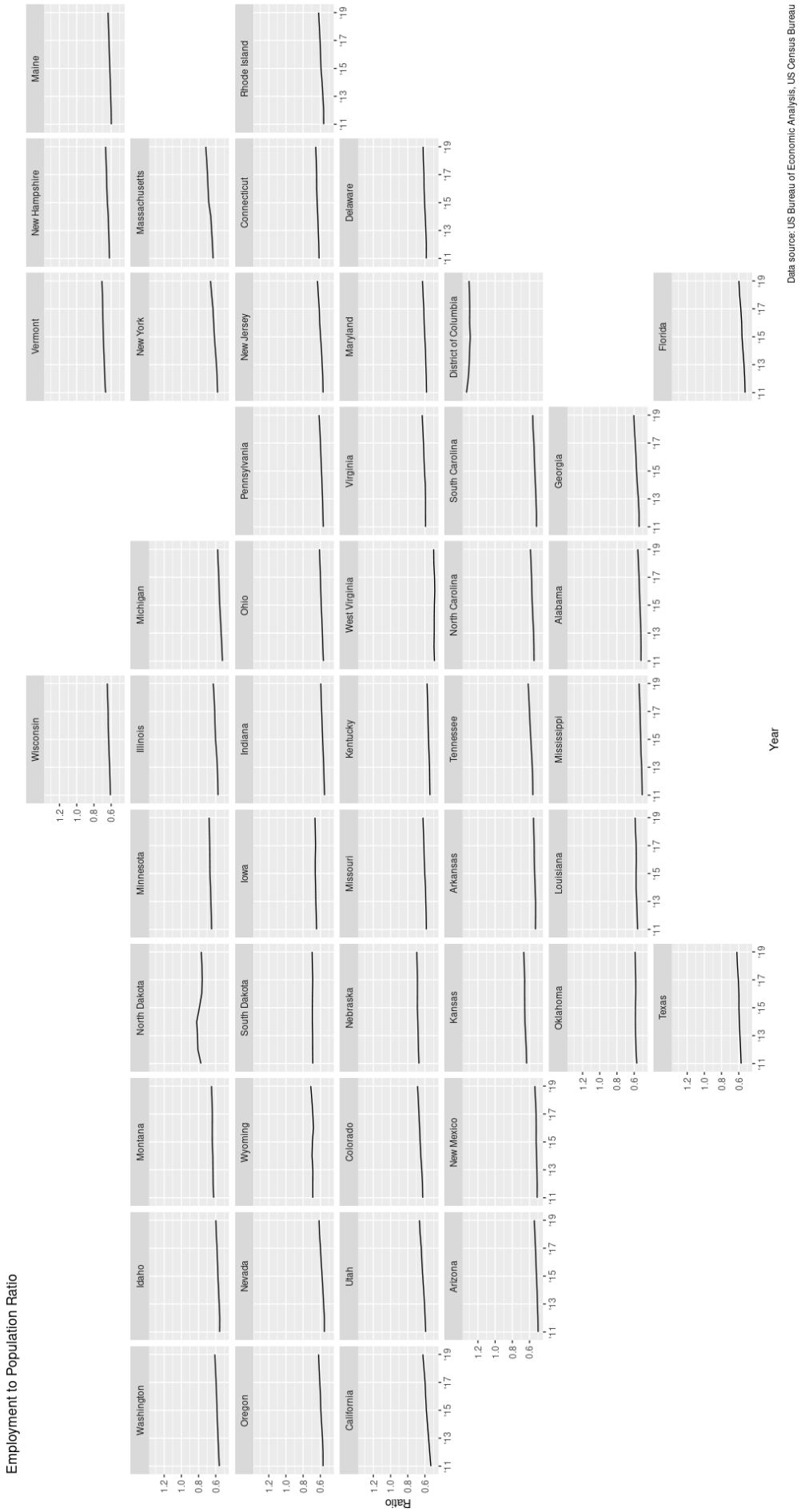


Figure 8: Employment to Population Ratio (2011-2019)

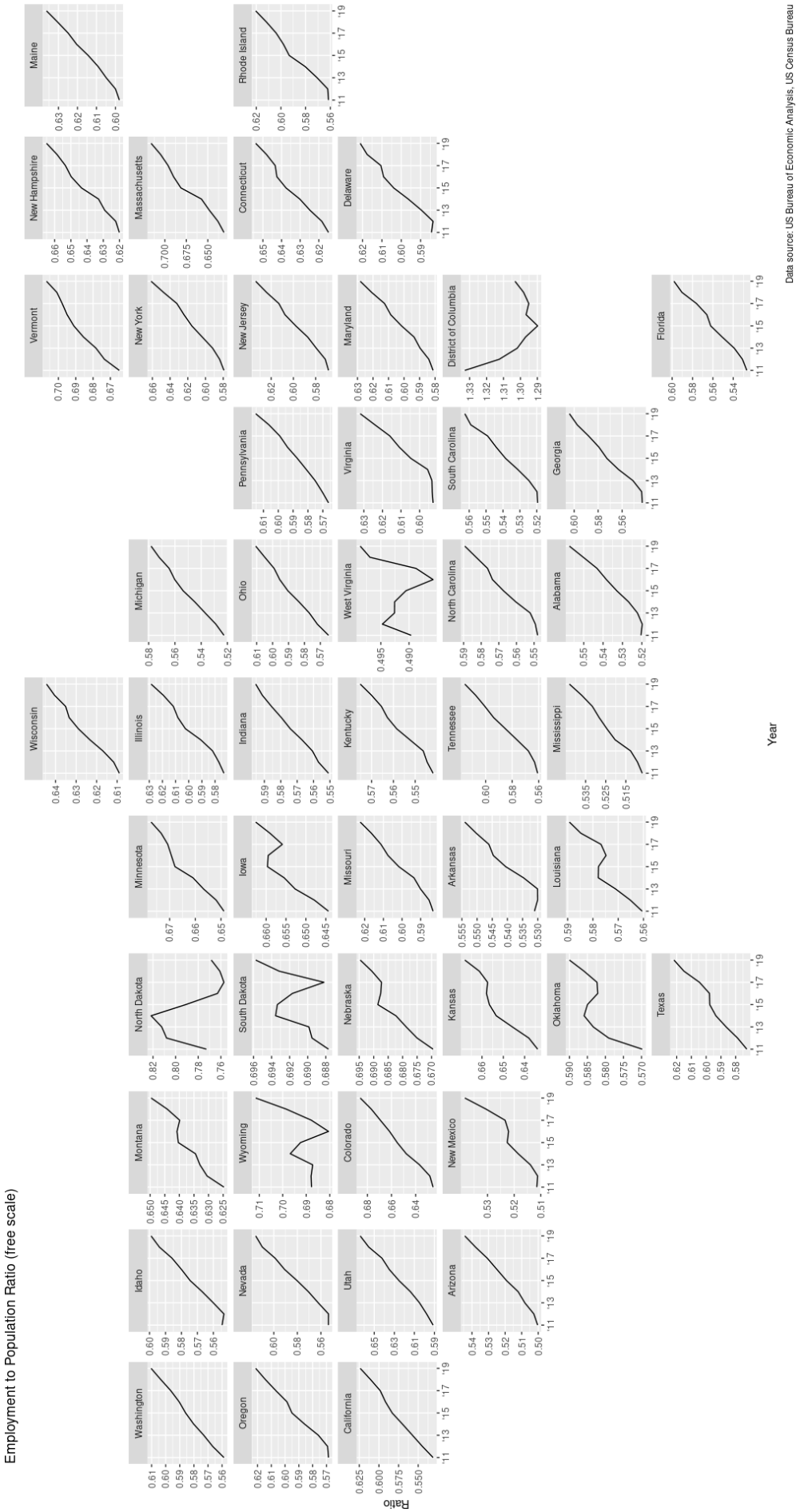


Figure 9: Employment to Population Ratio (free scale) (2011-2019)

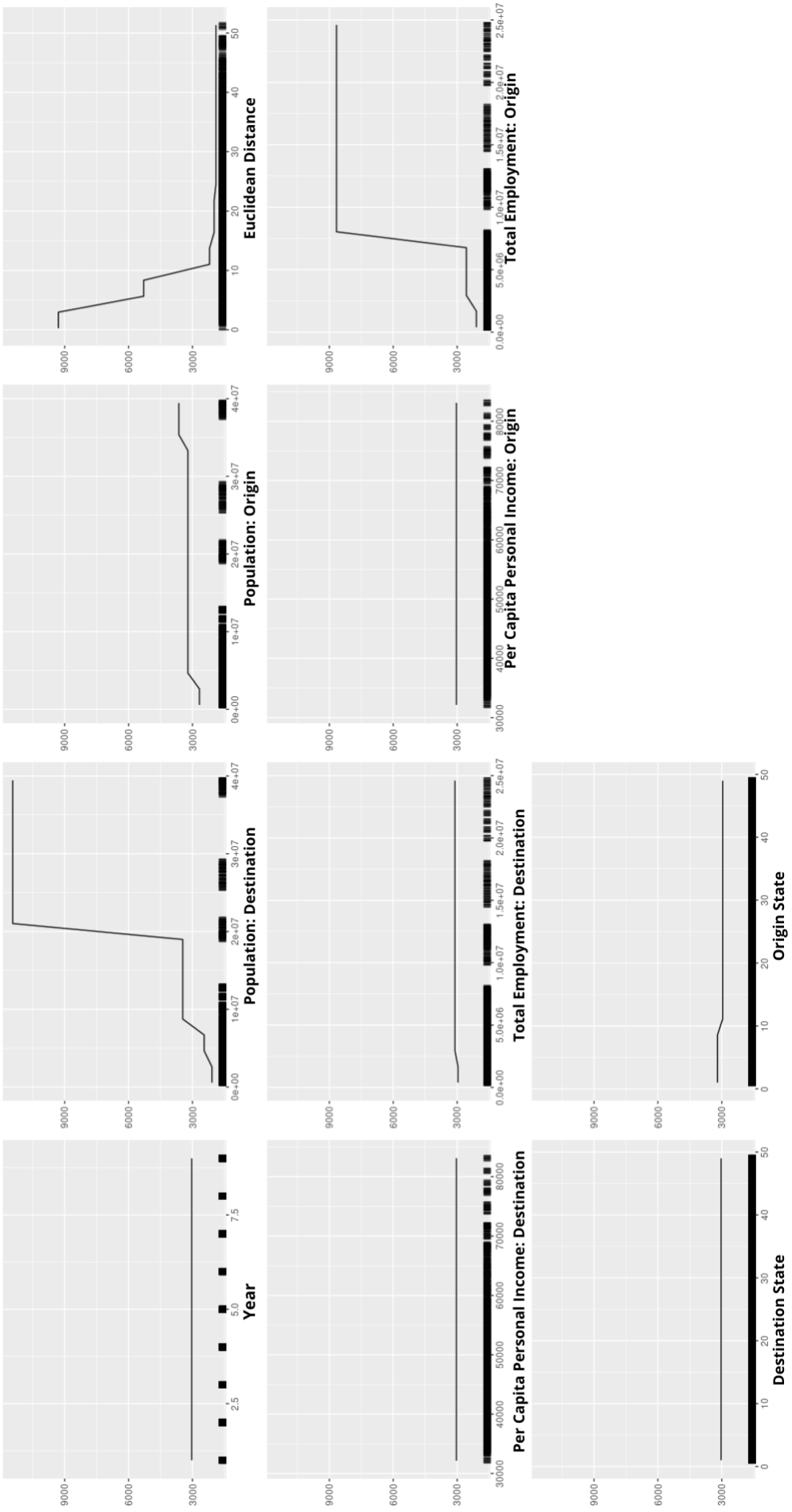


Figure 10: Partial Dependence Plots - Regression Tree

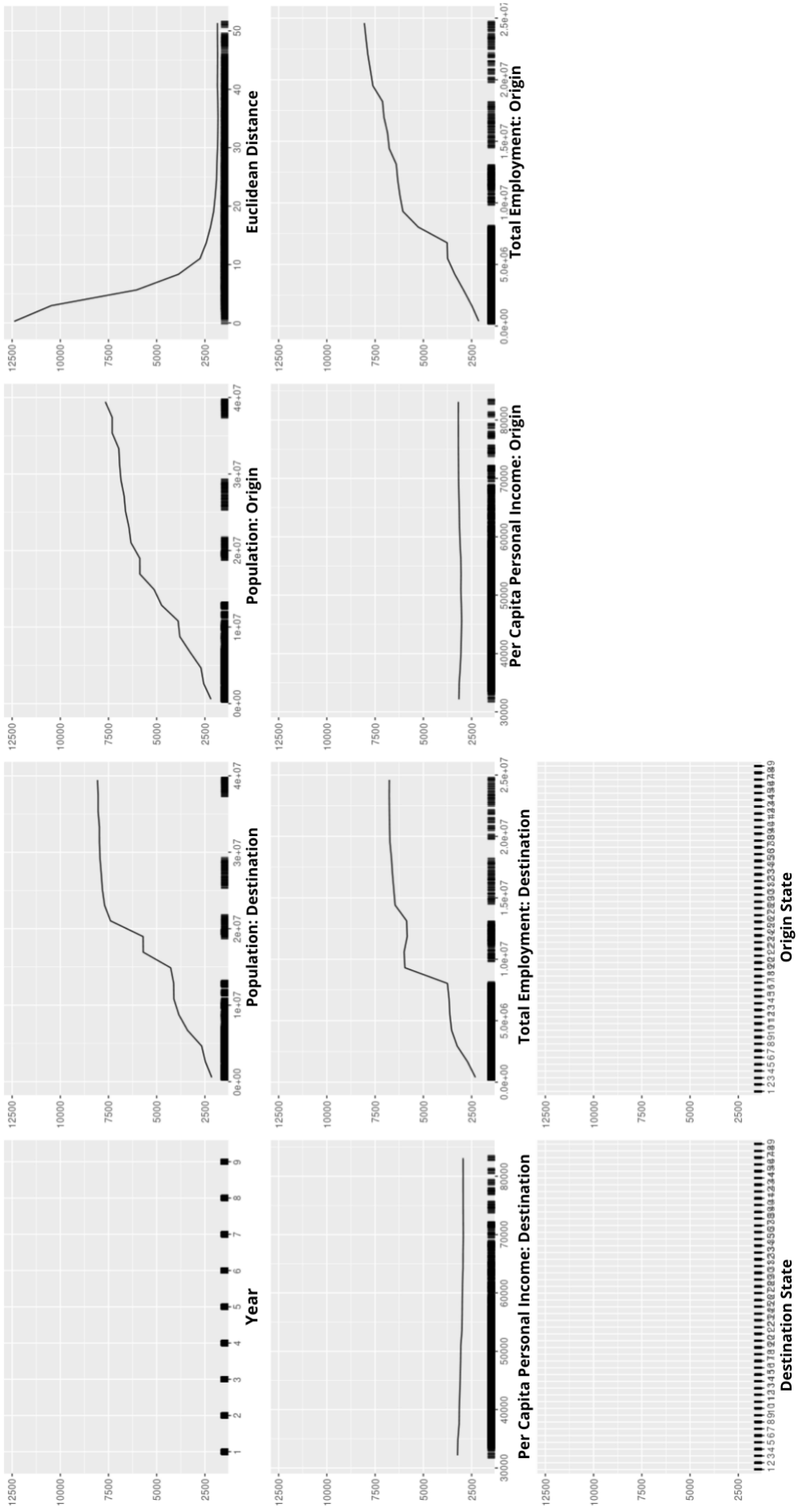


Figure 11: Partial Dependence Plots - Random Forests