**Perspectives on Computational Research: Final Paper**
**Xingyun Wu**
**6/4/2017**

# To Move or Not to Move:
# College Major and Geographical Mobility

## Abstract:

On the topic of college education and mobility, existing studies have examined both the relationship between years/levels of college education and geographical mobility and the relationship between college major and social mobility. But have rarely examined the relationship between college major and geographical mobility. This study attempts to analyze whether college major has impact on geographical mobility. If the answer were yes, what would be the effect? To answer these questions, this study combines decision tree and logistic regression to provide statistically supported classification of college majors, hypotheses testing, and prediction on geographical mobility. And the findings are that: (1) even after controlling the effect of original location, college major still has explanatory power on the occurrence of geographical mobility; (2) compared to other types of college majors, people with majors providing science/technology skills have higher probability of geographical mobility; (3) although the relationship between college major and geographical mobility is statistically significant, it needs to be used with other factors to gain the power of prediction.

**Key Words:**

College Major, Geographical Mobility, Decision Tree

## Introduction & Background

College graduates with different majors might have different skills required by the labor market. And opportunities might not be equally provided to people with

different kinds of skills. People with some certain skills might have different probabilities to obtain opportunities to move. However, existing studies on college education and mobility focus either on the relationship between years or levels of college education or the relationship between college major and social mobility. The relationship between college major and geographical mobility is rarely studied.

This study would attempt to analyze the relationship between college major and geographical mobility. And it would combine decision tree and logistic regression as its methods. Decision tree would be applied to statically classify college majors into several groups, on the basis of skills provided by college majors. Then logistic regression would be used for hypotheses testing. Finally, decision tree would be applied again to check college major's power of prediction.

The study would start with reviews on relevant theories and existing studies. Then would move on to hypotheses, data and methods used by this study, results, and finally conclusion and discussion.

## 1. Review on Migration or Geographical Mobility Theories

The concepts of migration and geographical mobility are different but highly related. Both of them refer to the movement of people from one location to another. But migration concerns moves across a boundary, while geographical mobility concerns short and long-distance moves (United States Census Bureau, 2017). In this study, I would treat migration as a subset of mobility, thus would refer to theories and literatures of both migration and geographical mobility.

This is an important field of demography, economy and sociology. Demographers at the U.S. Bureau of Census have made great contribution to this field (Bogue, 2009), such as Henry S. Shryock (Shryock, 1964; Thomas & Shryock, 1965) and Larry Long (1988). According to Larry Long (1988), Migration studies cares about who moves where, why they move, and how they decide to move or stay, and how they pick one area rather than another.

Theories of migration could date back to more than 100 years ago. E. G. Ravenstein's three articles form the basis of most modern research on migration (Grigg, 1977). He holds a linear theory of migration (Long, 1988). He predicted that most migration would come from rural areas and flow to urban areas, with differences by gender and age (Ravenstein, 1876, Ravenstein, 1885; Ravenstein, 1885). This was

supported by a classic statistical study (Weber, 1899). However, later researchers found that migration might also be influenced by other factors, such as shifting economic advantages of different areas (Goodrich et al., 1937), economic crises such as the Great Depression (Thompson, 1937), and features of people (Thomas, 1938). And these findings led to nonlinear theories (Long, 1988).

Since then, most theoretical attention has been shifted from the population of migration to the changing determinants of migration flows (Long, 1988). Currently, from the macro perspective, "push-pull" model still dominates contextual models of mobility, whose process is accompanied by environmental factors (Bogue et al., 2009). And from the micro perspective, the determinants could be gender, age, race-ethnicity, education, marital status, income level, and housing tenure (Bogur, 2009).

However, is the internal mobility within US still worth discussion, since international migration has become a greater concern of not only migration research but also social science research in this early 21st century (Korinek & Maloney, 2010)? Although the internal mobility rates are in a phase of long-term decline (due to aging, urbanization, decreased economic disparities among areas/regions, and technology), the amount of internal migration has a tendency to remain constant or to increase (Bogue, 2009). So it is still worthwhile to discuss the internal migration of the US.

## 2. Review on Education and Geographical Mobility

The effect of education on individual's mobility has long been discussed, especially social mobility. The classic status attainment model of Blau and Duncan (1967) and following studies indicate that individual's education attainment influences the differences in socioeconomic positions between generations (the father and the individual). However, is education related to individual's geographical mobility? Many studies have discussed this issue and proved that they are related, although the direction(s) of their relationship could be very complex.

An early study finds that part of the monetary return to schooling arises when people with more education could adapt to economic disequilibria more successfully (Bowles, 1970). Migration could be driven by economic incentives, while levels of education are involved in this process by influencing individual's ability of achieving this income gain (Bowles, 1970). However, education could also play a role in

making individual stay, through geographic linkage. A more recent study finds a modest link between attending college and working in the same state (Groen, 2004), but it still supports the view that the location of college influences the location of occupation. Some other researches hold a vague attitude towards the question how geographical mobility is influenced by education. For example, a recent population study verifies that internal migration age profiles closely mirror the age structure of key events in the life course, including exit from education and entry to the labor force (Bernard et al., 2014). This also indicates that education could be closely relevant to internal migration.

Therefore, existing researches support the influence of education on geographical mobility, through ability, geographic linkage, or some other factors.

## 3. Review on Major and Mobility

We have known from the previous part that education matters. But does majors of higher education matter? Generally speaking, returns to different college or graduate school majors are different (Altonji et al., 2015).

According to existing researches, college majors matters, but mainly from the perspective of social mobility rather than geographical mobility. Earning is a significant way of social mobility. Individuals may consider future earning-streams when making education investments in general (Berger, 1988), although the effect of expectation of earnings could be differentiated by gender (Eide & Waehrer, 1998; Montmarquette et al., 2002) and race (Montmarquette et al., 2002). Researcher has also found large differences in earning premiums among people from different majors, with natural science and business majors enjoying higher earning premiums (Arcidiacono, 2004). Actual earnings could be influenced by whether the occupation is related to major (Robst, 2007), but again, major matters for social mobility through earnings.

To express this issue in a more theoretical way, college major affects individual's placement on and movement along the social ladder, by inhibiting the degree to which people's education attainment allows them to be socially mobile (Wolniak et al., 2008). And this pattern persists. According to a more recent research, STEM and business-related majors still lead the way (Altonji et al., 2015). Thus, for college attendants, the choice of major is an important issue for their mobility, more

specifically social mobility. However, none of these researches above have referred to geographical mobility.

### 4. A Brief Summary

The importance of studying internal migration does not decrease with the rise of importance of immigration studies. Education, among several demographic or social factors, is an important determinant of migration or geographical mobility. Existing empirical researches have supported that education influences geographic mobility, and that major of higher education influences individual's social mobility. But studies about education and geographic mobility do not analyze the effect of majors in higher education, while studies about major and mobility mainly concern social mobility and ignoring geographical mobility. None of them seems to have taken a further step to examine whether major matters for individual's ability in geographic mobility.

Therefore, I would take this further step, by analyzing whether major of higher education influence individual's geographic mobility. To be more specific, this study would concern whether is any difference between individuals with different majors in geographic mobility.

## Hypotheses

My research question is: is there a relationship between college major and geographical mobility? If the relationship exists, what is the effect of college major on geographical mobility? More specifically, does the relationship remain the same among all majors?

However, the original location might influence the occurrence of geographical mobility. And people from different original location might hold different views on college majors when making decisions of education investment. So I would first test the influence of original location, then add college major into consideration. Correspondingly, I have two hypotheses to structure my questions.

**$H_1$: Hypothesis of Locational Dependence**

Original location has impact on geographical mobility

**H2: Hypothesis of Influence of College Major**

Majors of college education has impact on whether geographical mobility exists

# Data

## 1. Data Source

To conduct analysis on the above hypotheses, this study needs data with information of college degree, occupation, and demographics. The data I would use is from the National Survey of College Graduates, which could be directly downloaded on its official website.

It is a longitudinal biennial survey, particularly focusing on the science and engineering workforce. The respondents are individuals under the age of 76 by February 1 2015, with at least a bachelor's degree by January 1 2014, and living in the U.S. during the survey reference period. The survey has been conducted since the 1993, replacing the previous Survey of Natural and Social Scientists and Engineers (SSE) which began in 1972. This study would use the most recent collection, the 2015 NSCG.

The initial data was collected with a self-administered web survey. For the occurrence of nonrespondents, a self-administered mail survey would be held. And nonrespondents to the mail survey would receive computer-assisted telephone interviewing (CATI). With these efforts, the 2015 NSCG has a weighted response rate of 70%. According to data documentation on its official site, this dataset contains 135,000 sample cases. However, due to the nonresponses, its public version only contains 91,000 observations.

## 2. Descriptive Statistics

This study mainly uses 11 variables. 3 of them are used in original setting, 5 of them are adjusted, and the other three of them are constructed based on several other variables in the dataset.

Table 1: Classification of Variables

| Original Variables | Reclassified Variables | Constructed Variables |
|---|---|---|
| Age | Marital Status | Race |
| Gender | College Major | Whether Move |
| Annualized Salary | Location of High School | Distance of Mobility |
| | Locaton of Work | |

| Key Variables | Description |
|---|---|
| **Dependent Variables** | |
| Whether Moved | For logistic regression and prediction tree. 1 if original location is different from occupational location, 0 if not |
| Skills required by Job | For college major classification. Categorical, represents types of skills required by respondent's current job |
| **Independent Variables** | |
| College Major | Categorical, 20 college major categories |
| Age | Continuous, age of respondents |
| Female | 1 if the respondent is female, 0 if is male |
| Race | Categorical, race of respondents |
| Stable Relationship | 1 if the respondent is in a stable romantic relationship, 0 if not |
| Original Location | Categorical, 1 to 9 represents regions within U.S., 10 represents foreign countries |

Note that the annualized salary uses imputation to deal with missing values. According to data documentation, the general range of nonresponse rate for key items is from 0.0% to 0.6%, but salary has a nonresponse rate around 11%. Since this assignment is to provide initial results and insights, I use unconditional mean imputation to simply adjust this problem. The non-missing values and mean of the annualized salaries are not revised.

The original variable of marital status includes married, living in a marriage-like relationship, widowed, separated, divorced, and never married. To simplify the situation, I classify the married observations and living in a marriage-like relationship observations into one category with stable relationship. And I classified the others into another category without stable relationship.

Variables of college major and locations are also reclassified. The raw data of college major has more than 100 majors indicated. In this assignment, I categorized them into three categories: (1) engineering, computer science, math, and natural sciences; (2) social sciences; (3) others. And the raw data of locations contains many

abroad countries. I converge all the abroad countries into one category: abroad.

And I constructed three variables based on several variables in the raw data. The variable of race is constructed with variables indicating Asian, black and Hispanic in the raw data. The variable of whether individuals move is constructed on whether the observations report the same location of high school and work.

The variable of mobility distance is constructed on the basis of difference between the location of high school and the location of occupation. The distance is recorded as 1 when individuals move to their neighboring regions, and recorded as 2 when they move to regions next their neighboring regions. The rest is defined as the same rule, with the biggest internal mobility distance recorded as 5. If individuals move abroad, the distance is recorded as 8. Although the values of this variable are discrete, they represent the distance in order. So in data analysis, I would treat this variable as continuous variables.

Table 2: Descriptive Statistics

Continuous Variables

| Variable | Mean | Standard Deviation | Min | Max | Observat-ions |
|---|---|---|---|---|---|
| Age | 44.26 | 14.16 | 16 | 75 | 91,000 |
| Annual Salary | 83727.09 | 83349.38 | 0 | 1223166 | 91,000 |
| Mobility Distance | 1.85 | 2.54 | 0 | 8 | 91,000 |

Categorical Variables

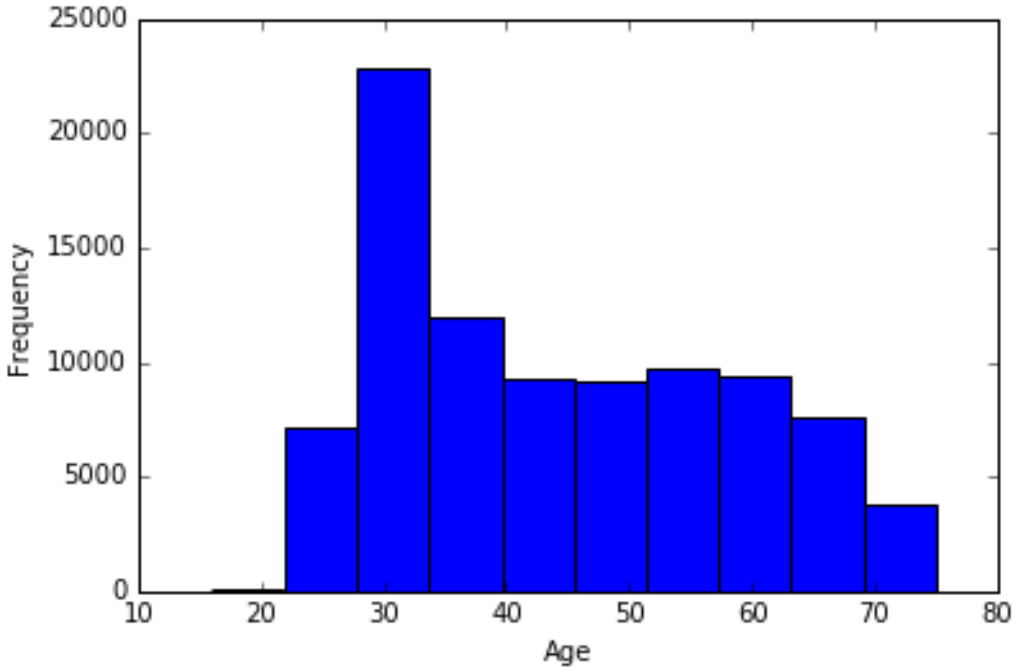
| Variable | Categories | Percent | Observations |
|---|---|---|---|
| Gender | | | 91,000 |
| | Male | 53.18 | 48,396 |
| | Female | 46.82 | 42,604 |
| Marital Status | | | 91,000 |
| | Married or in a married-like relationship | 72.58 | 66,052 |
| | Others | 27.42 | 24,948 |
| Race | | | 91,000 |

| | | | |
|---|---|---|---|
| | None of the below | 64.87 | 59,028 |
| | Asian | 16.62 | 15,122 |
| | Black | 8.35 | 7,594 |
| | Hispanic | 10.17 | 9,256 |
| College Major | | | 91,000 |
| | Engineering, CS, Math, and Natural Sciences | 48.42 | 44,063 |
| | Social Sciences | 21.07 | 19,171 |
| | Others | 30.51 | 27,766 |
| If Moved | | | 91,000 |
| | No | 56.32 | 51,251 |
| | Yes | 43.68 | 39,749 |
| Location of High School | | | 91,000 |
| | New England | 4.96 | 4,516 |
| | Middle Atlantic | 14.91 | 13,571 |
| | East North Central | 15.74 | 14,324 |
| | West North Central | 7.39 | 6,722 |
| | South Atlantic | 11.76 | 10,706 |
| | East South Central | 3.41 | 3,106 |
| | West South Central | 6.72 | 6,116 |
| | Mountain | 4.48 | 4,075 |
| | Pacific & US Territories | 13.84 | 12,596 |
| | Abroad | 16.78 | 15,268 |
| Location of Employer | | | 76,814 |
| | New England | 6.1 | 4,684 |
| | Middle Atlantic | 0.09 | 71 |
| | East North Central | 14.51 | 11,144 |
| | West North Central | 14.99 | 11,518 |
| | South Atlantic | 7.27 | 5,588 |
| | East South Central | 18.24 | 14,014 |
| | West South Central | 3.54 | 2,717 |
| | Mountain | 8.95 | 6,873 |

| | | | |
|---|---|---|---|
| | Pacific & US Territories | 6.38 | 4,899 |
| | Abroad | 19.93 | 15,306 |
| Location of Respondent | | | 91,000 |
| | New England | 5.98 | 5,446 |
| | Middle Atlantic | 14.45 | 13,146 |
| | East North Central | 14.89 | 13,550 |
| | West North Central | 7.18 | 6,530 |
| | South Atlantic | 18.26 | 16,614 |
| | East South Central | 3.61 | 3,287 |
| | West South Central | 8.92 | 8,121 |
| | Mountain | 6.54 | 5,955 |
| | Pacific & US Territories | 20.08 | 18,275 |
| | Abroad | 0.08 | 76 |

Table 2 shows descriptive statistics of the variables. And the plots below show the distribution of some key variables and some visualized relationships.

Figure 1: Distributon of Age




According to Figure 1, the distribution of age in the sample is not normal. This is because the 2015 NSCG has an oversample of young graduates. The group around the age of 30 has higher observations than the other groups. Simply adjust the age

distribution may cause more serious problems. By now, I have not found a suitable way to adjust this problem. So I would not apply variable transformation on age in data analysis.

Figure 4: Whether Individuals Move Across Age

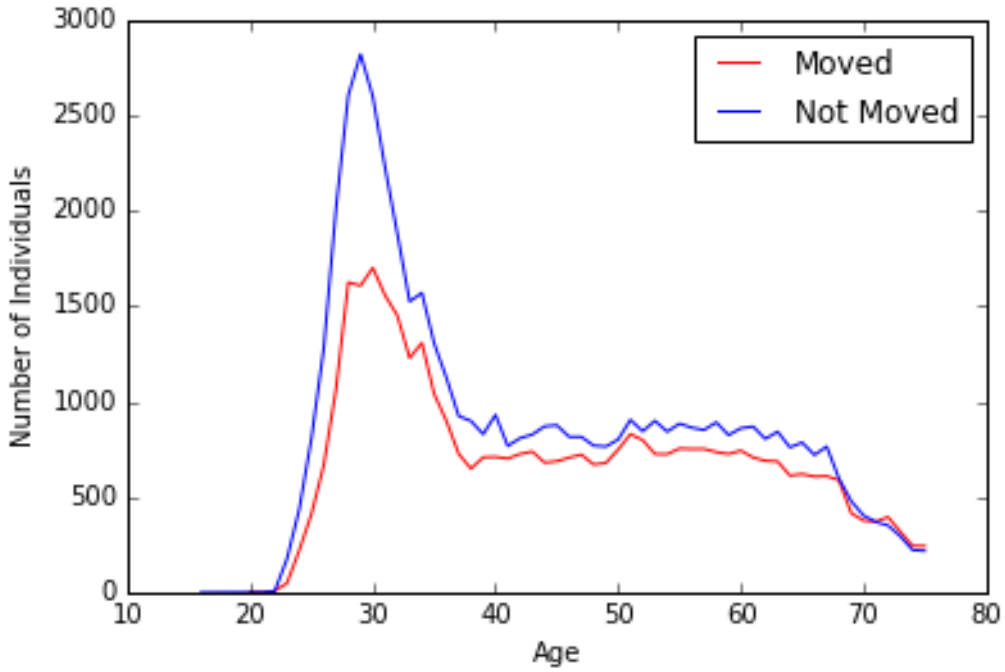
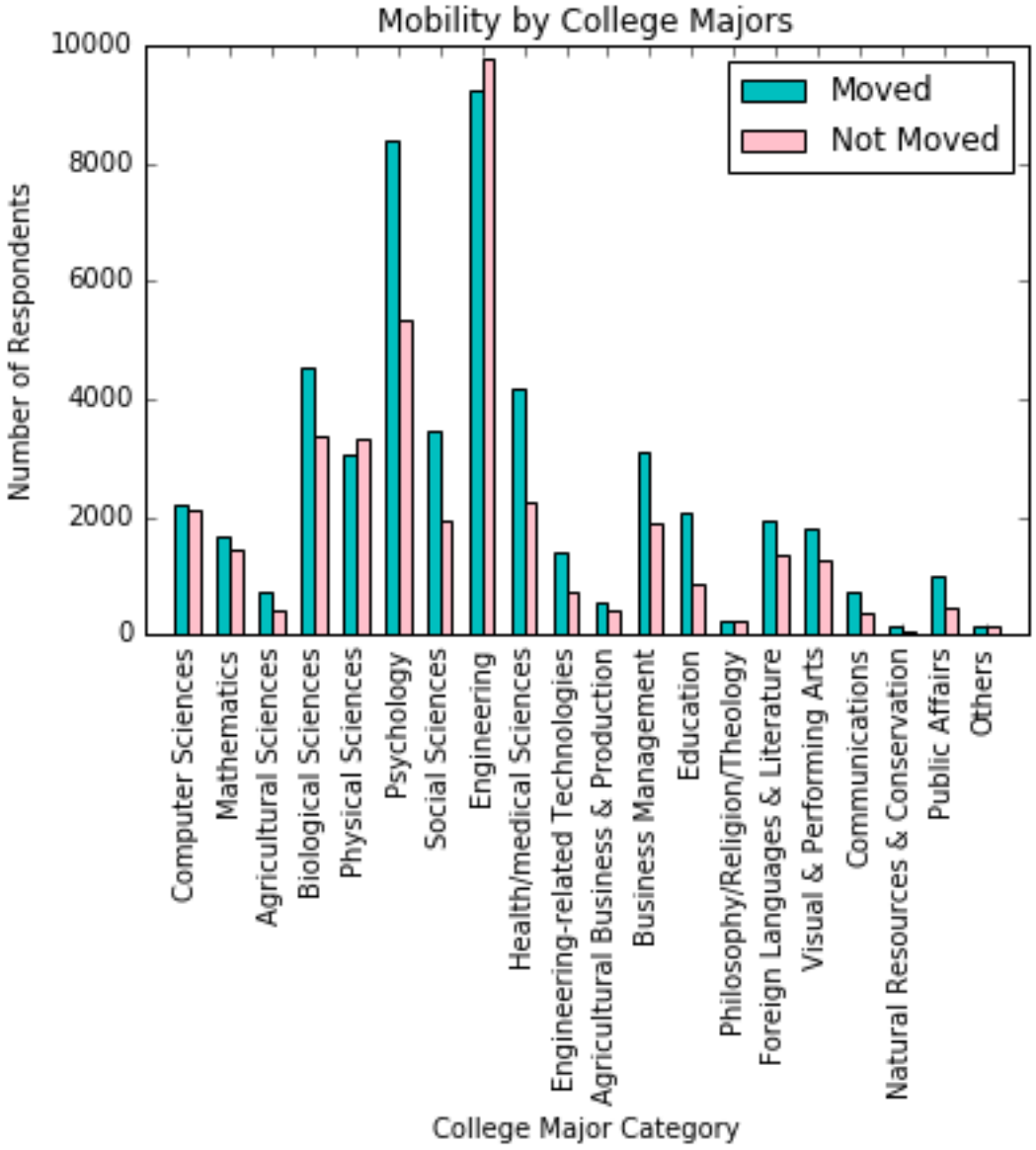



Figure 4 visualized the relationship between age and whether individuals move. Age distributions are generally similar for individual moved and not moved. Individuals with age around 30 have much higher frequency of mobility, which may due to the oversampling of the young graduates. Approximately after the age of 67, the frequencies of both moved observations and the not-moved observations drop, which may due to the age distribution of the sample. And almost across all ages, the not-moved observations are more than the moved observations.

In the raw data, there are 142 fields of study of respondents' first bachelor's degree. I categorize them into 20 college major categories. For example, in the category of computer and information sciences, there are computer and information sciences, computer science, computer system analysis, information services and systems, and other computer and information sciences. I categorize these fields into the category called computer sciences. And I also categorize other fields of study into corresponding college major categories, on the basis of NSCG's documentation. Please see the Appendix 1 for detailed information.

Figure 3 shows the distribution of college majors by the dummy variable of geographical mobility. It shows that in the dataset, there are more college graduates with majors of engineering, psychology and health/medical sciences than with other majors. And different college major categories have different proportions of the moved and the not-moved. This indicates that there might be a relationship between college major and geographical mobility.




## Methods

This study would use decision tree for classification and prediction, and use logistic regression for hypotheses testing. Although the fields of study have been

reduced to 20 college major categories, they need further aggregation. So I apply decision tree to classify the college major categories. Then use logistic regression to test my proposed hypotheses. Finally, I use decision tree to check whether college major has power of prediction to geographical mobility and how well it could be used.

**Decision Tree for Classification**

The contents learned in college would equip individuals with skills required by the labor market. So college graduates with different majors would go to jobs require different types of skills. In the 2015 NSCG, there are three variables regarding whether current job require certain technical expertise at the bachelor's level or hither. The first one asks about technical expertise of engineering, computer science, math or natural sciences. The second one asks about technical expertise of *other* fields such as health or business. And the third one asks about the technical expertise of social sciences. Based on these three variables, I construct a new variable to record types of match between college major and job.

Table 3: Construction of Major-Job Match Type

| Type of Match | Code | Sci/Tech Skills | Other Skills | Social Sciences Skills |
|---|---|---|---|---|
| None | 0 | ✗ | ✗ | ✗ |
| Only Sci/Tech | 1 | ✓ | ✗ | ✗ |
| Only Others | 2 | ✗ | ✓ | ✗ |
| Only Soc | 3 | ✗ | ✗ | ✓ |
| Sci & Oth | 4 | ✓ | ✓ | ✗ |
| Sci & Soc | 5 | ✓ | ✗ | ✓ |
| Oth & Soc | 6 | ✗ | ✓ | ✓ |
| All | 7 | ✓ | ✓ | ✓ |

Note1: the symbol ✗ means not required, and ✓ means required.

Note2: *Sci/Tech* is the abbreviation of Science/Technology, *Oth* is the abbreviation of Other Skills, and *Soc* is the abbreviation of Social Sciences.

Then in the tree model, I would use the constructed variable as the dependent variable and college major as the single predictor, to see what type of job would

college graduates with a certain major would take. College majors go to the same type of job would be categorized into the same type of majors. This approach would help reduce the complexity of the following models.

**Logistic Regression for Hypotheses Testing**

Logistic regression would be used for hypotheses testing. I would firstly apply a baseline model, add the variable of original location to test $H_1$ in the second model, and then add the variable of college major to test $H_2$ in the third mode.

Since demographic features may influence the occurrence of geographical mobility, age, gender and race are included in the baseline model. For race, I use three dummy variables regarding black, Asian and Hispanic. Since marital status might also impact on the probability of geographical mobility, a dummy variable about whether in marriage or marriage-like relationship is also included in the baseline model to control its effect. And parents' education, as indicators of original socioeconomic status, is also taken into consideration. This is because people from families with higher socioeconomic status might be more able to afford the cost or risks to move.

For the baseline model, the equation would be:

$$\log\left(\frac{\mathrm{P(moved)}}{\mathrm{P(not\text{-}moved)}}\right) = \beta_0 + \beta_1 * AGE_i + \beta_2 * GENDER + \beta_3 * \mathrm{BLACK}_i + \beta_4\mathrm{*ASIAN} + \beta_5\mathrm{*HISPANIC} + \beta_6 * \mathrm{STABLE\_RELATIONSHIP}_i + \beta_7 * \mathrm{FATHER\_EDU_i} + \beta_8\mathrm{*MOTHER\_EDU_i} + \varepsilon_i \quad (1)$$

The original location is added in the second model to test $H_1$.

$$\log\left(\frac{\mathrm{P(moved)}}{\mathrm{P(not\text{-}moved)}}\right) = \beta_0 + \beta_1 * AGE_i + \beta_2 * GENDER + \beta_3 * \mathrm{BLACK}_i + \beta_4\mathrm{*ASIAN} + \beta_5\mathrm{*HISPANIC} + \beta_6 * \mathrm{STABLE\_RELATIONSHIP}_i + \beta_7 * \mathrm{FATHER\_EDU_i} + \beta_8\mathrm{*MOTHER\_EDU_i} + \beta_9\mathrm{*ORIGINAL\_LOCATION} + \varepsilon_i \quad (2)$$

The variable of college major is added in the third model to test $H_2$.

$$\log\left(\frac{\text{P(moved)}}{\text{P(not-moved)}}\right) = \beta_0 + \beta_1 * AGE_i + \beta_2 * GENDER + \beta_3 * \text{BLACK}_i + \beta_4\text{*ASIAN} + \beta_5\text{*HISPANIC} + \beta_6 * \text{STABLE\_RELATIONSHIP}_i + \beta_7 * \text{FATHER\_EDU}_i + \beta_8\text{*MOTHER\_EDU}_i + \beta_9\text{*ORIGINAL\_LOCATION} + \beta_{10} * MAJOR_i + \varepsilon_i \quad (3)$$

**Decision Tree for Prediction**

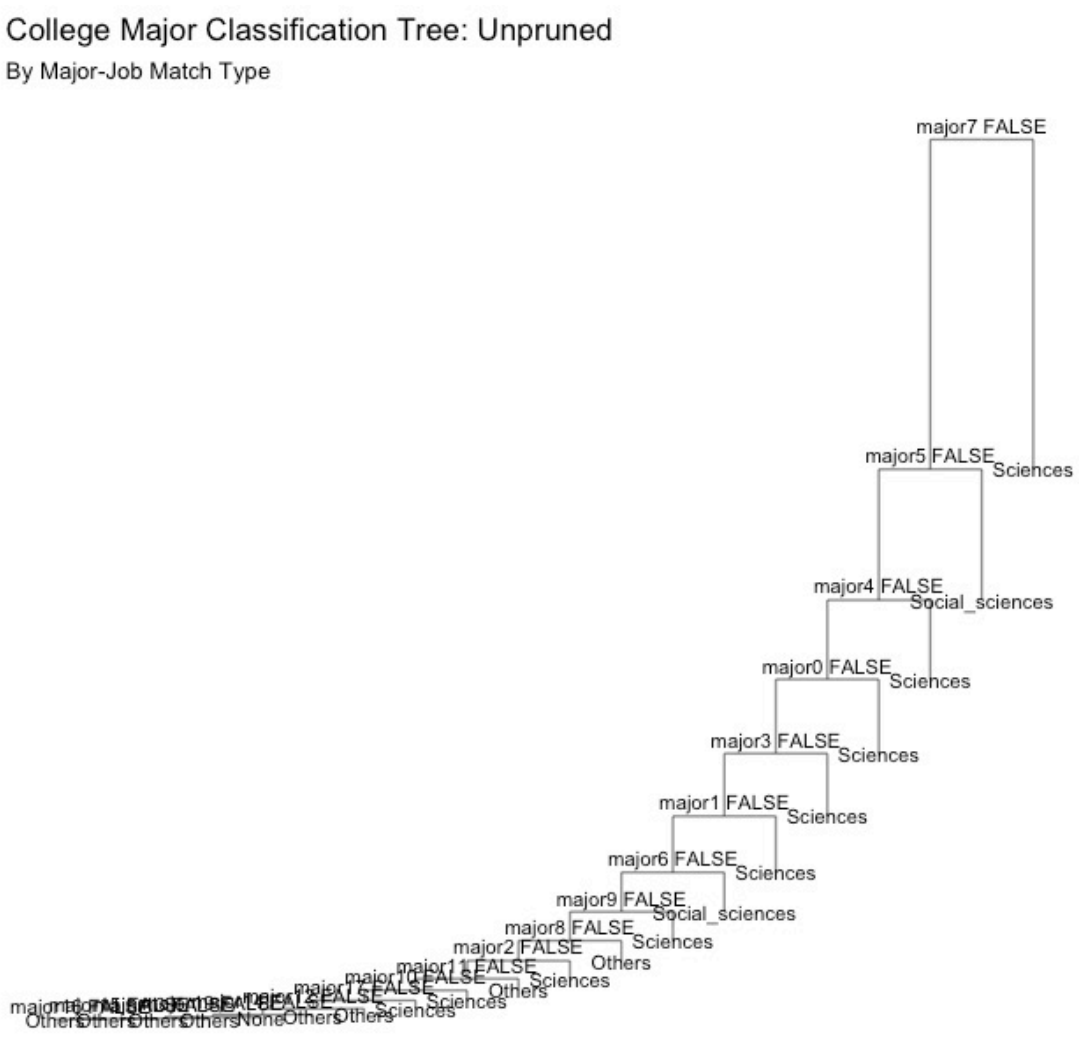
In the final step, I use decision tree to check college major's power of prediction. First, I use the type of college major as the single predictor, to check whether different types of college major could lead to distinguishing results. Further, I use a regression tree containing age, gender, race, relationship, and types of college major, to check the relative importance of college major in prediction. Since the tree model has a limited maximum of depth, which does not allow me to include all the variables used by logistic regression model in the previous step, I omit the variable of mother's education and original location in the tree model with multiple predictors.

## Results

### 1. College Major Classification

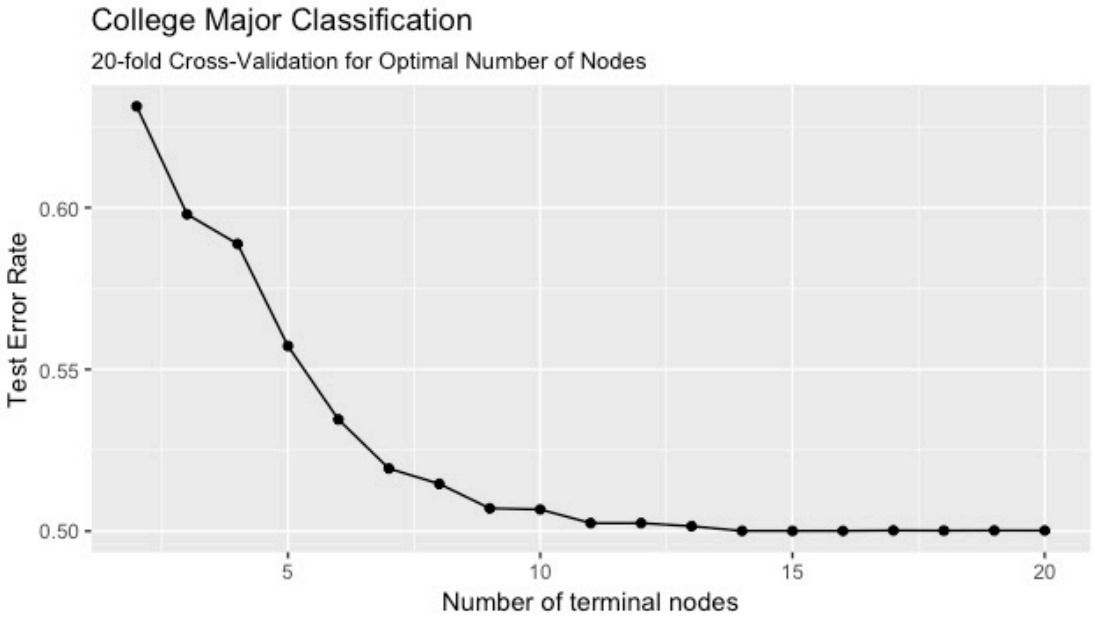
The tree plot below shows the result of classification tree. There are 20 college major categories. Correspondingly, the unpruned tree has 20 nodes, with each node going to one of the 8 major-job match types constructed in the Methods section. However, the 20 college major categories only go to 3 of the major-job match types: Only Sci/Tech, Only Oth, and Only Soc. Compared to the distribution of major-job match types, the result indicates that college major would generally equip individuals with single type of skill required by the labor market.

Figure: College Major Classification Tree (Unpruned)



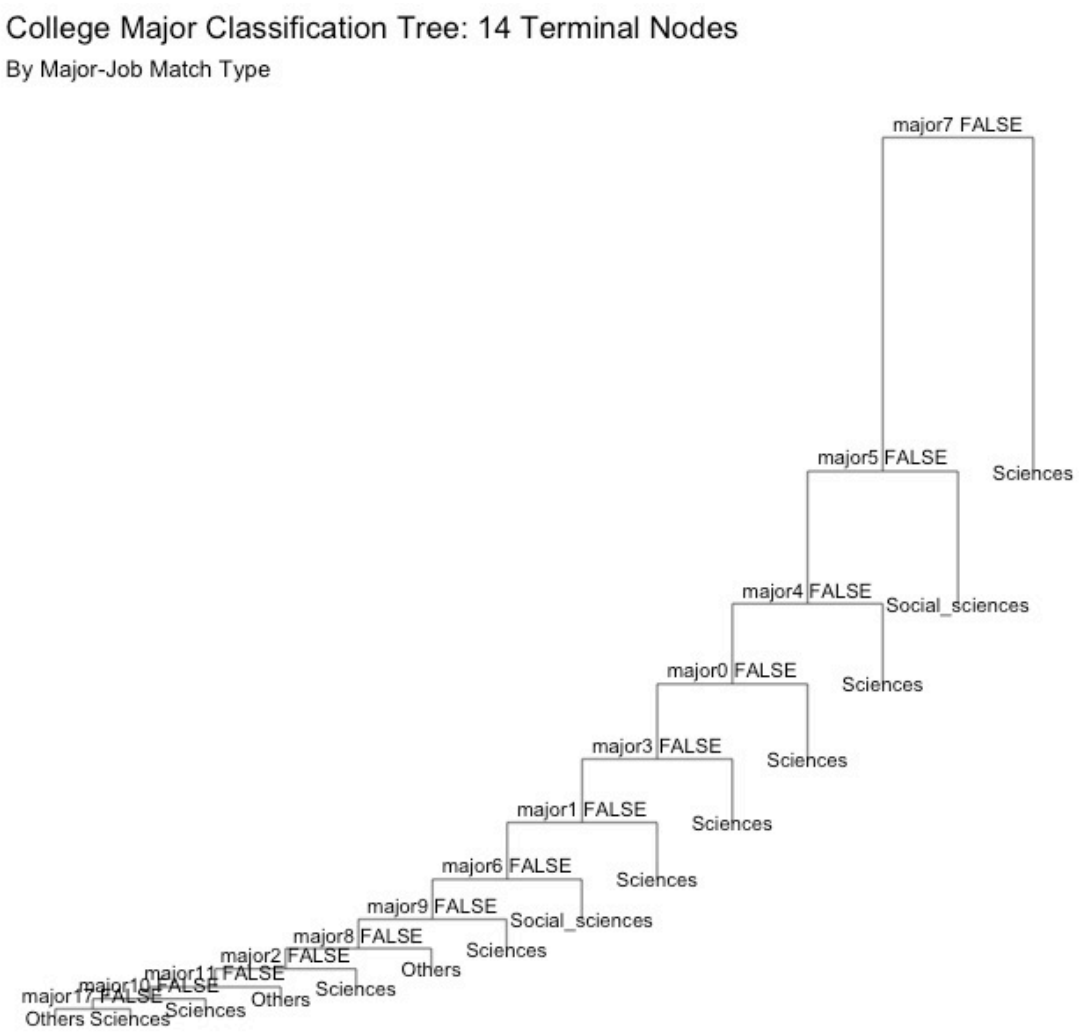

And I apply a 20-fold cross-validation for optimal number of nodes. The plot of test error rate shows that the optimal number of nodes would be 14. The test error rate would no longer decrease with the increase of numbers of tree nodes, after the number of nodes reaches 14.

Figure: 20-fold Cross-Validation for Optimal Number of Nodes




The tree plot with 14 terminal nodes is shown in the figure below. This provides better visualization than the plot of unpruned tree. Please refer to Appendix 2 for detailed classification of college major categories into 3 groups, by the major-job match type.

Figure: College Major Classification Tree (Optimal Number of Nodes: 14)




## 2. Hypotheses Testing: Logistic Regression

After aggregating college major categories into 3 groups, I apply logistic regression models to test the Hypothesis of Locational Dependence ($H_1$) and the Hypothesis of Influence of College Major ($H_2$). Model 1 is the baseline model, Model 2 is used to test $H_1$, and Model 3 is used to test $H_2$.

Table: Results of Logistic Regression on Whether Individuals Moved

| | Dependent variable: Whether Moved | | |
|---|---|---|---|
| VARIABLES | (1) | (2) | (3) |
| Age | 0.0166*** | 0.0162*** | 0.0172*** |
| Female | -0.173*** | -0.122*** | -0.0688*** |
| Black | 0.146*** | 0.027 | 0.0216 |
| Asian | 1.235*** | 0.0979*** | 0.0733** |

| | | | |
|---|---|---|---|
| Hispanic | 0.214*** | 0.152*** | 0.140*** |
| Stable Relationship | 0.298*** | 0.0893*** | 0.0873*** |
| Father's Years of Education | 0.0666*** | 0.0653*** | 0.0643*** |
| Mother's Years of Education | 0.00357 | 0.0515*** | 0.0507*** |
| Baseline: New England Region | | | |
| Middle Atlantic | | 0.0035 | 0.00481 |
| East North | | -0.167*** | -0.163*** |
| West North | | -0.182*** | -0.175*** |
| South Atlantic | | -0.528*** | -0.527*** |
| East South Central | | -0.0261 | -0.0187 |
| West South Central | | -0.562*** | -0.556*** |
| Mountain Region | | -0.121*** | -0.122*** |
| Pacific & US Territory | | -0.899*** | -0.904*** |
| Foreign Countries | | 6.291*** | 6.298*** |
| Majors with Technical Skills | | | 0.253*** |
| Majors with Social Sciences Skills | | | 0.201*** |
| Constant | -2.403*** | -2.878*** | -3.082*** |
| Pseudo R2 | 0.050 | 0.254 | 0.255 |
| Log Likelihood | -58110.072 | -45634.450 | -45549.698 |
| Observations | 89305 | 89305 | 89305 |
| *** p<0.01, ** p<0.05, * p<0.1 | | | |

Results of Model 2 confirm $H_1$, indicating that original location has impact on geographical mobility. I set original location as a categorical variable, and set the New England Region as the baseline. Compared to individuals from the New England Region, individuals from the Middle Atlantic Region and the East South Central Region do not have statistically significant different probabilities to move. But individuals from the East North Region, the West North Region, the South Atlantic Region, the West South Central Region, the Mountain Region and the Pacific & US Territory have lower probabilities to move, which is statistically significant on the level of 0.01. For example, the probability for people from the Pacific & US Territory Regions to move is only approximately 0.407 (since $\exp(0.899) \approx 0.407$) of the probability for people from the New England Region to move.

However, people from foreign countries have higher probability to move than people from the New England Region, which is statistically significant on the level of 0.01. The odds ratio is approximately 539.502 (since $\exp(6.291) \approx 539.502$). This is because the observations captured by this survey are in the US, and the subset from foreign countries must have moved from their original countries to the US.

Results of Model 3 have verified $H_2$, indicating that majors of college education have impact on whether geographical mobility occurs. I set the type of college majors

providing skills other than science/technology or social sciences as the baseline. Compared to the baseline type of majors, people with majors providing science/technology skills have approximately 1.287 (since exp(0.253) ≈1.287 times of probability to move. Still compared to the baseline type of majors, people with majors providing social sciences skills are 1.223 (since exp(0.201) ≈1.223) times of probability to move. Among the 3 types of college majors, people with the majors of science/technology skills have the highest probability to move.

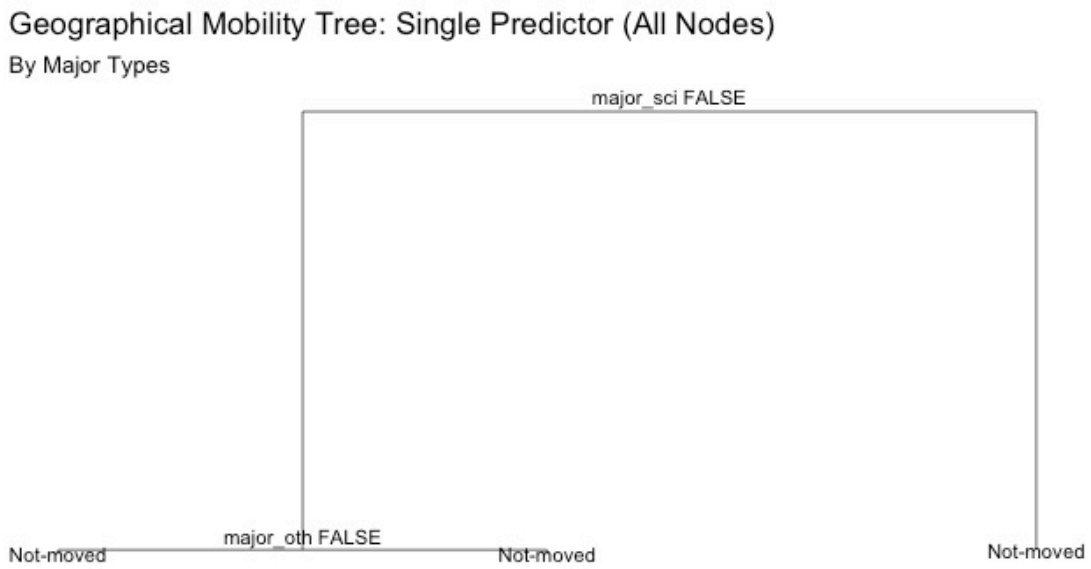
## 3. Prediction

Although logistic regression models have verified both of the proposed hypotheses, I apply decision tree models to check how well the college majors could be used in prediction of people's geographical mobility.

### 3.1 College Major as the Single Predictor

In this model, the dummy variable of whether people have moved is the response variable, and the variable of type of college majors is the single predictor. The result below shows no variance among different types of college majors. All nodes point to the same result. Thus, college major could not be used as the single predictor to geographical mobility.

Figure: Decision Tree of Geographical Mobility: Single Predictor (All Nodes)

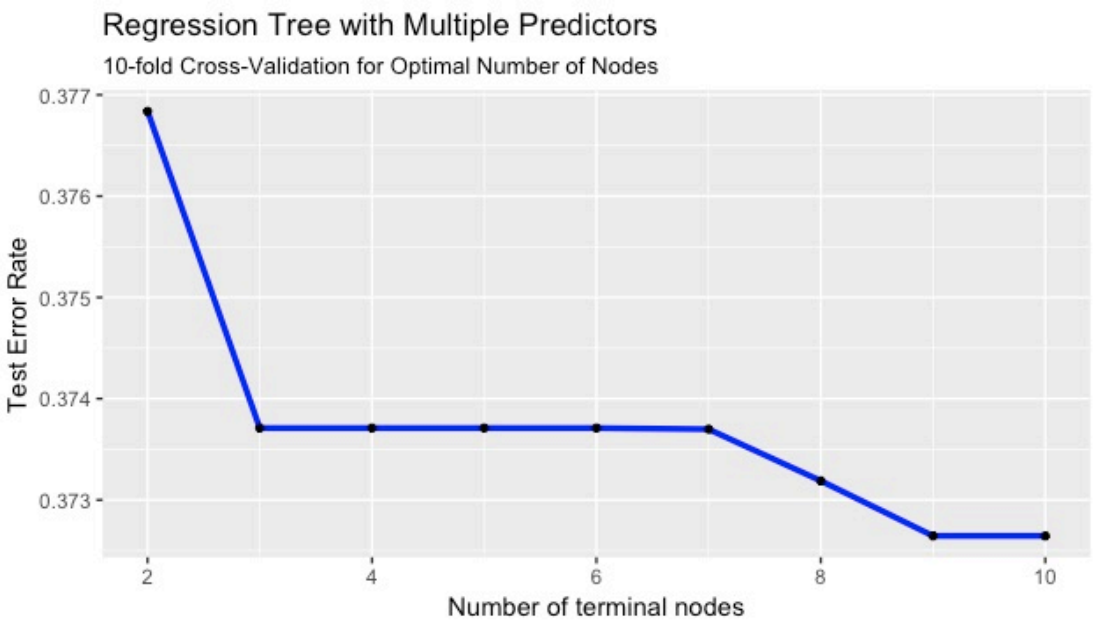


### 3.2 Multiple Predictors

In the decision tree model with multiple predictors, I add demographic features (age, gender, race) and the dummy variable regarding whether in a stable romantic relationship. Because of the limited maximum depth of tree, I have to omit variables of father's education, mother's education and original location.

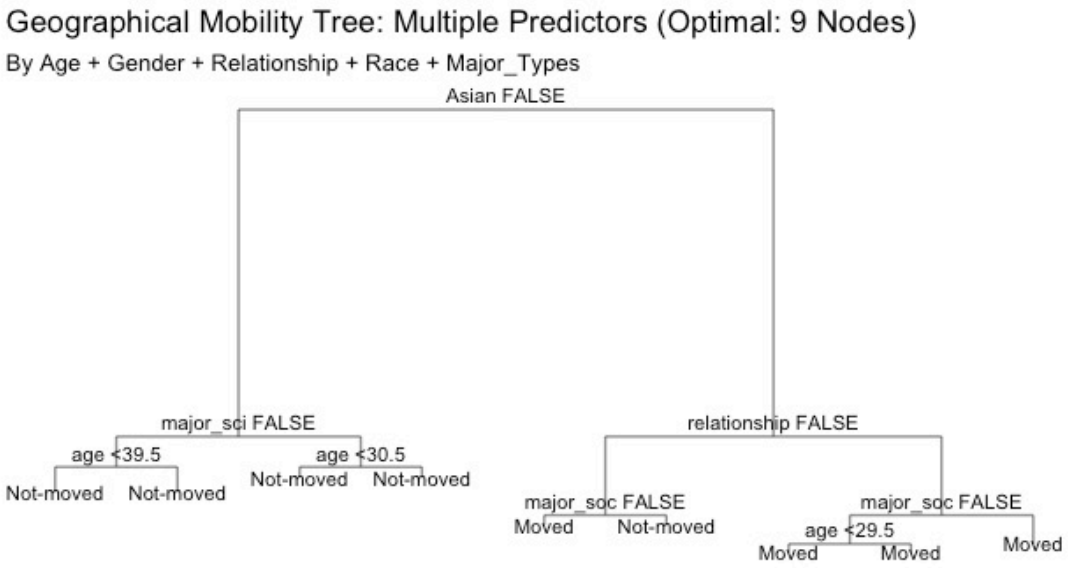
Since the tree plot with all nodes is too complex to read, I would not attach it on the paper. Instead, I apply 10-fold cross-validation to find optimal number of nodes.

Figure: 10-fold Cross-Validation for Optimal Number of Nodes




Results of the 10-fold cross-validation show the optimal number of nodes to be 9. With number of nodes as 9 and 10, the decision tree gets the lowest test error rate among the 10 models. And using 9 nodes instead of 10 nodes could omit an unnecessary node, which would makes the model cleaner.

Figure: Decision Tree of Geographical Mobility: Single Predictor (Optimal: 9 Nodes)




Results of the tree model with 9 terminal nodes show that, for prediction of geographical mobility, college major could be useful but its effect is not vital. For people who are not Asian, the prediction with different age and different majors would always go to not-moved. And for Asian, if people are in a stable romantic relationship, they always tend to have moved, no matter what age they are in and what major they are with. Only for Asian that is not in a stable relationship, if their major is in the group of social sciences, they have the tendency to have not moved, otherwise they have moved.

In brief, college major could be used to predict geographical mobility, but it needs to be used together with other predictors, such as race and relationship (or marital status). Statistical significance in its logistic regression model alone does not provide information for prediction.

## Conclusion & Discussion

In this study, I use the data of 2015 NSCG to analyze the relationship between college major and geographical mobility. Based on the research question regarding the effect of college major on geographical mobility, I come up with two hypotheses.

Three steps are taken in data analysis. Firstly, after manually aggregating the 142 fields into 20 categories on the basis of data documentation, I use decision tree to

classify the 20 college major categories into 3 types: majors with technical skills, majors with social sciences skills, and others. Then I use logistic regression for hypotheses testing. Results of logistic regression confirm both of the hypotheses. The probability for geographical mobility to occur could be influenced by the original places of individuals, which confirms $H_1$. Also, the probabilities for geographical mobility to occur are different for different types of college majors. For probabilities to gain geographical mobility, the descending order is: majors with science/technical skill, majors with social sciences skill, others. This confirms $H_2$. Finally, I use decision tree to check whether or how well college major could be used to predict geographical mobility. I find that college major could not be used as a single predictor. It only gains prediction power when it is used with age, gender, race and whether individual is in a stable romantic relationship. College major could be useful to prediction, but not vital.

However, there are three main limitations of the data. First, the geographic information used by this study is very limited, which not only prevents the use of spatial analysis tools but also has a negative impact on the accuracy of analysis. This is because the NSCG does not provide location in detail. It only provides regions, such as New England Region and East North Region, instead of specific location on state/county/city/community level. With the current data, I could only detect mobility from one region to another, but would miss mobility between smaller units, such as cities, within the same region. Second, for confidential concerns, the data does not provide information of the quality and locations of the colleges that the survey respondents attended. Not knowing the quality and locations of colleges, it is hard to get cleaner effect of college major. Third, the NSCG may not be representative to the population distribution, since it claims to have a special interest on young college graduates in science/technology labor force. This may negatively influence the power of prediction. If given more accurate and more representative data, the results could be more accurate and more would have better visualization.

There are several ways to improve this research in the future. First, it could take the dimension of time into consideration. With the development of technology, upgrade of industry and different stages of regional development, different cohorts may be faced with different labor market demands. So patterns might differ among cohorts. Second, more detailed analysis could have been applied. In the NSCG data,

some graduates have more than one higher education degrees, such as degrees with different majors or different levels. In the future, it would enrich the study by distinguishing them. Third, the criterion that is used to categorize college majors is not good enough. It is possible that college graduates with majors providing certain skills take jobs requiring different skills that they obtain from their minors or other activities.

In conclusion, this study is an attempt to analyze the relationship between college majors and geographical mobility. It uses the data of 2015 NSCG, combines logistic regression and decision tree, and finds college majors to be influential to geographical mobility under certain circumstances. Future study on the same topic could improve by using more accurate data, more advanced techniques, and better ways of classification.

**Reference**


Altonji, J. G., Arcidiacono, P., & Maurel, A. (2015). *The analysis of field choice in college and graduate school: Determinants and wage effects* (No. w21655). National Bureau of Economic Research.

Bernard, A., Bell, M., & Charles‐Edwards, E. (2014). Life‐Course Transitions and the Age Profile of Internal Migration. *Population and Development Review*, *40*(2), 213-239.

Berger, M. C. (1988). Predicted future earnings and choice of college major. *ILR Review*, *41*(3), 418-429.

Blau, P. M., & Duncan, O. D. (1967). The American occupational structure.

Bogue, D. J. (2009). Basics of contemporary US internal mobility and immigration. *Immigration, internal migration, and local mobility in the US. Northampton, MA, USA: Edward Elgar*, 1-30.

Bogue, D. J., Liegel, G., & Kozloski, M. (2009). Immigration, internal migration, and local mobility in the US. *Books*.

Bowles, S. (1970). Migration as investment: Empirical tests of the human investment approach to geographical mobility. *The Review of Economics and Statistics*, 356-362.

Eide, E., & Waehrer, G. (1998). The role of the option value of college attendance in college major choice. *Economics of Education review*, *17*(1), 73-82.

Geist, C., & McManus, P. A. (2008). Geographical mobility over the life course: Motivations and implications. *Population, Space and Place*, *14*(4), 283-303.

Goodrich, C., Allin, B. W., Brunck, H. D., Creamer, D. B., Hayes, M., & Thornthwaite, C. W. (1937). Migration and economic opportunity.

Greenwood, M. J. (1997). Internal migration in developed countries. *Handbook of population and family economics*, *1*, 647-720.

Grigg, D. B. (1977). EG Ravenstein and the "laws of migration". *Journal of Historical geography*, *3*(1), 41-54.

Groen, J. A. (2004). The effect of college location on migration of college-educated labor. *Journal of Econometrics*, *121*(1), 125-142.

Korinek, K., & Maloney, T. N. (Eds.). (2010). *Migration in the 21st century: rights, outcomes, and policy*. Routledge.

Long, L. (1988). *Migration and residential mobility in the United States*. Russell Sage Foundation.

Montmarquette, C., Cannings, K., & Mahseredjian, S. (2002). How do young people choose college majors?. *Economics of Education Review*, *21*(6), 543-556.

Ravenstein, E. G. (1876). *The birthplaces of the people and the laws of migration*. Trübner.

Ravenstein, E. G. (1885). The laws of migration. *Journal of the Statistical Society of London*, *48*(2), 167-235.

Ravenstein, E. G. (1889). The laws of migration. *Journal of the royal statistical society*, *52*(2), 241-305.

Robst, J. (2007). Education and job match: The relatedness of college major and work. *Economics of Education Review*, *26*(4), 397-407.

Shryock, H. S. (1964). *Population mobility within the United States* (Vol. 1). [Chicago]: Community and Family Study Center, University of Chicago.

Thompson, W. S. (1937). *Research Memorandum on Internal Migration in the Depression* (No. 30-31). Social science research council.

Thomas, D. S. T. (1938). *Research memorandum on migration differentials* (No. 43). Social Science Research Council.

Thomas, B., & Shryock, H. (1965). *The Milbank Memorial Fund Quarterly, 43*(2), 254-256. doi:10.2307/3349033

United States Census Bureau (2017). Website url: https://www.census.gov/topics/population/migration/about.html

Weber, A. F. (1899). *The growth of cities in the nineteenth century: A study in statistics* (No. 29). Columbia university.

Wolniak, G. C., Seifert, T. A., Reed, E. J., & Pascarella, E. T. (2008). College majors and social mobility. *Research in Social Stratification and Mobility*, *26*(2), 123-139.

## Appendix 1: Categories of College Major and Corresponding Fields of Study

| Appendix Table 1: Categories of College Major and Corresponding Fields of Study | |
|---|---|
| College Major Categories | Fields of study in the dataset |
| Computer sciences | Computer and information sciences<br>Computer science<br>Computer systems analysis<br>Information services and systems<br>Other computer and information sciences |
| Mathematics | Applied mathematics<br>Mathematics, general<br>Operations research<br>Statistics<br>Other mathematics |
| Agricultural sciences | Animal sciences<br>Food sciences and technology<br>Plant sciences<br>Other agricultural sciences |
| Biological sciences | Biochemistry and biophysics<br>Biology, general<br>Botany<br>Cell and molecular biology<br>Ecology<br>Genetics, animal and plant<br>Microbiological sciences and immunology<br>Nutritional sciences<br>Pharmacology, human and animal<br>Physiology and pathology, human and animal<br>Zoology, general<br>Other biological sciences |
| Physical sciences | Environmental science or studies<br>Forestry sciences<br>Chemistry, except biochemistry<br>Atmospheric sciences and meteorology<br>Earth sciences<br>Geology<br>Geological sciences, other<br>Oceanography<br>Astronomy and astrophysics<br>Physics<br>Other physical sciences |
| Psychology | Agricultural economics<br>Economics<br>Public policy studies<br>International relations<br>Political science and government |

| | |
|---|---|
| | Educational psychology<br>Clinical psychology<br>Counseling psychology<br>Experimental psychology<br>General psychology<br>Industrial/organizational psychology<br>Social psychology<br>Other psychology |
| Social sciences | Anthropology and archaeology<br>Criminology<br>Sociology<br>Area and ethnic studies<br>Linguistics<br>Philosophy of science<br>Geography<br>History of science<br>Other social sciences |
| Engineering | Aerospace, aeronautical and astronautical engineering<br>Chemical engineering<br>Architectural engineering<br>Civil engineering<br>Computer and systems engineering<br>Electrical, electronics and communications engineering<br>Industrial and manufacturing engineering<br>Mechanical engineering<br>Agricultural engineering<br>Bioengineering and biomedical engineering<br>Engineering sciences, mechanics and physics<br>Environmental engineering<br>Engineering, general<br>Geophysical and geological engineering<br>Materials engineering, including ceramics and textiles<br>Metallurgical engineering<br>Mining and minerals engineering<br>Naval architecture and marine engineering<br>Nuclear engineering<br>Petroleum engineering<br>Other engineering |
| Health/medical sciences | Audiology and speech pathology<br>Health services administration<br>Health/medical assistants<br>Health/medical technologies<br>Medical preparatory programs (e.g. pre-dentistry, -medical, -veterinary)<br>Medicine (dentistry, optometry, osteopathic, podiatry, veterinary)<br>Nursing (4 years or longer program)<br>Pharmacy<br>Physical therapy and other rehabilitation/therapeutic services |

| | |
|---|---|
| | Public health (including environmental health and epidemiology)<br>Other health/medical sciences |
| Engineering-related technologies | Computer teacher education<br>Mathematics teacher education<br>Science teacher education<br>Social science teacher education<br>Computer programming<br>Data processing<br>Electrical and electronic technologies<br>Industrial production technologies<br>Mechanical engineering-related technologies<br>Other engineering-related technologies |
| Agricultural business and production | Architecture/environmental design<br>Actuarial science<br>Other agricultural business and production |
| Business management/administrative services | Accounting<br>Business administration and management<br>Business, general<br>Business and managerial economics<br>Financial management<br>Other business management/administrative services |
| Education | Education administration<br>Counselor education and guidance services<br>Elementary teacher education<br>Physical education and coaching<br>Pre-school/kindergarten/early childhood teacher education<br>Secondary teacher education<br>Special education<br>Other education |
| Philosophy, religion, theology | Other philosophy, religion, theology |
| Foreign languages and literature | Social work<br>Business marketing/marketing management<br>Marketing research<br>English language, literature and letters<br>Other foreign languages and literature |
| Visual and performing arts | Liberal arts/general studies<br>History, other<br>Dramatic arts<br>Fine arts, all fields<br>Music, all fields<br>Other visual and performing arts |
| Communications | Communications, general<br>Journalism<br>Other communications |
| Natural resources and conservation | Other natural resources and conservation |
| Public affairs | Criminal justice/protective services<br>Home economics |

| | |
|---|---|
| | Law/prelaw/legal studies<br>Library science<br>Parks, recreation, leisure, and fitness studies<br>Public administration<br>Other public affairs |
| Others | Other fields (not listed) |

## Appendix 2: College Major Types and Corresponding College Major Categories

| College Major Types | College Major Categories |
|---|---|
| With Science/Technology Skills | Computer sciences |
| | Mathematics |
| | Agricultural sciences |
| | Biological sciences |
| | Physical sciences |
| | Engineering |
| | Engineering-related technologies |
| | Agricultural business and production |
| | Natural resources and conservation |
| With Other Skills | Health/medical sciences |
| | Business management/administrative services |
| | Education |
| | Philosophy, religion, theology |
| | Foreign languages and literature |
| | Visual and performing arts |
| | Communications |
| | Public affairs |
| | Others |
| With Social Sciences Skills | Psychology |
| | Social sciences |