Selective Migration, Occupational Choice, and the Wage Returns to College Majors*

Job Market Paper

for most recent version please visit

http://sites.duke.edu/transom/files/2016/10/roymajors.pdf

Tyler Ransom†

Duke University

December 27, 2016

Abstract

I examine the extent to which the returns to college majors are influenced by selective migration and occupational choice across locations in the US. To quantify the role of selection, I develop and estimate an extended Roy model of migration, occupational choice, and earnings where, upon completing their education, individuals choose a location in which to live and an occupation in which to work. In order to estimate this high-dimensional choice model, I make use of machine learning methods that allow for model selection and estimation simultaneously in a non-parametric setting. I find that OLS estimates of the returns to business and STEM majors relative to education majors are upward biased by 15% on average and by as much as 30%. Using estimates of the model, I characterize the migration behavior of different college majors and find that migration flows are twice as sensitive to occupational concentration as they are to wage returns. This finding has important implications for local governments seeking to attract or retain skilled workers.

JEL Classification: I2, J3, R1

Keywords: College major, migration, occupation, Roy model

*I would like to thank Peter Arcidiacono, Esteban Aucejo, Pat Bayer, Rob Garlick, Arnaud Maurel, and seminar participants at Cal State Fullerton, Duke Economics, Duke SSRI, Oklahoma State, the US Naval Academy, the 1st IZA Workshop on the Economics of Education: Higher Education, and the 2016 SEA Meetings (Washington, DC) for their helpful discussions and comments. Special thanks to Jamin Speer for generously providing helpful code for classifying college majors in the ACS. All errors are my own.

†Contact: Social Science Research Institute, Duke University, Box 90989, Durham, NC 27708-0989. E-mail: *tyler.ransom@duke.edu*

1 Introduction

One lesser-known characteristic of the US labor market is that the wage returns to different college majors are highly heterogeneous across space. For example, among males in the 2010-2015 American Community Survey, the return to STEM and business majors each range from 23% to 43%, relative to education majors.[1] While much work has examined sorting of majors into occupations, occupational sorting does little to narrow this gap: the return to a STEM major in a STEM occupation relative to a STEM major in a non-STEM occupation ranges from 10% in Texas to 26% in Oregon, with a similar range for other majors.[2] This broad range in returns to majors and occupations suggests that post-college migration, and in particular its interaction with post-college occupational choice, might be a significant driver of the observed range in returns.

The objectives of this paper are (i) to uncover the extent to which selection into residence location and occupation biases the observed wage returns to college majors (relative to education majors); and (ii) to assess the extent to which migration flows respond to cross-location differences in wage returns, occupational availability, and local amenities. This is the first paper to examine the spatial dimension of college major and occupation decisions, and the first to examine how the interaction of the two influences migration flows.[3] In doing so, I find that correcting for selection tends to reduce the measured returns by up to 30%. I also find that migration of college majors is influenced twice as much by occupational density as it is by wage returns.

It is important to know how college graduates make post-college migration decisions. The answer is of interest not only to students who select their major, but also to local governments who seek to attract and retain a skilled workforce. Certain majors may be more likely to move out of the state from which they graduated, possibly in search of employment in a related occupation. If this is the case, then a state government seeking to retain its college graduates could respond in two ways: (i) increasing the tuition rate of the majors that are more likely to leave; or (ii) increasing the density of occupations related to the majors that are more likely to leave. Knowing how sensitive graduates of specific majors are to occupation relatedness can inform the effectiveness of such policies.

Using data on male college graduates from the 2010-2015 American Community Survey, I document substantial differences in earnings, occupational choice, and locational choice across

[1] Returns calculated using a Mincerian regression of log earnings on a cubic in potential experience, demographic indicators, and MSA fixed effects.

[2] For overviews of the literature on college major choice and consequences, including post-college occupational choice, see Altonji, Blom, and Meghir (2012) and Altonji, Arcidiacono, and Maurel (2016a). Lemieux (2014), Kinsler and Pavan (2015), and Altonji, Kahn, and Speer (2016b) examine the effect of occupational choice on major earnings premiums. Each of these studies abstracts from location.

[3] Winters (Forthcoming) is the only paper in the literature analyzing the migration behavior of college majors. He examines the migration response of different college majors to birth-state earnings shocks to workers in the same major.

college majors. These differences provide support for the existence of different location-occupation complementarities for different majors. As an example, I show that STEM and business majors earn the highest returns to and are much more likely to work in occupations related to their major. However, business majors are much less likely to live outside their state of birth. These results are consistent with a model where college graduates have preferences for working in an occupation related to their field of study, but where occupational concentration varies across space.

Additional evidence on the importance of location and occupation for college majors can be seen by examining flows between specific locations. For example, education majors who originate in New York are highly unlikely to work as teachers in New York unless they hold a master's degree. As a result, there is a large outflow of bachelor's-level education majors from New York to areas where working as a bachelor's-level teacher is more common, but where the wage returns to doing so are much lower. Migration flows such as these show that non-wage factors, specifically related occupation availability, are potentially strong determinants of the observed returns to college majors.

One would expect selection to result in naïve estimates being upward biased if certain majors are more prone to migrate or choose a particular occupation in response to favorable wage shocks. On the other hand, naïve estimates may be downward biased if certain majors have strong non-wage preferences for a particular location or occupation. Estimating the direction and magnitude of this bias is one of the primary empirical questions of this paper.

To account for the various factors described above, I estimate an extended Roy (1951) model that allows for nonpecuniary tastes in both the location and occupation dimensions.[4] The model divides occupations for each major into related and unrelated, and divides the United States into 15 groups of states. This paper bridges previous work that has examined the role of selective migration on the wage returns to a college degree (Dahl, 2002; Bayer, Khan, and Timmins, 2011) and the role of selective occupational choice on the returns to college major (Lemieux, 2014; Kinsler and Pavan, 2015).

Estimation of an extended Roy model is difficult in a model with nonpecuniary preferences and many choice alternatives. To estimate the model, I implement methods pioneered by Lee (1983) and Dahl (2002) which show that a control function approach, where the control function includes a polynomial of a small number of observed choice probabilities, is able to account for a variety of patterns in selection.[5] This polynomial serves as a multidimensional analog of the inverse Mill's ratio in the classic Heckman (1979) correction model. As a result, the researcher can obtain unbiased and consistent estimates of the selection-corrected returns using OLS. With the selection-corrected returns in hand, I then examine the responsiveness of migra-

[4]For surveys on the Roy model and its empirical content, see Heckman and Vytlacil (2007a,b) and French and Taber (2011).

[5]The assumption that a small number of probabilities can form a sufficient statistic for selection is referred to by Dahl as the *index sufficiency assumption*.

tion flows to spatial differences in wage returns, occupational density, and non-wage amenities.

I implement the Lee and Dahl approach with a machine learning method that shows the usefulness of machine learning in economics, in the spirit of Varian (2014) and Bajari et al. (2015). The specific method used in this paper is the conditional inference classification tree. While existing methods have utilized nonparametric bin estimation to derive selection probabilities, tree classification of this type has the advantage of using statistical hypothesis tests to determine which covariates should be included, and where bin cut points should be made. The algorithm is especially useful in settings where it would be infeasible to include all covariates. I assess the performance of the classification tree relative to classical econometric estimators and show that it performs better both in simulations and in practice.

Using these empirical methods, I find that OLS estimates of the returns to college majors (relative to education majors) are upward biased. Correcting for selective migration and occupational choice tends to lower the measured returns, by up to 30% in some locations and consistent with other studies (Dahl, 2002; Bayer et al., 2011). The bias is the strongest among business and STEM majors who hold advanced degrees. Controlling for selection does not narrow the spatial gaps in measured returns by very much. These findings imply that spatial dispersion in the returns to major is likely primarily due to innate productivity differences or compensating differentials.

With the corrected returns to major in hand, I analyze the determinants of migration flows for different majors. I find that, in addition to differences in the wage returns to major, migration flows for all majors are responsive both to the availability of occupations related to the major, and to non-wage amenities such as distance, weather, and local government characteristics. Surprisingly, the elasticity of migration with respect to occupational density is more than twice the elasticity with respect to earnings.

The findings of this paper have important implications for local governments seeking to attract or retain skilled workers. Specifically, the results highlight the importance of employment in related occupations as a means of attracting college-educated workers. For example, state governments who enact tuition subsidies that are geared towards certain majors may not be able to retain those students if there is not a sufficient density of occupations related to those majors in that location.[6] Moreover, as discussed in Moretti (2012) and Kline and Moretti (2014), the success of place-based policies is not guaranteed and often comes at significant cost. One potential solution could be to offer different tuition by major that is indexed to the local concentration of related occupations.

The remainder of the paper is organized as follows: Section 2 details the Roy model which serves as the empirical basis of understanding selection. Section 3 outlines the statistical framework that allows me to reduce the dimensionality of the choice set. Section 4 describes the data construction and key variables used in the estimation, and Section 5 discusses the estimation of the model, including the non-parametric machine learning decision tree algorithm. Section 6

[6]For further discussion on the implementation of major-specific tuition rates, see Stange (2015).

discusses the main empirical findings, and Section 7 concludes.

2 A Roy Model of Migration, Occupation, and Earnings

In this section, I introduce an extended Roy (1951) model of college major, occupational choice, and locational choice, using the framework developed in Dahl (2002).[7] It extends Roy's original model in two ways: (i) both pecuniary and nonpecuniary factors influence an individual's decision; and (ii) there are more than two alternatives in the choice set.[8]

The focus of this paper is on how selective migration and occupational choice in the United States affects the measured returns to the human capital investment of college major. The objective is to examine how sensitive earnings in a particular major are to post-college location and occupational choice. Existing models in the literature on college major and occupation have treated location as fixed (Lemieux, 2014; Kinsler and Pavan, 2015; Ransom and Phipps, Forthcoming). At the same time, there is strong evidence that location is an increasingly important determinant of labor market outcomes, particularly for the college educated (Moretti, 2012; Diamond, 2016). This paper serves to fill the gap between these two literatures.

An extended Roy model serves as an appropriate lens through which to view the joint location and occupation decisions of college graduates because it allows for the inclusion of nonpecuniary components. Factors such as amenities and distance have been shown to be important determinants of migration decisions (Kennan and Walker, 2011; Ransom, 2016; Zabek, 2016), while nonpecuniary considerations have also been shown to be important to occupational choice among college graduates (Arcidiacono et al., 2014).

2.1 Model

This section formalizes each component of the Roy model and how each of the components interact with each other. The primary components of the model are earnings (the outcome equation) and preferences (the selection equation). In contrast with most of the Roy model literature, this paper emphasizes the empirical results of the outcome equation as opposed to the selection equation. As such, it is appropriate to view the model as a reduced-form approximation of a Roy model because I make no attempt to structurally model the selection equation.

The framework of the model is as follows. A geographical area (e.g. the United States) is divided into L mutually exclusive locations (e.g. groups of states). The model has two periods.

[7] The Roy model has also been used in the migration literature by Borjas (1987) and Falaris (1987), among others.

[8] See Heckman and Taber (2008) for an overview of the original Roy (1951) model and its various extensions. Heckman and Honoré (1990) discusses identification of the Roy model, including the assumptions on the distribution of earnings that are required to generate empirical content of the Roy model. D'Haultfœuille and Maurel (2013) perform inference on an extended Roy model of schooling decisions in France. Eisenhauer et al. (2015) discuss how to use the generalized Roy model to separately identify costs and benefits of treatment.

In the first period, individuals are born and make human capital investment decisions. In the second period, individuals choose where to live and in which occupation to work, and receive utility from both earnings and nonpecuniary aspects of the chosen location and occupation.[9]

2.1.1 Earnings

The potential log annual earnings for individual i residing in location ℓ and working in occupation k are given by the following equation:

$$w_{i\ell k} = x_i\gamma_{1\ell k} + s_i\gamma_{2\ell k} + \eta_{i\ell k}, \quad \ell = 1,\ldots,L\ ,\ k = 1,\ldots,K \tag{2.1}$$

where x_i is a vector of individual characteristics and s_i is an S-dimensional vector of dummy variables indicating i's college major and advanced degree attainment. The parameter of interest in (2.1) is $\gamma_{2\ell k}$, which measures the link between earnings, college major, and location and occupational choice. However, because $\eta_{i\ell k}$ is only observed in the chosen (ℓ, k) combination, and because the chosen (ℓ, k) is the result of a non-random selection process, OLS estimates of $\gamma_{1\ell k}$ and $\gamma_{2\ell k}$ will be biased. I next outline the preferences of individuals, which govern the selection process.

2.1.2 Preferences

Individuals have preferences for both earnings and nonpecuniary factors:

$$V_{ij\ell k} = w_{i\ell k} + u_{ij\ell k}, \quad \ell = 1,\ldots,L\ ,\ k = 1,\ldots,K \tag{2.2}$$

where j indexes birth location, ℓ indexes current location, and k indexes occupation. $u_{ij\ell k}$ encompasses all nonpecuniary utility components that could determine the utility of residing in location ℓ and working in occupation k given origin j. These include location characteristics such as climate, crime, commuting time, distance from j, geographical and cultural amenities, and many others. Also included are occupational characteristics such as working conditions, relevance to previous human capital investments, coincidence with personal preferences, and flexibility of hours, among many others.

[9] The choice to model location and occupation as once-and-for-all decisions is primarily due to data limitations: longitudinal surveys containing data on college major, location, and occupation do not have sufficient sample size to allow for meaningful estimation of location-specific outcomes. Work by Kennan (2015) examines the interaction between migration and college completion in a dynamic setting using the National Longitudinal Survey of Youth 1979 (NLSY79), but is unable to capture heterogeneity across majors because of data limitations.

Preferences can be rewritten as follows:

$$\begin{aligned} V_{ij\ell k} &= \underbrace{\mathbb{E}\left[w_{i\ell k} \mid x_i, s_i\right] + \mathbb{E}\left[u_{ij\ell k} \mid z_i\right]}_{v_{j\ell k}} + \underbrace{\eta_{i\ell k} + \varepsilon_{ij\ell k}}_{e_{ij\ell k}} \\ &= v_{j\ell k} + e_{ij\ell k} \end{aligned}$$

where z_i is a vector of individual characteristics that affect preferences, $\eta_{i\ell k}$ represents measurement error in earnings, and $\varepsilon_{ij\ell k}$ represents preference shocks for choosing to live in ℓ and work in occupation k given birth location j. $v_{j\ell k}$ is referred to as either the subutility function (in the selection literature) or the conditional value function (in the dynamic discrete choice literature).[10]

2.1.3 Utility maximization

Individuals maximize utility such that

$$d_{ij\ell k} = 1\left[v_{j\ell k} + e_{ij\ell k} \geq v_{jmn} + e_{ijmn} \;\forall\, (m,n) \neq (\ell,k)\right] \tag{2.3}$$

where $1[A]$ is an indicator variable that takes a value of 1 when condition A is true and 0 otherwise. (2.3) emphasizes that utility depends not only on the location of residence, but also on the deterministic and stochastic elements of utility in *each* location, including the location of birth. Furthermore, earnings are observed only in the location that is selected:

2.1.4 Selection rule

The selection rule is given by

$$w_{i\ell k} \text{ observed} \iff d_{ij\ell k} = 1 \tag{2.4}$$

Specifically, earnings are only observed if all L selection equations in (2.3) are simultaneously satisfied. Thus, individuals observed to reside in ℓ are not a random sample of the population;

[10] The model assumes that individuals have no uncertainty regarding their earnings or tastes in other locations. While it is possible to allow for imperfect information, doing so would require, e.g. assuming that the individual's information set is shared by the econometrician. On the other hand, the approach taken here to model migration in response to individual earnings shocks departs from much of the migration literature, which assumes that migration decisions are influenced by the deterministic portion of earnings (Kennan and Walker, 2011; Bishop, 2012; Ransom, 2016). This assumption is typically made for tractability of dynamic models.

hence

$$\begin{aligned}\mathbb{E}\left[\eta_{i\ell k} \mid w_{i\ell k} \text{ observed}\right] &= \mathbb{E}\left[\eta_{i\ell k} \mid d_{ij\ell k} = 1, x_i, z_i\right] \\ &= \mathbb{E}\left[\eta_{i\ell k} \mid e_{ijmn} - e_{ij\ell k} \leq v_{j\ell k} - v_{jmn},\ \forall (m,n) \neq (\ell,k)\right] \\ &\neq 0\end{aligned} \tag{2.5}$$

where $\mathbb{E}\left[\eta_{i\ell k} \mid \cdot\right]$ is the selectivity bias for i.

Equations (2.1) through (2.5) comprise an extended Roy model of earnings, migration, and occupational choice.

Unfortunately, this extended Roy model is difficult to estimate without making additional assumptions about how the subutility functions affect the selection term (i.e. the conditional expectation in (2.5)). There are two reasons for this: (i) the number of locations L needs to be sufficiently large in migration models in order to accurately reflect the actual choice set faced by individuals, thus effecting the curse of dimensionality; and (ii) individuals derive utility from both earnings and nonpecuniary aspects of the location, meaning that the researcher is required to account for individual preferences. The problem with the latter reason is that there are a large number of variables that are important factors in the nonpecuniary dimension, but which are unobserved or poorly measured.

In the next section, I explain how I avoid these issues by implementing existing estimation methods (Lee, 1983; Dahl, 2002) which are designed to circumvent parametric estimation of the subutility functions, and which work well on choice sets that are otherwise prohibitively large.

3 Reducing the Dimensionality of the Problem

Estimating the problem described in Section 2 is infeasible without making additional assumptions. The difficulty arises out of the curse of dimensionality due to the large set of locations and occupations in which a person can choose to live and work. In this section, I provide intuition and a brief formal derivation on how to feasibly estimate the aforementioned extended Roy model. I also informally discuss how the model is identified. The key point is that I follow the strategy developed by Lee (1983) and refined by Dahl (2002) to express the selection in the earnings equation as a function of a small number of observed choice probabilities.

3.1 Overview

The intuition of this approach is as follows: examining equations (2.3) and (2.4) reveals that the probability of observing an individual's earnings in location ℓ and occupation k is related to the probability that $V_{j\ell k}$ is the maximum of all subutility functions. Thus, the joint distribution between the error term in the earnings equation ($\eta_{i\ell k}$) and the differenced subutility error terms ($e_{j11} - e_{jmn}, \ldots, e_{jLK} - e_{jmn}$) can be reduced from $L \times K$ dimensions to two dimensions:

the first dimension is the earnings error and the second is the maximum order statistic of the differenced subutility functions. The key assumption is that this bivariate distribution does not depend on the subutility functions themselves, except through a small number of choice probabilities.[11] This allows the researcher to express the selection correction term in the earnings equation (analogous to the inverse Mills ratio term in the canonical Heckman selection model) as a function of a small number of observed choice probabilities. Without this assumption, the researcher would be required to estimate an $(LK-1)$-dimensional integral. This becomes quickly infeasible as L grows large, as is the case in the current setting.

3.2 Technical details

To aid the exposition, I now briefly formalize the above intuition. Readers interested in a full derivation should consult Dahl (2002) and Lee (1983).

First consider a reformulation of (2.3) and (2.4):

$$\begin{aligned} w_{i\ell k} \text{ observed} &\iff v_{j\ell k} + e_{ij\ell k} \geq v_{jmn} + e_{ijmn} \;\forall (m,n) \neq (\ell,k) \\ &\iff \left(v_{j11} - v_{j\ell k} + e_{ij11} - e_{ij\ell k}, \ldots, v_{jLK} - v_{j\ell k} + e_{ijLK} - e_{ij\ell k}\right)' \leq 0 \\ &\iff \max_{m,n} \left(v_{jmn} - v_{j\ell k} + e_{ijmn} - e_{ij\ell k}\right) \leq 0 \end{aligned} \tag{3.1}$$

Now consider the joint density of the earnings error term and the max of the subutility differences, evaluated at the error realizations. We have the following one-to-one mapping between the LK-dimensional density $f_{j\ell k}$ and the two-dimensional density $g_{j\ell k}$. This mapping is made possible by implementing maximum order statistics (see Lee, 1983):

$$\begin{aligned} & f_{j\ell k}\left(\eta_{i\ell k}, e_{ij11} - e_{ij\ell k}, \ldots, e_{ijLK} - e_{ij\ell k}\right) \\ = \; & g_{j\ell k}\left(\eta_{i\ell k}, \max_{m,n}\left(v_{jmn} - v_{j\ell k} + e_{ijmn} - e_{ij\ell k}\right) \mid v_{j11} - v_{j\ell k}, \ldots, v_{jLK} - v_{j\ell k}\right) \end{aligned} \tag{3.2}$$

where the expression for $g_{j\ell k}$ in (3.2) is written as being conditional on the differences in the subutility functions in order to emphasize this dependence.

In order to reduce the dimensionality of $g_{j\ell k}(\cdot)$, Dahl proposes an *index sufficiency assumption* as follows:

$$\begin{aligned} & g_{j\ell k}\left(\eta_{i\ell k}, \max_{m,n}\left(v_{jmn} - v_{j\ell k} + e_{ijmn} - e_{ij\ell k}\right) \mid v_{j11} - v_{j\ell k}, \ldots, v_{jLK} - v_{j\ell k}\right) \\ = \; & g_{j\ell k}\left(\eta_{i\ell k}, \max_{m,n}\left(v_{jmn} - v_{j\ell k} + e_{ijmn} - e_{ij\ell k}\right) \mid p_{ij\ell k}, p_{ijmn}\right) \end{aligned} \tag{3.3}$$

where $p_{ij\ell k}$ and p_{ijmn} are two probabilities that are readily observed in the data. I discuss later

[11] Dahl (2002) refers to this assumption as the *index sufficiency assumption*, which I discuss below in more detail.

how to choose these probabilities. The implicit assumption in (3.3) is that the probabilities $p_{ij\ell k}$ and p_{ijmn} contain all of the information about how the index of subutility functions influences the joint distribution of the earnings error term and the maximum of the subutility errors.

Applying the assumption in (3.3) to the earnings equation gives the following corrected earnings equations that account for selective migration and occupational choice, and that are feasibly estimated:

$$w_{i\ell k} = x_i\gamma_{1\ell k} + s_i\gamma_{2\ell k} + \sum_{j=1}^{L} d_{ij\ell k}\lambda_{j\ell k}\left(p_{ij\ell k}, p_{ijmn}\right) + \omega_{i\ell k}, \tag{3.4}$$

The implication of the assumption in (3.3) is that $\mathbb{E}\left[\omega_{i\ell k} \mid x_i, s_i, p_{ij\ell k}, p_{ijmn}, d_{ij\ell k} = 1\right] = 0$, meaning that the selection problem has been resolved. Note also that the index sufficiency assumption reduces the dimensionality of the selection correction functions from LK, LK-dimensional control functions to LK bivariate control functions.

Because index sufficiency is an assumption, it is important to recognize the restrictions that it levies. Index sufficiency holds, for example, if earnings errors are composed of an individual fixed effect that is invariant to the location of residence. On the other hand, this assumption is less likely to hold in a setting where, for example, an individual's fixed effect on earnings could vary with location. I discuss in Appendix A the results of Monte Carlo simulations that show that this assumption holds for a variety of scenarios.

In Section 5, I discuss details of the estimation of equation (3.4) including how to estimate the probabilities of interest, and how to estimate the unknown correction functions $\lambda_{j\ell k}$, including additional assumptions made to reduce the number of control functions entering (3.4).

3.3 Identification

I now informally discuss how the model is identified. As discussed in other implementations of the Roy model (Dahl, 2002; D'Haultfœuille and Maurel, 2013; Bayer et al., 2011), separately identifying nonpecuniary preferences from earnings in most cases requires an exclusion restriction—a covariate which appears in the choice probabilities but does not affect wages.

Crucial to identification in this model is the existence of two such exclusion restrictions: one for locational choice and one for occupational choice. I use two related exclusion restrictions inspired by Kinsler and Pavan (2015). To separately identify preferences for location from earnings, I use the fraction of demographically similar (including college major and advanced degree status) individuals from the same birth state who stayed in their birth state, net of the national rate of staying. To separately identify preferences for occupation from earnings, I compute a similar number, but instead calculate the share who choose to work in an occupation related to their major.

The ideal exclusion restriction for location or occupational choice would be an adequate

measure of search frictions. The rationale for this is as follows: individuals have preferences for a certain location or occupation, but are unable to secure employment in the preferred alternative because there are not enough vacancies. While not a perfect measure of search frictions, the proposed exclusion restriction recovers a reduced-form approximation of such.

Another advantage of using the above exclusion restrictions is that it allows me to include birth location directly into the wage equation. Previous literature has shown that certain locations do a better job of educating their residents, which implies that stayers in those locations may receive higher wages than movers (Card and Krueger, 1992; Heckman, Layne-Farrar, and Todd, 1996; McHenry, 2011). Allowing stayers to earn different wages than movers improves on the previous approaches of Dahl (2002) and Bayer et al. (2011) which require birth location to be excluded from wages. Finally, research by Zabek (2016) finds that there is substantial heterogeneity across states in the fraction of people who reside in their state of birth. This result gives further credence to the exclusion restriction explained above.[12]

In addition to the peer share exclusion restrictions, I also allow distance moved and other demographic characteristics to influence the nonpecuniary portion of utility. Specifically, these covariates are: an indicator for birth location in the same Census region as the location of residence, and separate indicators for each of the following: co-residence with a family member, spouse's work status (if applicable), spouse born in residence location, and presence of children aged 0-4 or 5-18. In results not shown, I find that these demographic characteristics have much less predictive power in the first stage than the two primary exclusion restrictions.[13]

The primary threat to the validity of these exclusion restrictions is if the location or occupation decision of the demographic cell is driven by advantageous draws from the earnings distribution in the home location or in a certain occupation. This is unlikely to be a strong driver of decisions because there is a strong correlation across majors within a given state in the propensity to stay in that state. Thus, propensity to stay in the state of birth appears to be driven more by nonpecuniary factors.

[12]The exclusion restriction rests on the assumption that certain states retain their natives at higher frequencies than others for purely idiosyncratic reasons. For example, Texas is the "stickiest" state, retaining 77% of its natives. On the other hand, Wyoming is the least sticky, retaining just 37% of its natives. Stickiness is positively correlated with state population, but not very strongly ($\rho = 0.45$, rank correlation = 0.72), indicating that this is not simply a mechanical relationship. Finally, stickiness is strongly correlated across majors within state, indicating that preferences for staying in one's state of birth have more to do with nonpecuniary factors.

[13]For example, I estimate separate linear probability models for moving out of one's birth location and for working in a related occupation. The first-stage F-statistic for leaving the birth location is 51,386 ($R^2 = .081$) when including only the migration peer share variable and 21,150 ($R^2 = .1533$) when including all other excluded variables, with spousal birth location accounting for all of the latter's explanatory power. For related occupation, the F-statistic is 28,349 ($R^2 = .046$) for the occupation peer share variable and 413 ($R^2 = .003$) for the other excluded variables.

4 Data and Descriptive Analysis

I now discuss the data used in the estimation procedure. I also present a descriptive analysis of the data trends which, when compared with the model estimates, will be used to quantify the amount of selection in migration and occupation decisions.

4.1 Data

I use data from the American Community Survey (ACS) as compiled by Ruggles et al. (2015) over the years 2010-2015. The ACS is an annual stratified random sample of 1% of US households produced by the US Census Bureau. Sampled households respond to the survey either on paper or via the internet, and non-responding households receive a follow-up telephone call or visit by a Census employee.

The ACS collects detailed data for each adult household member on income, employment, education, demographic characteristics, and health. It also collects information about the household, such as household and family structure and housing unit characteristics. In this analysis, I focus on the following variables: location of birth, location of residence, demographic characteristics (e.g. age, gender, race, ethnicity, household composition), college major, advanced degree attainment, occupation, and earnings.[14]

The analysis sample consists of all native-born males between the ages of 22 and 54 with at least a bachelor's degree, and who have observed earnings within a reasonable range, who have observed college major, who are not in school, do not live in group quarters, and who do not have imputed values for any of the variables of interest. This corresponds to a 6% sample of the US population for this subgroup. The estimation sample of the data comprises 583,913 individuals. Details on the number of observations deleted with each criterion are listed in Table B1.

4.1.1 Definitions of majors, occupations, and locations

I now discuss aggregation of majors, occupations, and locations in order to preserve tractability in estimation.

Majors I aggregate majors into five categories, crossed with advanced degree status so that s_i in equation (2.1) is a 10-dimensional vector. The ACS records hundreds of distinct college major fields following the Classification of Instructional Programs (CIP) established by the National Center for Education Statistics (NCES). In order to focus the analysis and to maintain statistical power, I aggregate majors into groups with similar pre- and post-graduation outcome character-

[14]Information on college major began to be collected in 2009. I focus on the years 2010-2015 in order to maximize sample size while avoiding the most severe part of the Great Recession, because previous work has shown that migration is sensitive to business cycle conditions (Molloy and Wozniak, 2011; Ransom, 2016).

istics. The set of aggregated majors is: education, social sciences, business, STEM, and all others. A detailed mapping of the 51 Department of Education major fields to these five aggregated fields is provided in Table B2. Notably, the business field includes economics majors and the STEM field includes pre-med majors.

Occupations I define occupation as having two values: *related* or *unrelated* (i.e. $K = 2$). An occupation is related to a major if it is reported to have a 2% or larger share of all 3-digit occupation codes within a detailed definition of major (i.e. the 51 Department of Education codes).[15] The set of occupations that are related to an aggregated major category is then the union of the set of related occupations for each of the detailed majors corresponding to the aggregate. I allow the set of related occupations to differ based on advanced degree status.

The cutoff of 2% was chosen so as to ensure that highly specialized majors (i.e. majors with high concentration in few occupations) would have their most concentrated occupations defined as related. To provide further intuition for this approach, I present in Figure 1 the frequency distribution of occupations (sorted from most to least frequent) for non-advanced-degree holders in four majors: primary education, history, economics, and computer programming. For each panel of the figure, I include a vertical line along with the frequency distribution, which serves to mark the cutoff between related and unrelated occupations. Figure 1 shows that the primary education and computer programming majors are highly specialized, with 30%-40% of majors working in the most common occupation (elementary school teachers and software developers, respectively). Furthermore, computer programming majors are observed in many fewer occupations than the other majors included in the figure, by a factor of four. On the other hand, economics and history majors do not have clear-cut occupations corresponding to them, as the most frequent occupation contains only 10% of majors (miscellaneous managers for both). Figure 2 reports the same information but for advanced degree holders only. The results are similar. The exact occupation titles that are related to each of these majors are listed in Tables B3 and B4 respectively by advanced degree status.

While the 2% cutoff for defining related occupations may seem arbitrary, the rule results in a construction of majors and occupations that aligns with common sense and other papers in the literature.[16] A list of related occupations for each of the five aggregate college major categories

[15] This is similar to the "Top 5" occupation distinction made by Altonji et al. (2016b), but is more flexible in defining relatedness by taking into account the distribution of occupations within a given major.

[16] As an example, Kinsler and Pavan (2015) use a self-reported measure of occupational relatedness and find that there is considerable overlap across majors among workers who report being in the same related occupation. The difference between my definition of relatedness and the self-reported definition in Kinsler and Pavan is that my approach restricts all individuals in an occupation-major category to be either related or unrelated. In contrast, the self-reported definition of relatedness allows for both unrelated and related jobs to be observed in every occupation-major category.

Abel and Deitz (2015) pursue a different approach by mapping college majors to occupations using the Department of Labor's O*NET data and crosswalks provided by the Department of Education's National Center for Education Statistics (NCES). They distinguish between occupations that are a "college degree match" and occupations that are a "college major match" and find that college graduates with better job matches earn higher wages

is included in Table B5 for bachelor's degree holders and Table B6 for advanced degree holders. Importantly, the definition of relatedness explained here does not preclude the same occupation from being related to two different majors. This distinction allows for the occupation relatedness definition to match what is observed in the data.

To further illustrate my definition of occupation relatedness, I discuss four different extremes observed from Tables B5 and B6. First, almost all engineering occupations are not considered to be related to any major except STEM.[17] Second, salespersons and miscellaneous administrators are considered to be related to every major. Third, lower-level service jobs in food services, tourism, and administrative support tend to only be related to other majors, reflecting the occupations that aspiring performing artists and authors tend to work in. Finally, accountants and auditors are related to business majors, other majors, and STEM majors. Of additional note is that Table B6 includes a set of occupations not included in Table B5 such as actuaries, pharmacists, and lawyers. These occupations all have the expected relatedness with bachelor's degree major: actuaries and pharmacists are related only to STEM, while lawyers are related to all majors except education. Based on this set of illustrative examples, the definition of occupation relatedness posed here is reasonable.

Locations Because the empirical method employed in this paper does not work well in small samples, I aggregate locations as another way of maintaining statistical power. Specifically, I divide the United States into 15 locations, corresponding to states or groups of adjacent states. The 15 locations consist of the five largest states (California, Texas, Florida, New York, and Illinois), followed by the nine Census divisions, with the South Atlantic division being divided in two. The resulting locations range in population from 11.5 million to 39 million. A detailed list of each location is included in Table B7.

4.2 Descriptive Analysis

To motivate the modeling approach described in Section 2, I now discuss descriptive evidence of the heterogeneity of migration and occupational choice across majors at the national level, and heterogeneity in migration flows across certain locations by college major, advanced degree status, and occupation.

4.2.1 Summary statistics

This subsection details the main differences across majors in earnings, propensity to leave one's state of birth, and propensity to work in a related occupation.

Table 1 lists differences across major in the three outcomes considered in this paper. The results in the odd-numbered rows of the table are regression coefficients on major dummies,

and that better matches are more likely to occur in larger labor markets.

[17] Civil engineers and industrial engineers are also related to the "other" category of majors.

estimated at the national level and controlling for demographics, advanced degree status, CBSA fixed effects, and a cubic in potential experience. The results in parentheses are standard deviations of the distribution of state-level coefficients.

The results of Table 1 show that education majors earn the least, leave their birth state at the lowest rates, and work in related occupations at the highest rates. What is interesting from the table is that there is no clear monotonicity among these three outcomes. For example, STEM and business majors each earn about the same amount and work in related occupations at similar rates. However, STEM majors are much more likely to leave their state of birth.

Finally, the standard deviations in Table 1 show that there is substantial heterogeneity in these outcomes across states, and that state-specific variation in migration and availability of related occupations is as large as state-specific variation in earnings. While the spatial variation in earnings is well known, variation in migration and concentration of related occupations is much less known. As discussed previously, variation in these latter two outcomes is a crucial component of identification of the extended Roy model.

4.2.2 Transition Matrix

The results of the previous subsection indicate that there is sizable variation across locations in all three outcomes that I consider. In this section, I present evidence of how migration flows are related to the variation in location-specific outcomes previously documented.

Figure 3 displays the migration transition matrix by major for the five largest states, for those who do not hold an advanced degree. Rows indicate birth location, while columns indicate residence location. Each row and column contains five bars, which correspond to the five majors. Each bar is divided in two, with the bottom section corresponding to the share of individuals choosing the related occupation.

Examining Figure 3 reveals a number of facts that support the model. First, the flow of workers from New York to Florida is remarkable. Underscoring this pattern is the fact that Florida is disproportionately popular for New Yorker education majors. Furthermore, it is especially evident of non-pecuniary factors because the education majors who stay in New York disproportionately leave the teaching occupation, while the those who move to Florida are disproportionately in the education occupation. The reverse is also true: education majors who leave Florida (see the second row) are almost all those who choose the non-education occupation. This fact is evident of nonpecuniary preferences, because, as will be shown, I find that education majors who work as teachers in Florida face a wage cut for doing so. This nonpecuniary dimension of the choice is likely to affect the observed earnings distribution in a significant way.

Figure 4 is the transition matrix for advanced degree holders. While there are high flows from New York to Florida among this group, there are equally high flows from New York to California. Furthermore, the education majors in New York who earn master's degrees stay in

New York and work as teachers at much higher rates than their counterparts who do not hold master's degrees. These findings are further evidence of self-selection in location and occupation decisions that differ by college major and advanced degree status.

It is worth noting one other observation from Figures 3 and 4. Examining the middle bar of the off-diagonal elements of columns 1 and 4 shows the fraction of other majors who choose to move to California and New York. Of the movers who choose these two locations, other majors are disproportionately represented. This likely reflects the fact that other majors are composed of performing arts majors, and California and New York are hubs for such occupations. This is consistent with migration being a function not only of earnings, but also of availability of related occupations. I formally show this effect in more detail later.

Taken together, the results from Figures 3 and 4 provide additional evidence of the presence of nonpecuniary factors on the decision of where to live and in which occupation to work. These nonpecuniary factors are likely to cause the observed earnings distribution to look much different than if individuals were placed randomly into locations and occupations.

5 Estimation

In this section, I discuss how to estimate the final equation (3.4) of the model discussed in Sections 2 and 3. The estimation proceeds in two stages. First, I estimate the choice probabilities $\left(p_{ij\ell k}, p_{ijmn}\right)$. Second, I estimate the parameters of equation (3.4), including the unknown correction functions $\lambda_{j\ell k}$.

5.1 Choice probabilities

There are a variety of ways in which one can estimate the choice probabilities. Some alternatives include the conditional logit model, the conditional probit model, or non-parametric estimation techniques.

The conditional logit model is by far the most popular method used to estimate choice probabilities (and in migration models in particular, because the dimension of the choice set tends to be large) due to its simple closed-form expression for the underlying choice probabilities. The primary drawback of this model is that it suffers from the independence of irrelevant alternatives property.[18]

The conditional probit model (Hausman and Wise, 1978) allows for arbitrary correlations among the choice alternatives, but is unsuitable for settings such as this where the choice set is large. This is because the conditional probit model requires estimation of a $(J-1)$-dimensional integral, where J is the number of alternatives. Using this model would would eliminate the

[18]For tractability reasons, dynamic migration models such as Kennan and Walker (2011) and Ransom (2016) assume that migration probabilities take a conditional logit form. Davies et al. (2001) assume this form in a static setting. Monras (2015) argues that a nested logit is more appropriate for characterizing migration decisions.

gains afforded by the index sufficiency assumption discussed in Section 3.

Non-parametric estimation has two advantages. First, it does not require the researcher to model location-specific characteristics, of which there are a large number and many of which are poorly measured. Second, it does not require the researcher to specify the dependence structure of the choice alternatives as would be required with the conditional probit model or a nested logit (or GEV) model.[19]

The primary drawback to non-parametric estimation is deciding how finely and in which ways to divide the state space. Probabilities that are estimated from cells that are too small will introduce a large amount of error into the estimation. On the other hand, failure to create enough cells will result in probabilities that do not accurately represent the data.

5.1.1 Non-parametric estimation using machine learning

I estimate the location and occupational choice probabilities non-parametrically using a method from the machine learning literature called conditional inference recursive partitioning, developed by Hothorn et al. (2006) and implemented in the R programming language by Hothorn and Zeileis (2015).

The algorithm is designed to overcome the drawbacks associated with non-parametric estimation. The main advantage is that it prevents the researcher from being required to make ad hoc assumptions about how the state space should be divided when creating probability bins. It also has the advantage of automatically merging together sparse bins such that the algorithm does not return any empty bins or any bins of excessively small size. I detail the conditional inference tree algorithm in the following subsection.

Generally speaking, machine learning is the practice of allowing computers to learn for themselves without having to be explicitly programmed. In statistical applications, machine learning amounts to using methods that combine estimation with model selection to enhance out-of-sample prediction of statistical models. The result is an algorithm which automatically selects which covariates to include while also estimating their parameters. In the current setting, the conditional inference recursive partitioning algorithm selects which variables and which levels of the variables matter most in predicting migration and occupations. For other settings where the set of covariates is larger than the sample size, model selection methods automatically choose which covariates should be included such that standard rank and order conditions for identification are satisfied.[20] Varian (2014) provides an overview of basic machine learning algorithms and suggests ways in which they can be used to improve existing research methods

[19] Hausman and Wise (1978) note that the conditional probit model produces inconsistent estimates of the choice probabilities if dependence among the alternatives is incorrectly specified. Likewise, the conditional logit model produces inconsistent estimates if there is in fact any dependence among the alternatives. Estimates from the nested logit or other generalized extreme value (GEV) models are also inconsistent if the wrong nesting structure is specified.

[20] This setting applies to Bajari et al. (2015) who show how a variety of machine learning methods can be used in demand estimation to evaluate advertising effectiveness.

in economics. Asher et al. (2016) prove consistency of classification trees for heterogeneous moment-based models. Other examples of machine learning applications in economics include Athey and Imbens (2015), Gentzkow et al. (2015), and Belloni et al. (2011).[21]

5.1.2 Conditional inference recursive partitioning algorithm

The conditional inference recursive partitioning algorithm is a classification tree algorithm designed to non-parametrically predict a dependent variable from a set of covariates. The algorithm takes as inputs the dependent variable and the covariates, and returns as outputs combinations of the covariates that form clusters (nodes of the tree) or cells. Using an internal stopping criterion based on hypothesis testing, it optimally trades off bias (creating too few clusters and, as a result, poorly fitting the estimation data) and variance (creating too many clusters and, as a result, poorly fitting out of sample) such that out-of-sample prediction is maximized.[22] The algorithm works for both continuous and categorical variables on both sides of the equation.[23] The current application contains a categorical dependent variable and covariates that are primarily categorical, but some of which are continuous.

Below, I detail the algorithm, which recursively iterates on the following two steps:

1. *Selection.* The algorithm begins by testing whether the dependent variable is independent of the covariates (i.e. testing whether the distribution of the dependent variable Y is different from the conditional distribution $Y|X_j$ for all covariates). If any member of this set of conditional distributions is significantly different from the unconditional distribution, then the algorithm selects the covariate with the strongest association with Y as measured by a p-value.

2. *Splitting.* Once a covariate has been selected, the algorithm optimally splits it. This is done in a similar fashion as the selection, only the algorithm at this phase selects among different *subsets* of the specified covariate. The optimal split is the one that creates the most distinct pair of distributions of the dependent variable, as measured by a p-value. There are other criteria involved in determining if a candidate split is carried out; namely how large the resultant cluster will be. Clusters that are too small will predict poorly out-of-sample and are skipped accordingly.

[21] Athey and Imbens (2015) show how machine learning methods can be used to estimate heterogeneous treatment effects. Gentzkow et al. (2015) illustrate how to use model selection to estimate polarization in high-dimensional textual data. Belloni et al. (2011) develop methods for using model selection in instrumental variables models when the number of instruments is larger than the sample size.

[22] Hothorn et al. (2006) emphasize that the internal stopping criterion acts similarly to pruning or cross-validation methods that are commonly used in other machine learning settings to penalize complexity.

[23] In the case of a continuous dependent variable, the algorithm minimizes the sum of squared errors within each cluster to find the optimal cluster division. In the case of a continuous covariate, the algorithm creates bins by choosing cut points. The algorithm can also be used in survival analysis.

The algorithm then iterates on these two steps until at least one of the following criteria is met:[24]

- No additional covariates can be selected because they fail to reject the null hypothesis of independence.
- Any further splits of the already-selected covariates would fail to reject the null hypothesis of equality in the dependent variable across the split
- Any further splits would result in clusters with too few observations (i.e. unsuitable for out-of-sample prediction)
- The candidate cluster already perfectly predicts the dependent variable
- No further splits are possible because the candidate cluster is composed of a single combination of all independent variables

As an example of what the output of this algorithm looks like, I include Figure 5, which depicts a simple example of the output from a fictitious migration dataset. Individuals are characterized only by their level of work experience and can choose to live in 3 locations: New York, Texas, or elsewhere. The algorithm shows that experience is the strongest predictor of location choice, and that the most distinct difference occurs when splitting at three, followed by an additional split that occurs at eight. The algorithm shows that New York is entirely composed of individuals with less than four years of work experience, that Texas is composed nearly perfectly of individuals with experience levels between four and eight years, and that workers with nine or more years of experience almost certainly live elsewhere. In the actual estimation, each tree will have many more than three terminal nodes.

5.1.3 Implementation of the non-parametric estimation algorithm

I now discuss in detail the estimation of the choice probabilities and which variables are used to predict migration and occupational choice. Following Dahl (2002), I use cell decision probabilities, where the cells are computed from the recursive partitioning algorithm detailed above. The implicit assumption with this approach is that observably similar people face similar unobserved earnings and preference shocks. Importantly, this implies that the researcher need not model the characteristics of the alternatives, only the characteristics of the individuals.

[24] There are a few tuning parameters of the algorithm that the researcher can adjust. One is the p-value that determines splitting, another is the smallest number of observations allowed in a cluster, and a third is the smallest number of observations allowed in a candidate node split (i.e. the minimum number of observations required in each resulting subset of the split). I choose 5% for the p-value parameter, 50 observations for the minimum cluster size, and 50 observations for the minimum candidate node split size. These were chosen via cross-validation, but in practice the predictive accuracy of the tree algorithm was not sensitive to these tuning parameters.

Formally, the cell decision probability for all individuals, all origin locations j, and all destination locations ℓ and occupations k is

$$\begin{aligned} p_{ij\ell k} &= \Pr\left(d_{ij\ell k} = 1 \,|\, v_{j1k} - v_{j\ell k}, \ldots, v_{jLK} - v_{j\ell k}\right) \\ &= \Pr\left(d_{ij\ell k} = 1 \,|\, \text{cell}\right) \end{aligned} \tag{5.1}$$

The conditional inference tree algorithm assigns cells based on the following characteristics: whether the individual was born in the location of residence or in the same Census region; college major; advanced degree status; age; race; marital status; whether or not the individual is living with a family member or relative; whether or not the individual's spouse is working (if married); the presence of children ages 0-4 and ages 5-18; and the two exclusion restrictions discussed in Section 3.3. I estimate the cell probabilities using the so-called "one-vs-all" classification method: for each residence location and occupation, I compute the probability of choosing the alternative under consideration vs. all others. I then assign individuals into cells based on the terminal node of the tree in their chosen alternative.

5.1.4 Tree algorithm performance relative to more commonly used methods

A valid question regarding the conditional inference tree algorithm is how it compares with the traditional non-parametric bin estimator or with the logit estimator, the latter of which is by far the most popular estimation method for discrete choice models.

The primary benefits of the tree algorithm are twofold: (i) it allows the researcher to consider a large number of candidate covariates without having to worry about encountering the curse of dimensionality (i.e. the result of which would be empty bins); and (ii) it allows the sample space to be divided into irregularly shaped bins. The first benefit arises out of model selection and could be accomplished with other parameter regularization methods such as LASSO (Belloni et al., 2011). The second benefit arises out of the algorithm's recursive nature: by not making all splits simultaneously, the division of the state space can contain non-rectangular shapes. A final benefit of the algorithm is that it performs slightly better at out-of-sample prediction than existing methods.[25] A summary of this is given in Table B8.[26]

[25] Another general benefit of the tree algorithm is that it can inform structural modeling by providing the researcher with an ordered list of predictors. Traditionally, researchers have used theory to choose a set of candidate covariates. Decisions should continue to be based on theory; however, machine learning approaches can be combined with theory to give researchers an improved way of informing structural models.

[26] To assess the performance of each of the estimators, I estimate the first-best choice probabilities for each algorithm using the 2010-2015 ACS sample discussed previously. I then test the out-of-sample predictive performance of each algorithm using a holdout sample of the 2010-2015 ACS. The results from this exercise are detailed in Table B8. Each of the classification algorithms performs roughly similarly in terms of raw predictive accuracy as well as penalized predictive accuracy, with the tree algorithm slightly outperforming both of the alternatives. The definitions of each of these accuracy metrics are detailed in Table B8. The superior performance of the tree classifier is due to the inability of the bin estimator handle heterogeneous splits of the continuous covariates, foremost of which are the two exclusion restrictions. In the bin scenario, the researcher must choose cut points of this continu-

The benefits of the tree algorithm are manifest in Appendix A where I compare the small- and large-sample performance of various algorithms and error structures. The tree algorithm performs about as well as the bin estimator in large samples, but much better in small samples. Furthermore, if the researcher misspecifies the bins (because of the curse of dimensionality), then the tree algorithm significantly outperforms the simple bin estimator.

While the tree algorithm performs better in Monte Caarlo simulations, does it substantially alter the estimates of selection bias in the ACS data? The answer is yes. In results not shown, but available from the author upon request, I find that using either a bin or a logit estimator causes the degree of selection bias to be understated. That is, the model estimates when using these two estimators tend to fall in between those of OLS and the tree algorithm. This evidence is further support for the appropriateness of the tree algorithm in this particular application.

5.2 Correction functions

I now describe how to feasibly estimate the unknown selection correction functions in (3.4). As written, this equation contains LK bivariate correction functions for each location ℓ and occupation k. To further simplify this, I make the assumption that the selection correction functions are the same for everyone. In formal terms, this assumption imposes that the correction term in (3.4) be rewritten as $\lambda_{j\ell k}(\cdot) = \lambda_{\ell k}\left(p_{ij\ell k}, p_{ijmn}\right)$. While this assumption is restrictive, it allows me to estimate the wage effect of staying in the birth location. In results not shown, I also test the sensitivity of the measured returns to major when allowing separate correction functions for stayers and movers. The estimates change very little.

I now discuss my choice for the probabilities $p_{ij\ell k}, p_{ijmn}$. I assign as $p_{ij\ell k}$ the first-best choice probability, which is readily observable in the data. For p_{ijmn}, I use the probability that individual i would stay in the first-best location, but work in the non-chosen occupation. This is simply $p_{ij\ell k'}$, where k' denotes the non-chosen occupation.

To estimate the unknown correction functions $\lambda_{j\ell k}$, I use a flexible polynomial function of the probabilities as discussed in Dahl (2002). Extensive specification testing leads me to choose a polynomial of degree three in each of the probabilities. Also included are second- and third-order interactions between the two choice probabilities. Including a higher degree polynomial or a larger number of probabilities results in much less precise estimates with no appreciable increase in the Wald test statistic of joint significance of the polynomial. Lower degree polynomials do not appear to be flexible enough to adequately capture selection patterns. The final estimating equation is of the same form as (3.4), except that there are no summation operators because of the assumption that $\lambda_{j\ell k}(\cdot) = \lambda_{\ell k}\left(p_{ij\ell k}, p_{ijmn}\right)$, as discussed above.

ous variable in which to categorize the data. This process of discretization throws out useful variation. In contrast, the tree algorithm allows different splits of the exclusion restriction to be made at different combinations of the covariates.

5.3 Earnings equation

The earnings equation parameters in (2.1) are estimated by OLS (separate equations for each location and occupation) after making use of the index sufficiency assumption in (3.3) and the dimensionality reduction assumptions discussed in the previous section.

5.3.1 Standard errors

The standard errors of the parameters associated with the selection functions must be adjusted to account for two elements of the estimation: (i) the selection probabilities are not i.i.d. across individuals because of the cell assumption in (5.1); and (ii) the estimation of the cell probabilities induces estimation error into the coefficients because the true probabilities are not observed. To resolve this, I follow Dahl (2002), who proposes the following feasible estimator of the asymptotically correct covariance matrix:

$$\hat{V} = \hat{\sigma}^2 (X'X)^{-1} + (X'X)^{-1} \hat{\Gamma} \hat{V}(P) \hat{\Gamma}' (X'X)^{-1} \tag{5.2}$$

where X is the matrix of earnings equation covariates (including the correction function terms), $\hat{\Gamma}$ is a block-diagonal matrix containing the derivatives of the polynomial correction functions with respect to the probabilities, evaluated at the estimated polynomial coefficient and estimated probabilities. $\hat{V}(P)$ is a block diagonal matrix with each block containing the 2×2 covariance matrix for the estimated first-best and occupation probabilities within the given cell.

At present, I present standard errors that are corrected for the fact that probabilities are based on migration cells, but which do not explicitly correct for estimation error in the choice probabilities. This is done by clustering the standard errors at the cell level.

6 Empirical Results

In this section, I discuss the results of the estimation procedure described in the previous section and their implications. I first present results on the estimation of the decision probabilities, followed by a discussion of the selection-corrected estimates of the returns to majors and occupational relatedness. Finally, I analyze migration flows across space

6.1 Choice probabilities

The cell assumption in (5.1) states that the choice probability of a given cell is the probability that all individuals in the cell make the same choice.[27] Thus, deviations from the cell mean correspond to a reduced-form measure of preference shocks, which allow me to separate prefer-

[27] For parametric choice models, the analogous assumption is that the choice probability is the same for all individuals with the same values for all covariates.

ences from earnings. Because of their key role in identification, I present in Table 2 moments of the distributions of average cell probabilities, conditional on major, occupation, and move-stay decision. The table also reports the number of individuals in each migration-occupation-major classification and the number of different cells contributing to each classification.

An implication of the earlier discussion on identification is that identification requires the decision probabilities across majors within a migration-occupation bin to be overlapping. Intuitively, the returns to major can be calculated by comparing individuals in two different majors who have the same preferences, as measured by the cell probabilities. Examination the table reveals that there is a wide range of overlap in probabilities across majors.

The probabilities listed in Table 2 also confirm the earlier descriptive analysis of Figures 3 and 4. The cell probabilities in panels (a) and (c), which correspond to working in a related occupation, are highest among education, business, and STEM majors. Another way to see this is to compare the difference in average cell probabilities for working in a related occupation relative to working in an unrelated occupation. This difference is much higher for education, business, and STEM majors than for the remaining two.

Finally, note that the number of cells is larger for movers than for stayers, and that the number of cells roughly corresponds to the number of individuals within a major-occupation category. The difference in the number of cells is much less stark than if a bin estimator were to be used, because the tree algorithm automatically merges together sparse bins, or bins that are not statistically distinct, to avoid overfitting.

6.2 Earnings

I now discuss the estimates of the earnings equation with and without selection correction. The primary parameters of interest are the college major dummies and their interaction with a dummy for advanced degree attainment. The primary research question is how these parameter estimates change once I account for self-selection into locations and occupations. Throughout, I treat bachelor's-level education majors as the reference category.

6.2.1 Estimates for specific states

Table 3 lists the full estimates of equation (3.4) with the implemented simplifications discussed in Section 5.2. While I estimate 30 equations, I present detailed results for only three of the five most populous states. I later present aggregate results for all 15 locations.

Table 3 reports the earnings equation estimates for each occupation in the three states, for both the naive case and the corrected case. The first column within each state and occupation reports the estimated returns to each major assuming no selection bias, while the second column reports the estimated returns after correcting for selection. The OLS estimate is upward biased for the vast majority of all measured returns. The magnitude of the upward bias differs from state to state, with the largest differences in New York and the smallest differences in Florida.

As noted previously, I am able to separately identify the earnings effect of stayers. These estimates are reported on the last row of Table 3. There is actually a wage penalty for stayers in each of these three states. This penalty gets erased once controlling for selection, indicating that what naively appears to be a compensating differential for staying in one's birth state is actually a selection effect.

It is important to keep in mind the interpretation of what generates the direction of the bias in returns. As noted in Dahl (2002) and Bayer et al. (2011), an upward bias in the returns to schooling is the result of individuals responding to above-average earnings shocks. This comes about in the model through the selection correction terms: if someone moves to a location when observationally similar individuals do not, then it must be because of a favorable earnings shock. Put differently, moves in response to favorable earnings shocks will overstate the treatment effect of randomly assigning individuals to live in a given location.

A remaining question upon examining the results in Table 3 is whether or not the differences are statistically significant. I test for this in two ways: (i) I conduct a Wald test for joint significance of the polynomial of choice probabilities; and (ii) I conduct a Hausman-type test where the null hypothesis is that the baseline OLS is efficient and consistent, while the corrected estimates are consistent but inefficient. The former is a necessary condition of the presence of bias, while the latter is a sufficient condition. I present the Wald test statistics in the bottom of Table 3. In all cases, the Wald tests have p-values smaller than 0.003. I present the results of the Hausman test for all locations for select major-occupation combinations in Tables B9 through B20. In Florida, for example, the following returns have Hausman p-values less than 10%: advanced degree STEM, business, and social science majors in related occupations; and BA social science and STEM majors in unrelated occupations.

On aggregate, about one-third of all returns are significantly different. The returns to major among advanced degree holders who work in related occupations tend to be the most significantly different from OLS.

6.2.2 Estimates for all locations

I now present results on the returns to major for all locations for each of the two occupations and advanced degree statuses. Figures 6 through 9 contain plots for all majors and degree attainments. The figures include a 45 degree line with the corrected return on the vertical axis and the uncorrected return on the horizontal axis. Circles or dots represent pairs of returns, where circles indicate that the difference is not significant while dots indicate that the corrected estimate is statistically significantly different at the 10% level or lower, using the Hausman-type test described previously.

In all figures there is upward bias in the returns to major for almost all locations and majors. The magnitude of the bias is larger for business and STEM majors in related occupations, and especially so for advanced degree holders.

To assess the magnitude of bias, I present in Table 4 the percentage change in returns when correcting for selection. Some of the returns have very low bases on which the percent change is calculated, particularly for the low-earnings majors. Thus, I focus on STEM and business majors, and find that the magnitude of the bias ranges from 0% to 45% with a median value of between 5% and 17%. As shown in the earlier graphs, the magnitude of the bias is largest among advanced degree holders who work in related occupations.

Finally, a valid question is whether or not correcting for selection bias in the returns narrows the gap in returns across locations? Examining the graphs in Figures 6 through 9 shows that the range of values is roughly the same for both the horizontal and vertical axes. This means that the cross-location range in returns is largely unaffected by the selection correction. This finding implies that wage differences across locations are most likely due to other factors such as innate productivity differences or compensating differentials.

6.3 Migration flows

With estimates of individual-level migration probabilities and location-occupation-level selection-corrected returns to major, I now investigate the cross-major comparative responsiveness of migration flows to earnings, occupational availability, and local amenities.

To formalize ideas, consider the log migration flow from location j to ℓ for college graduates in major m with advanced degree status $a \in \{0, 1\}$. This is assumed to be

$$\begin{aligned} \ln \hat{p}^m_{aj\ell} =& \psi^m_{a0} + \psi_1 \sum_a \sum_k \left(w^m_{a\ell k} - w^m_{ajk} \right) + \psi^m_{a2} \ln \text{dist}_{j\ell} + \\ & \psi_{a3} \left(\ln R^m_{a\ell} - \ln R^m_{aj} \right) + \psi^m_4 \left(\ln A_\ell - \ln A_j \right) + \nu^m_{aj\ell} \end{aligned} \tag{6.1}$$

where $\hat{p}^m_{aj\ell} = \frac{1}{N_{a,m}} \sum_i \sum_k \hat{p}_{ij\ell k} 1[s_i = (a, m)]$ is the average of the individual estimated migration probabilities for all individuals of major m with advanced degree status a. The estimates are taken from the procedure outlined in Section 5.

Equation (6.1) states that migration flows for individuals of a particular major are a function of cross-location differences in the following characteristics: four log earnings terms ($w_{a\ell k}, a \in \{0, 1\}, k \in \{\text{unrelated, related}\}$); log distance (Great Circle formula, in miles); the log fraction of individuals in major m who work in a related occupation, R; and log measures of local amenities A.[28] Because each right hand side variable is expressed in logs, the corresponding coefficients represent elasticities. Importantly, I restrict the log earnings elasticity to be the same for both

[28]Specifically, A includes a number of climate, geographical, and local government amenities to capture differences in quality of life across locations. These amenities include: climate measures such as cloudiness, average wind speed, heating degree days, cooling degree days, morning humidity, and precipitation; quality of life measures such as per-pupil schooling expenditures, population density, health care expenditures per capita, and violent crime rates; and local spending measures such as state budget expenditures per capita and higher education expenditures per full-time-equivalent student. Variables are measured either at the city or state level and are aggregated to the regional level by weighting by the component populations.

occupations and both advanced degree statuses because there is a high level of correlation in these measures within locations. The theoretical implication of this assumption is that all individuals value money in the same way. A similar argument explains why I restrict ψ_4 to be the same for both bachelors and advanced degree holders.

In order to feasibly estimate the effect of earnings on migration flows, I difference (6.1) with respect to individuals of a different major m'. Similar analyses have been used by Dahl (2002) to investigate migration behavior and by Wiswall and Zafar (2015) to investigate college major choice. Rewriting (6.1) gives

$$\begin{aligned}\ln \hat{p}^{m}_{aj\ell} - \ln \hat{p}^{m'}_{aj\ell} = \left(\psi^{m}_{a0} - \psi^{m'}_{a0}\right) + \psi_1 \sum_a \sum_k \left[\left(w^{m}_{a\ell k} - w^{m'}_{a\ell k}\right) - \left(w^{m}_{ajk} - w^{m'}_{ajk}\right)\right] + \\ \left(\psi^{m}_{a2} - \psi^{m'}_{a2}\right) \ln \text{dist}_{j\ell} + \\ \psi_{a3}\left[\left(\ln R^{m}_{a\ell} - \ln R^{m'}_{a\ell}\right) - \left(\ln R^{m}_{aj} - \ln R^{m'}_{aj}\right)\right] + \\ \left(\psi^{m}_{4} - \psi^{m'}_{4}\right)\left(\ln A_\ell - \ln A_j\right) + \left(\nu^{m}_{aj\ell} - \nu^{m'}_{aj\ell}\right)\end{aligned} \tag{6.2}$$

Note that, according to equation (3.4), earnings differences only vary across majors through the parameter on schooling, $\gamma^{a}_{2\ell k}$. Thus, (6.2) can be rewritten in a more compact form as

$$\begin{aligned}\Delta^m \ln \hat{p}^{m}_{aj\ell} = \widetilde{\psi}_{a0} + \psi_1 \sum_a \sum_k \Delta^m \Delta^\ell \hat{\gamma}^{a,m}_{2\ell k} + \widetilde{\psi}_{a2} \ln \text{dist}_{j\ell} + \\ \psi_{a3} \Delta^m \Delta^\ell \ln R^{m}_{a\ell} + \psi_4 \Delta^\ell \ln A_\ell + \widetilde{\nu}_{aj\ell}\end{aligned} \tag{6.3}$$

where $\Delta^m \Delta^\ell$ signifies the difference-in-differences operator across majors and locations.

Estimates of (6.3) are reported in Table 5. Each column is based on $2\left(L^2 - L\right)$ observations, or the number of off-diagonal pairwise location combinations for both advanced degree groups. In all columns, the normalized major m' is education. Thus, the estimates should be interpreted as determinants of migration for individuals with major m and advanced degree a relative to education majors of advanced degree group a. For each major, I report results with and without the local amenity variables. My preferred results include the amenity variables, as these are shown to contain a large amount of predictive power as measured by the R squared statistic.

The results of Table 5 indicate that, in addition to differences in the wage returns to major, migration flows for all majors relative to education are responsive to distance and availability of related occupations. Most surprising is that, for all majors, the elasticity for related occupation concentration is larger than the elasticity of earnings, by a factor of 2 or more. These large elasticities are strongest among STEM and business majors with advanced degrees and among social science and other majors without advanced degrees.

The response of migration to wage returns is strongest for social science majors in certain specifications. In others, the earnings elasticity is roughly equal among business, other, and social science majors. All majors show positive elasticities with respect to distance relative to

education majors, but only for advanced degree holders. This implies that education majors with master's degrees face the largest costs to migration, which is likely a result of state-level education policies.[29] This result, coupled with the low distance elasticity for business majors, is consistent with descriptive evidence presented in Figures 3 and 4 which show that education and business majors are the least likely to leave their state of birth, and that master's level education majors are even less likely to leave.

Comparing the odd- and even-numbered columns of Table 5 reveals the effect of including amenity measures on the estimated impacts. Including these measures has the effect of lowering both the wage and related occupation elasticities for all majors except STEM. The robustness of the earnings and occupation elasticities even after accounting for a wide variety of amenities confirms their importance to migration decisions.

The findings of this section have important implications for local governments seeking to retain their educated workers, either through higher education subsidies or other place-based policies. In particular the results underscore the importance of employment in related occupations as a substantial component of the migration decision of college graduates. If state governments wish to retain the college graduates whose tuition they have partially subsidized, then it may be optimal to index the tuition of different majors to the local concentration of related occupations for those majors.[30]

The findings of this section also have important implications for the literature on college major and occupational choice. The fact that college graduates have preferences for working in related occupations and that these occupations are not uniformly distributed across space implies that post-college outcomes are dependent on local labor market characteristics. Thus, location preferences may be a large component of the non-pecuniary factors that other studies have found when studying the determinants of college major choice.

7 Conclusion

This paper examines the extent to which selection into residence location and occupation biases the wage returns to college majors. To analyze this question, I develop and estimate an extended Roy model where individuals have preferences for both wage and non-wage aspects of given location-occupation pairs. Using estimates of the model, I examine how sensitive migration flows of different majors respond to cross-location differences in wage returns to major, availability of occupations related to the major, and non-wage local amenities.

To estimate the model, I implement the framework of Dahl (2002) and Lee (1983) which

[29] For example, Ashworth (2015) studies teachers' decisions to obtain master's degrees in North Carolina where the associated wage premium is legislated by the state. Many states have similar fixed salary schedules for primary and secondary school teachers. Thus, moving states may cause a reduction in wages either because the worth of a master's degree might be lower in the new state, or because work experience may not be counted in the same way in the new state.

[30] For a review of major-specific tuition pricing policies, see Stange (2015).

allows for feasible estimation of the extended Roy model by expressing the selection in terms of a small number of observed choice probabilities. I estimate the model using data on college-educated men from the American Community Survey from years 2010-2015. I also illustrate the advantages of using machine learning methods to non-parametrically estimate the selection probabilities. The primary advantage of this is in combining model selection and estimation.

I find that selective migration and occupational choice cause an upward bias in the measured wage returns to college major, relative to education majors. The percent change in the corrected returns ranges from 0% to 45% for STEM and business majors, is strongest among advanced degree holders, and is statistically significant in about one-third of all locations. Correcting for selection bias does little to narrow the range in returns. This implies that cross-location differences in the wage returns are due to other reasons, such as innate productivity differences or compensating differentials.

My analysis of migration flows shows that migration decisions of college graduates are determined by the concentration of related occupations in addition to wage returns and non-wage amenities. The elasticity for related occupation concentration is twice that of earnings, and is strongest among advanced degree holders who are business or STEM majors and among bachelors degree holders who are social science or other majors.

Given that migration flows are sensitive to occupational density, these results imply that place-based policies designed to retain or attract skilled workers may not be successful without taking into account workers' preferences for occupations. The results also point to the importance of considering migration and local occupation concentration when determining major-specific tuition rates at universities (Stange, 2015).

These results also raise questions about what is in a student's information set at the time of college major choice. Research on the determinants of college major choice indicates that non-monetary preferences for class subject or post-college occupation are a significant part of the major decision (Arcidiacono et al., 2014; Wiswall and Zafar, 2015). Such non-monetary occupational preferences might also reflect preferences for a location because occupations are not distributed evenly across space.

References

Abel, Jaison R. and Richard Deitz. 2015. Agglomeration and job matching among college graduates. *Regional Science and Urban Economics* 51:14–24.

Altonji, Joseph G., Peter Arcidiacono, and Arnaud Maurel. 2016a. The analysis of field choice in college and graduate school: Determinants and wage effects. In *Handbook of the economics of education*, vol. 5, eds. Eric Hanushek, Stephen Machin, and Ludger Wößmann. North Holland: Elsevier Science, 305–396.

Altonji, Joseph G., Erica Blom, and Costas Meghir. 2012. Heterogeneity in human capital investments: High school curriculum, college major, and careers. *Annual Review of Economics* 4, no. 1:185–223.

Altonji, Joseph G., Lisa B. Kahn, and Jamin D. Speer. 2016b. Cashier or consultant? Entry labor market conditions, field of study, and career success. *Journal of Labor Economics* 34, no. S1:S361–S401.

Arcidiacono, Peter, V. Joseph Hotz, Arnaud Maurel, and Teresa Romano. 2014. Recovering ex ante returns and preferences for occupations using subjective expectations data. Working Paper 20626, National Bureau of Economic Research.

Asher, Sam, Denis Nekipelov, Paul Novosad, and Stephen P. Ryan. 2016. Classification trees for heterogeneous moment-based models. Working Paper 22976, National Bureau of Economic Research.

Ashworth, Jared. 2015. Teacher education and career outcomes. Working paper, Pepperdine University.

Athey, Susan and Guido W. Imbens. 2015. Machine learning methods for estimating heterogeneous causal effects. Working paper, Stanford University.

Bajari, Patrick, Denis Nekipelov, Stephen P. Ryan, and Miaoyu Yang. 2015. Demand estimation with machine learning and model combination. Working Paper 20955, National Bureau of Economic Research.

Bayer, Patrick, Shakeeb Khan, and Christopher Timmins. 2011. Nonparametric identification and estimation in a Roy model with common nonpecuniary returns. *Journal of Business & Economic Statistics* 29, no. 2:201–215.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2011. Lasso methods for gaussian instrumental variables models. Working paper, Duke Fuqua School of Business, Massachusetts Institute of Technology, and Chicago Booth School of Business.

Bishop, Kelly. 2012. A dynamic model of location choice and hedonic valuation. Working paper, Washington University in St. Louis.

Borjas, George J. 1987. Self-selection and the earnings of immigrants. *American Economic Review* 77, no. 4:531–553.

Card, David and Alan B. Krueger. 1992. Does school quality matter? returns to education and the characteristics of public schools in the United States. *Journal of Political Economy* 100, no. 1:1–40.

Dahl, Gordon B. 2002. Mobility and the return to education: Testing a Roy model with multiple markets. *Econometrica* 70, no. 6:2367–2420.

Davies, Paul S., Michael J. Greenwood, and Haizheng Li. 2001. A conditional logit approach to U.S. state-to-state migration. *Journal of Regional Science* 41, no. 2:337–360.

D'Haultfœuille, Xavier and Arnaud Maurel. 2013. Inference on an extended Roy model, with an application to schooling decisions in France. *Journal of Econometrics* 174, no. 2:95–106.

Diamond, Rebecca. 2016. The determinants and welfare implications of US workers' diverging location choices by skill: 1980-2000. *American Economic Review* 106, no. 3:479–524.

Eisenhauer, Philipp, James J. Heckman, and Edward Vytlacil. 2015. The generalized roy model and the cost-benefit analysis of social programs. *Journal of Political Economy* 123, no. 2:413–443.

Falaris, Evangelos M. 1987. A nested logit migration model with selectivity. *International Economic Review* 28, no. 2:429–443.

French, Eric and Christopher Taber. 2011. Identification of models of the labor market. In *Handbook of labor economics*, vol. 4, Part A, eds. Orley Ashenfelter and David Card. North Holland: Elsevier, 537–617.

Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy. 2015. Measuring polarization in high-dimensional data: Method and application to congressional speech. Working paper, Stanford University, Brown University, and Chicago Booth School of Business.

Hausman, Jerry A. and David A. Wise. 1978. A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and heterogeneous preferences. *Econometrica* 46, no. 2:403–426.

Heckman, James, Anne Layne-Farrar, and Petra Todd. 1996. Human capital pricing equations with an application to estimating the effect of schooling quality on earnings. *Review of Economics and Statistics* 78, no. 4:562–610.

Heckman, James J. 1979. Sample selection bias as a specification error. *Econometrica* 47, no. 1:153–161.

Heckman, James J. and Bo E. Honoré. 1990. The empirical content of the Roy model. *Econometrica* 58, no. 5:1121–1149.

Heckman, James J. and Christopher Taber. 2008. Roy model. In *The new palgrave dictionary of economics*, eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2 ed., 1–9.

Heckman, James J. and Edward J. Vytlacil. 2007a. Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation. In *Handbook of econometrics*, vol. 6, Part B, eds. James J. Heckman and Edward E. Leamer. North Holland: Elsevier, 4779–4874.

———. 2007b. Econometric evaluation of social programs, part II: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments. In *Handbook of econometrics*, vol. 6, Part B, eds. James J. Heckman and Edward E. Leamer. North Holland: Elsevier, 4875–5143.

Hothorn, Torsten, Kurt Hornik, and Achim Zeileis. 2006. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics* 15, no. 3:651–674.

Hothorn, Torsten and Achim Zeileis. 2015. partykit: A modular toolkit for recursive partytioning in R. *Journal of Machine Learning Research* 16, no. 12:3905–3909.

Kennan, John. 2015. Spatial variation in higher education financing and the supply of college graduates. Working Paper 21065, National Bureau of Economic Research.

Kennan, John and James R. Walker. 2011. The effect of expected income on individual migration decisions. *Econometrica* 79, no. 1:211–251.

Kinsler, Josh and Ronni Pavan. 2015. The specificity of general human capital: Evidence from college major choice. *Journal of Labor Economics* 33, no. 4:933–972.

Kline, Patrick and Enrico Moretti. 2014. Local economic development, agglomeration economies, and the big push: 100 years of evidence from the tennessee valley authority. *Quarterly Journal of Economics* 129, no. 1:275–331.

Lee, Lung-Fei. 1983. Generalized econometric models with selectivity. *Econometrica* 51, no. 2:507–512.

Lemieux, Thomas. 2014. Occupations, fields of study and returns to education. *Canadian Journal of Economics* 47, no. 4:1047–1077.

McHenry, Peter. 2011. The effect of school inputs on labor market returns that account for selective migration. *Economics of Education Review* 30, no. 1:39–54.

Molloy, Raven S. and Abigail Wozniak. 2011. Labor reallocation over the business cycle: New evidence from internal migration. *Journal of Labor Economics* 29, no. 4:697–739.

Monras, Joan. 2015. Economic shocks and internal migration. Discussion Paper 8840, IZA.

Moretti, Enrico. 2012. *The new geography of jobs*. New York: Houghton Mifflin Harcourt.

Ransom, Michael R and Aaron Phipps. Forthcoming. The changing occupational distribution by college major. *Research in Labor Economics* 45, no. 1.

Ransom, Tyler. 2016. The effect of business cycle fluctuations on migration decisions. Working paper, Duke University.

Roy, A.D. 1951. Some thoughts on the distribution of earnings. *Oxford Economic Papers* 3, no. 2:135–146.

Ruggles, Steven, Katie Genadek, Ronald Goeken, Josiah Grover, and Matthew Sobek. 2015. *Integrated public use microdata series: Version 6.0 [machine-readable database]*. Minneapolis: University of Minnesota.

Stange, Kevin. 2015. Differential pricing in undergraduate education: Effects on degree production by field. *Journal of Policy Analysis and Management* 34, no. 1:107–135.

Varian, Hal R. 2014. Big data: New tricks for econometrics. *Journal of Economic Perspectives* 28, no. 2:3–28.

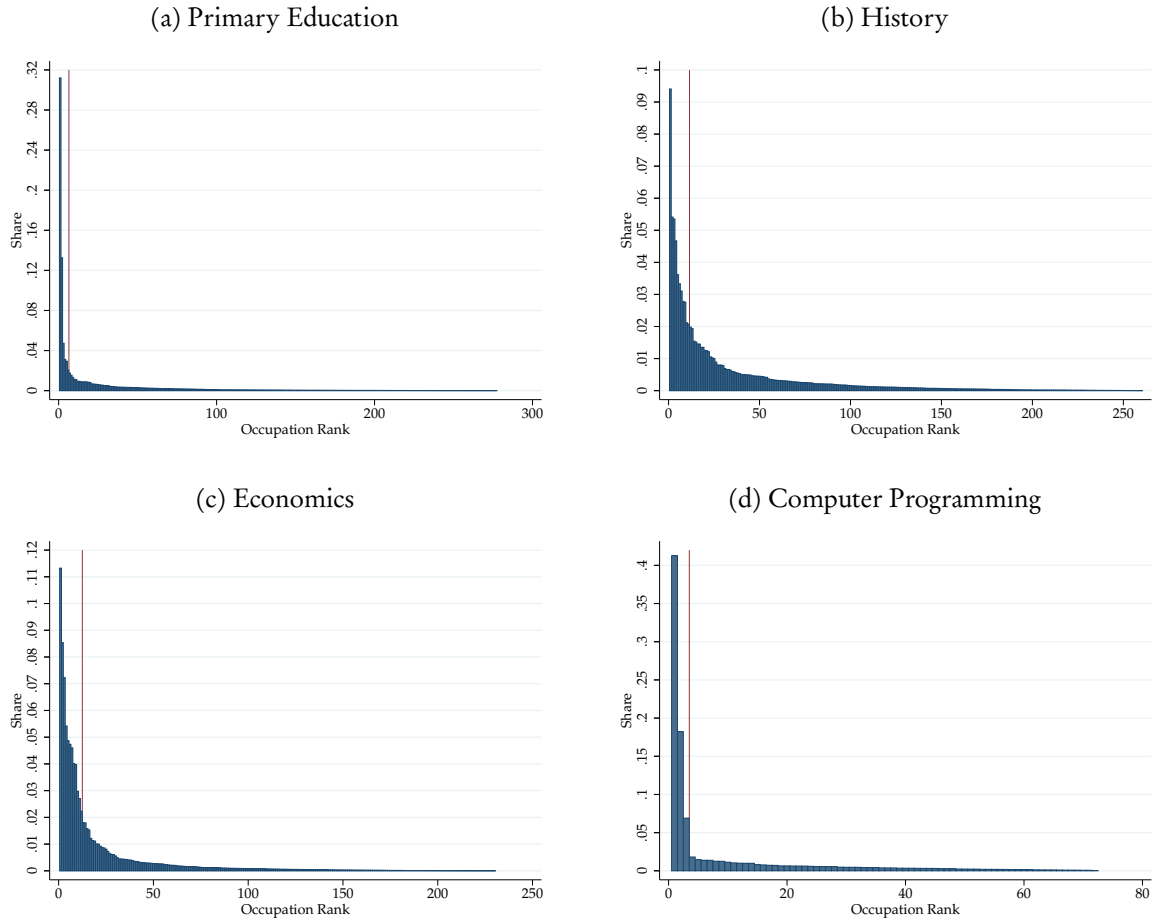
Winters, John V. Forthcoming. Do earnings by college major affect college graduate migration? *Annals of Regional Science* .

Wiswall, Matthew and Basit Zafar. 2015. Determinants of college major choice: Identification using an information experiment. *Review of Economic Studies* 82, no. 2:791–824.

Zabek, Mike. 2016. Population growth, decline, and shocks to local labor markets. Working paper, University of Michigan.

Figures and Tables

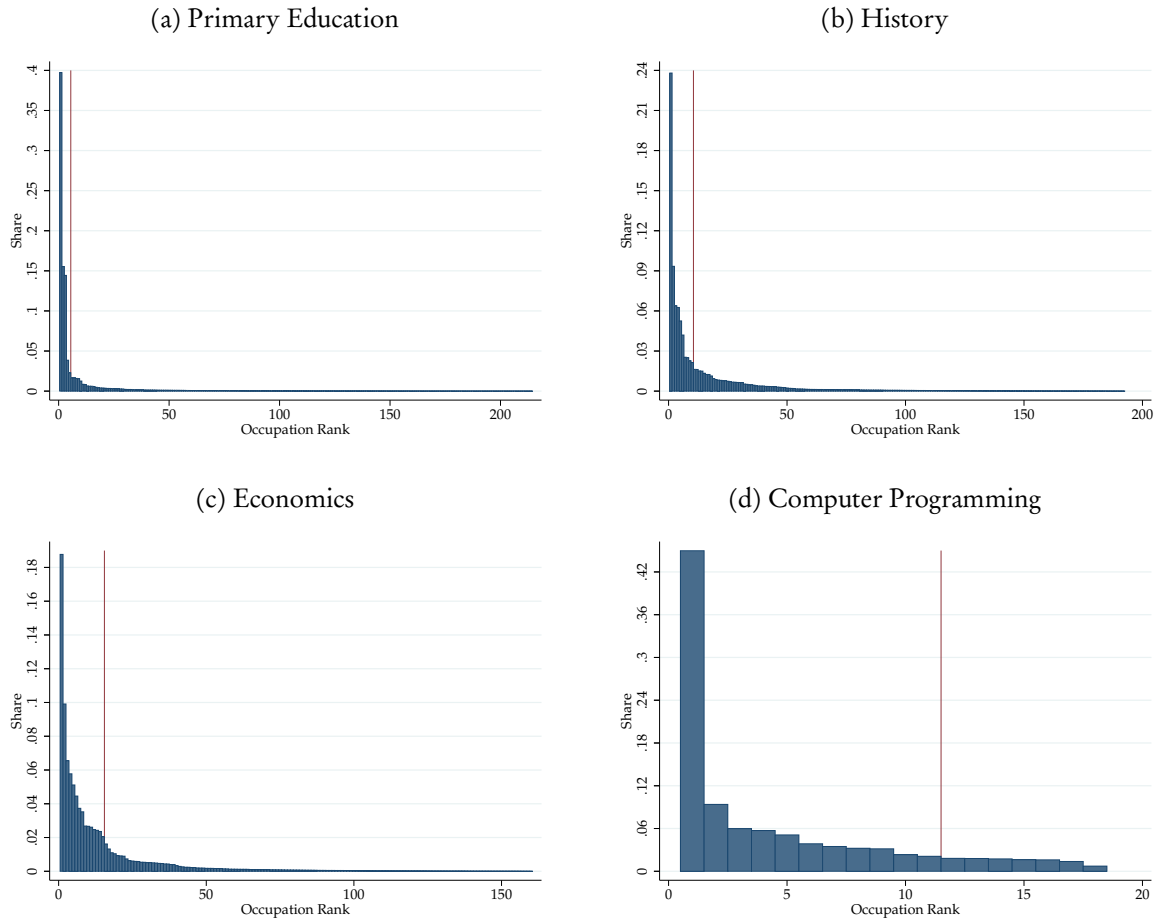
Figure 1: Occupation distributions for select detailed majors: Non-advanced degree holders



Notes: Graphs represent occupation distributions conditional on detailed major. Vertical lines represent the cutoff between related and unrelated occupations: those to the left of the line are related, while those to the right are unrelated. Additional details regarding the definition of occupation relatedness are provided in the text and the appendix.

Source: Author's calculations from American Community Survey, 2010-2015.

Figure 2: Occupation distributions for select detailed majors: Advanced degree holders



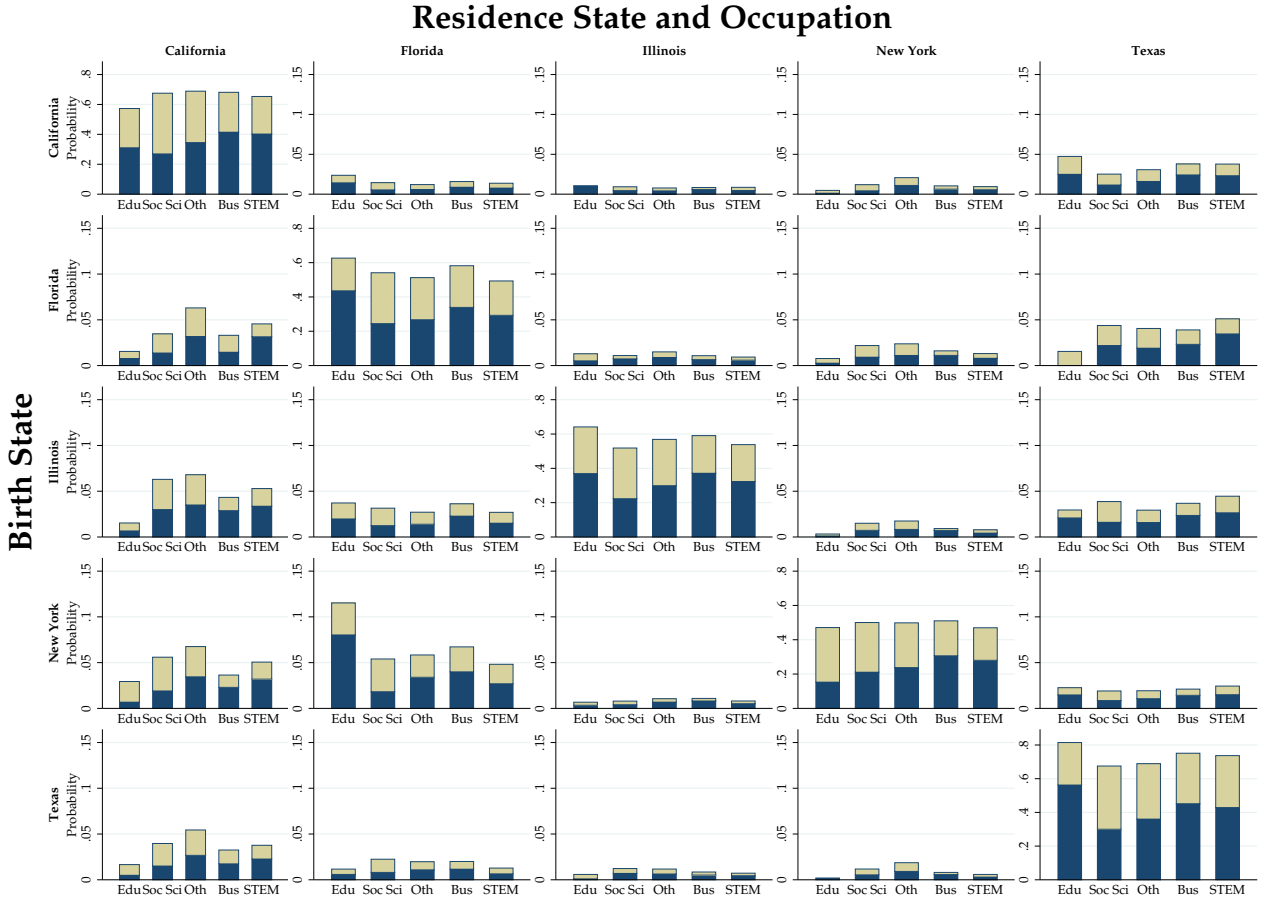
Notes: Graphs represent occupation distributions conditional on detailed major. Vertical lines represent the cutoff between related and unrelated occupations: those to the left of the line are related, while those to the right are unrelated. Additional details regarding the definition of occupation relatedness are provided in the text and the appendix.
Source: Author's calculations from American Community Survey, 2010-2015.

Table 1: Differences in outcomes by college major, relative to education majors

	Education	Soc Sci	Other	Business	STEM
Log Earnings	0.00	0.184	0.154	0.388	0.402
	(—)	(0.099)	(0.079)	(0.071)	(0.080)
Pr (Lives outside birth state)	0.00	0.115	0.124	0.077	0.134
	(—)	(0.060)	(0.072)	(0.064)	(0.062)
Pr (Works in related occupation)	0.00	-0.154	-0.110	-0.021	-0.029
	(—)	(0.078)	(0.085)	(0.089)	(0.091)
Frequency	5.32	11.35	20.90	28.53	33.91
N	31,043	66,276	122,015	166,569	198,011

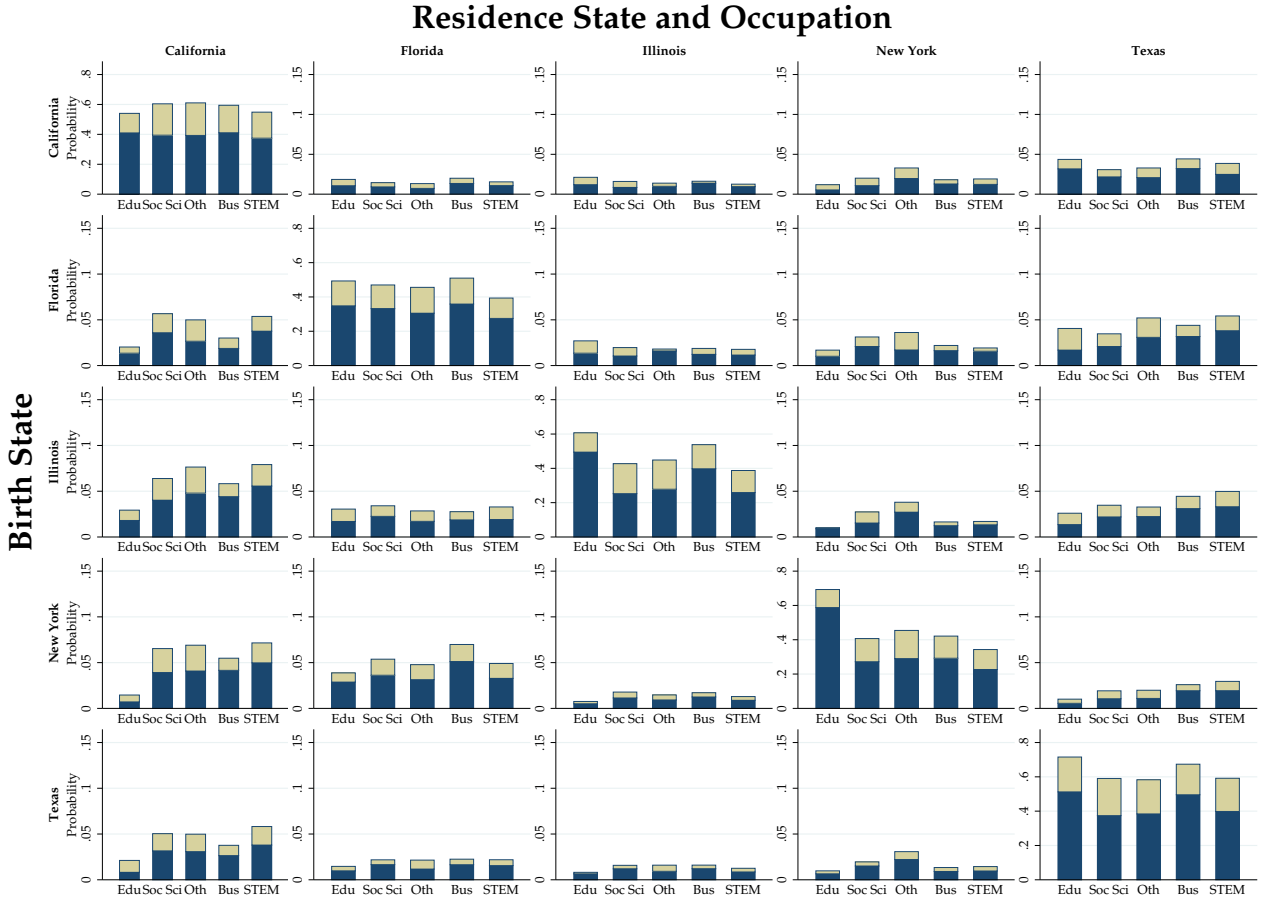
Notes: Regression estimates at national level, controlling for demographics, advanced degree status, CBSA dummies, and a cubic in potential experience. Standard deviation of state-specific estimates reported below in parentheses. All variables except for log earnings and distance are expressed in percentage points and estimated from linear probability models. Sample taken from the 2010-2015 American Community Survey and is restricted to males ages 22-54 with a bachelor's degree or higher. Sample weights are included in the computation. Additional details on sample selection can be found in Table B1.

Figure 3: Migration and occupation transition matrix by major for the five largest states: Non-adv. deg. holders



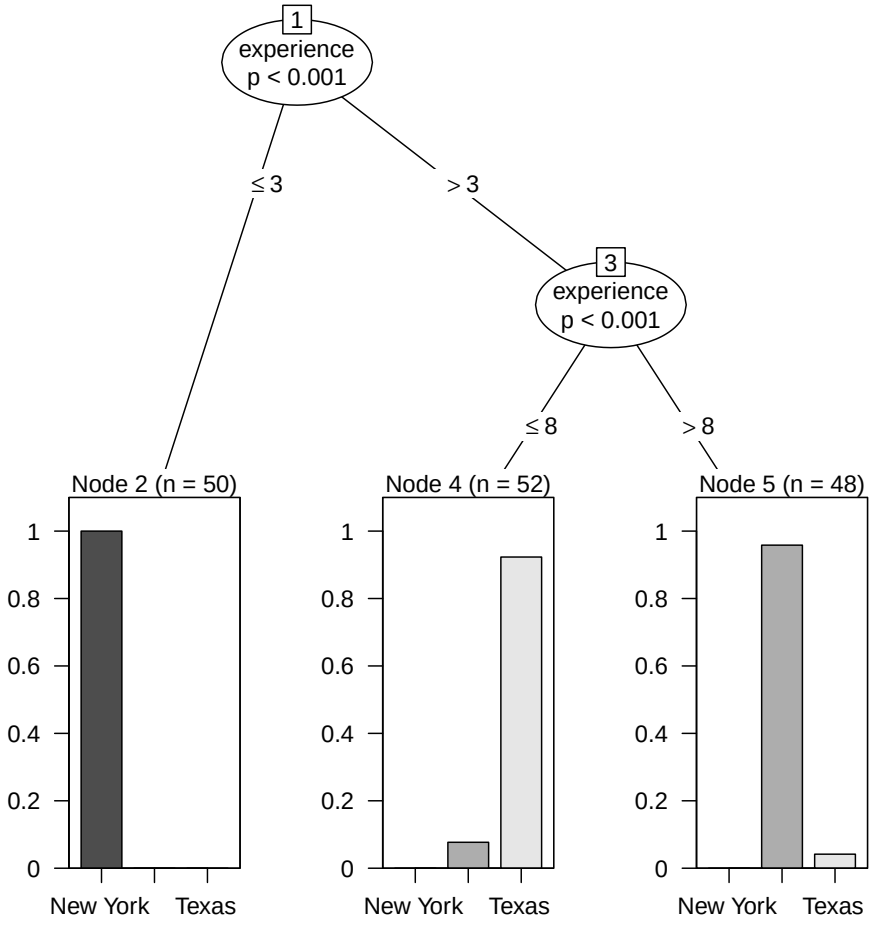
Notes: Markov transition matrix probabilities of living in a particular location and working in a particular occupation, by major, for the five largest US states. Light-colored bar segments represent proportion working in an unrelated occupation. Dark-colored bar segments represent proportion working in a related occupation.

Figure 4: Migration and occupation transition matrix by major for the five largest states: Adv. degree holders



Notes: Markov transition matrix probabilities of living in a particular location and working in a particular occupation, by major, for the five largest US states. Light-colored bar segments represent proportion working in an unrelated occupation. Dark-colored bar segments represent proportion working in a related occupation.

Figure 5: Simple example of tree structure from conditional inference recursive partitioning algorithm



Note: Sample tree output from fictitious data using the algorithm described in Section 5.1.2

Table 2: Summary of cell probabilities of observed decisions

(a) Stayers, Related occupation

Education Level	Cells	Individuals	Mean	Std. Dev.	10th Percentile	90th Percentile
Education Major	306	14,359	0.5285	0.1662	0.2841	0.7220
Social Sciences Major	328	16,121	0.3625	0.1316	0.1965	0.5214
Other Major	342	32,107	0.3616	0.1164	0.2058	0.4924
Business Major	363	55,839	0.4295	0.1158	0.2623	0.5554
STEM Major	371	61,008	0.4031	0.1158	0.2345	0.5284

(b) Stayers, Unrelated occupation

Education Level	Cells	Individuals	Mean	Std. Dev.	10th Percentile	90th Percentile
Education Major	374	6,061	0.3006	0.1505	0.1235	0.5182
Social Sciences Major	381	15,922	0.3284	0.1376	0.1542	0.5068
Other Major	411	27,934	0.3089	0.1148	0.1612	0.4421
Business Major	424	33,417	0.2724	0.1014	0.1467	0.3830
STEM Major	441	36,431	0.2609	0.1023	0.1319	0.3760

(c) Movers, Related occupation

Education Level	Cells	Individuals	Mean	Std. Dev.	10th Percentile	90th Percentile
Education Major	637	7,469	0.1728	0.1952	0.0089	0.4698
Social Sciences Major	783	17,976	0.1031	0.1266	0.0088	0.2839
Other Major	799	33,598	0.1009	0.1259	0.0088	0.2859
Business Major	860	47,331	0.1269	0.1521	0.0083	0.3484
STEM Major	885	67,631	0.1167	0.1416	0.0103	0.3259

(d) Movers, Unrelated occupation

Education Level	Cells	Individuals	Mean	Std. Dev.	10th Percentile	90th Percentile
Education Major	627	4,439	0.0906	0.1162	0.0050	0.2603
Social Sciences Major	743	16,316	0.0831	0.1099	0.0067	0.2418
Other Major	777	26,769	0.0789	0.1024	0.0065	0.2283
Business Major	761	26,243	0.0750	0.0986	0.0054	0.2244
STEM Major	810	36,942	0.0694	0.0905	0.0056	0.1999

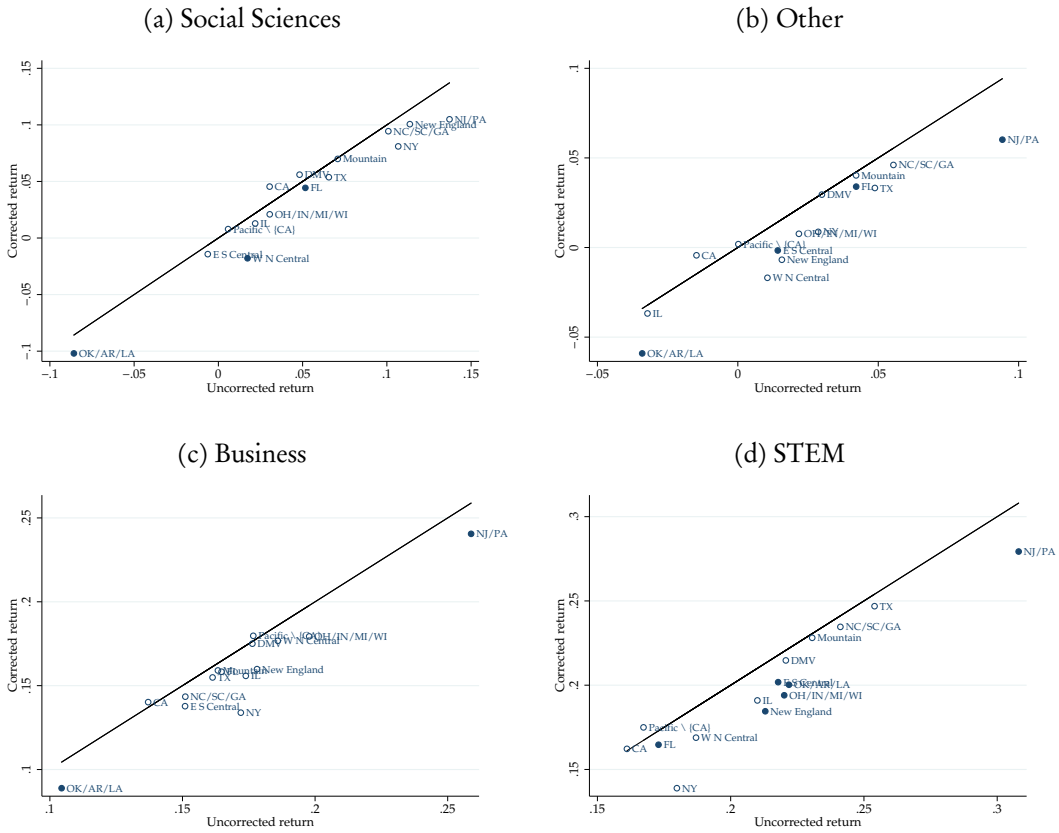
Note: Estimated decision probabilities and cell structure from the conditional inference recursive partitioning algorithm described in Section 5.1.2. Probabilities correspond to the probability of making the decision that is observed in the data. Source: Author's calculations from American Community Survey, 2010-2015.

Table 3: Uncorrected vs. corrected earnings equation estimates for select states

	Florida				New York				Texas			
	Unrelated Occupation		Related Occupation		Unrelated Occupation		Related Occupation		Unrelated Occupation		Related Occupation	
	Uncorrected	Corrected	Uncorrected	Corrected	Uncorrected	Corrected	Uncorrected	Corrected	Uncorrected	Corrected	Uncorrected	Corrected
Bachelor's degree												
Education major	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Social sciences major	0.052	0.044	0.313***	0.312***	0.107***	0.081*	0.212***	0.181***	0.066*	0.054	0.156***	0.153***
	(0.042)	(0.059)	(0.037)	(0.043)	(0.041)	(0.045)	(0.053)	(0.064)	(0.036)	(0.043)	(0.029)	(0.041)
Other major	0.042	0.034	0.279***	0.276***	0.029	0.009	0.149***	0.108*	0.049	0.033	0.133***	0.130***
	(0.040)	(0.049)	(0.032)	(0.032)	(0.038)	(0.038)	(0.050)	(0.061)	(0.034)	(0.036)	(0.024)	(0.032)
Business major	0.165***	0.158***	0.500***	0.500***	0.172***	0.134***	0.467***	0.413***	0.161***	0.155***	0.380***	0.385***
	(0.040)	(0.051)	(0.030)	(0.037)	(0.039)	(0.045)	(0.049)	(0.065)	(0.033)	(0.035)	(0.023)	(0.034)
STEM major	0.173***	0.165***	0.460***	0.456***	0.180***	0.139***	0.393***	0.333***	0.254***	0.247***	0.344***	0.349***
	(0.040)	(0.044)	(0.031)	(0.035)	(0.039)	(0.040)	(0.050)	(0.063)	(0.033)	(0.041)	(0.023)	(0.033)
Advanced degree (interaction)												
Education major	0.135	0.133	0.114	0.093	0.048	-0.061	0.130*	0.052	0.150*	0.140	-0.125**	-0.140**
	(0.105)	(0.100)	(0.074)	(0.084)	(0.079)	(0.100)	(0.067)	(0.087)	(0.084)	(0.098)	(0.058)	(0.054)
Social sciences major	0.165*	0.167**	0.198***	0.175*	0.184***	0.111	0.210***	0.157*	0.089	0.086	0.016	-0.003
	(0.088)	(0.073)	(0.066)	(0.092)	(0.063)	(0.081)	(0.051)	(0.086)	(0.072)	(0.090)	(0.052)	(0.073)
Other major	0.102	0.107	0.176***	0.154**	0.122**	0.049	0.154***	0.120	0.029	0.029	-0.044	-0.063
	(0.086)	(0.075)	(0.063)	(0.075)	(0.060)	(0.078)	(0.047)	(0.076)	(0.071)	(0.083)	(0.049)	(0.063)
Business major	0.097	0.099	0.205***	0.181**	0.136**	0.074	0.236***	0.211***	0.071	0.071	0.051	0.028
	(0.086)	(0.073)	(0.061)	(0.074)	(0.061)	(0.085)	(0.045)	(0.071)	(0.068)	(0.085)	(0.046)	(0.062)
STEM major	0.149*	0.154**	0.288***	0.265***	0.250***	0.183**	0.135***	0.097	0.108	0.101	0.098**	0.076
	(0.085)	(0.070)	(0.061)	(0.083)	(0.060)	(0.081)	(0.045)	(0.060)	(0.067)	(0.090)	(0.046)	(0.057)
Born here	-0.063***	-0.056**	-0.056***	0.020	-0.110***	-0.048	-0.113***	0.014	-0.079***	0.004	-0.077***	0.008
	(0.014)	(0.026)	(0.012)	(0.021)	(0.012)	(0.047)	(0.010)	(0.028)	(0.010)	(0.051)	(0.008)	(0.017)
Cubic in experience	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Demographics	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CBSA fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Wald test for λ terms		4.79		3.37		4.87		13.87		5.80		6.92
		[0.000]		[0.002]		[0.000]		[0.000]		[0.000]		[0.000]
R^2	0.171	0.174	0.216	0.218	0.231	0.236	0.237	0.244	0.209	0.212	0.233	0.235
Observations	10,626	10,626	15,984	15,984	14,878	14,878	22,210	22,210	16,883	16,883	26,591	26,591

Note: Standard errors are listed below coefficients in parentheses. *p*-values of statistical tests are listed below test statistics in brackets. *** $p<0.01$; ** $p<0.05$; * $p<0.10$.

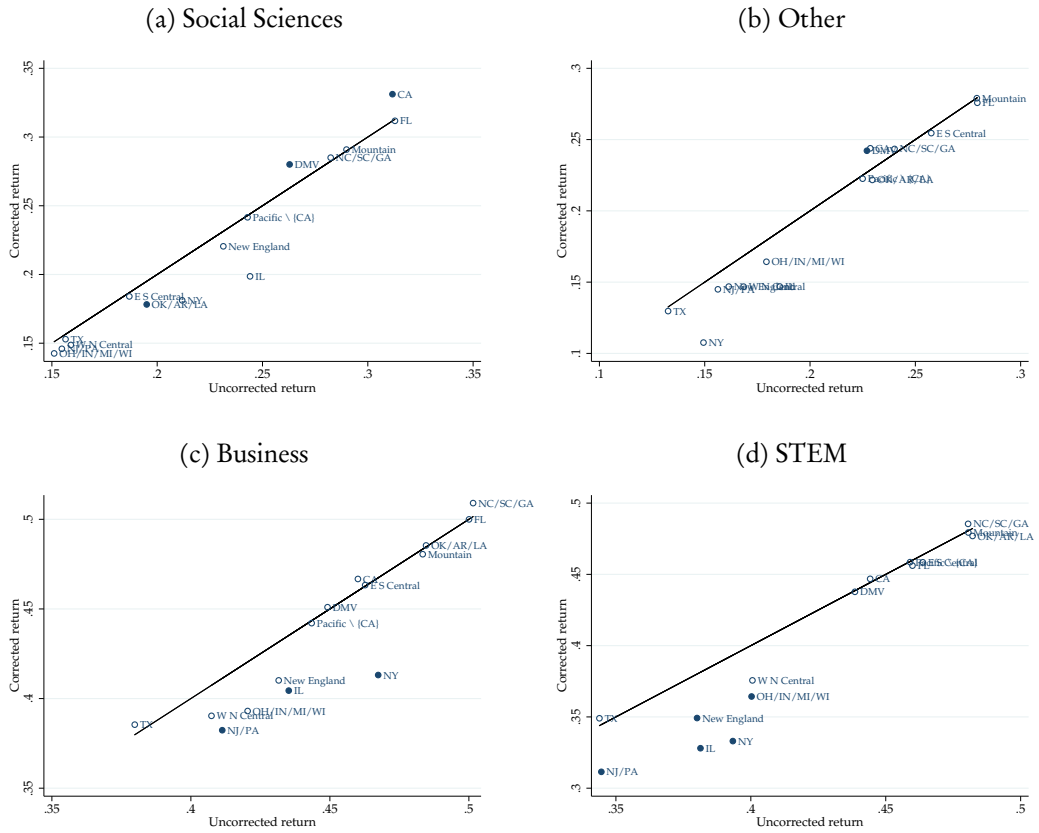
Figure 6: Scatter plots of uncorrected and corrected returns to major and working in an unrelated occupation



Notes: Scatter plots of return to major for those working in an unrelated occupation. Solid black lines are 45-degree lines. Blue circles or dots are state-specific pairs marking the uncorrected and corrected returns. Solid dots indicate statistically significant difference at the 90% level or higher.

Source: Author's calculations from American Community Survey, 2010-2015.

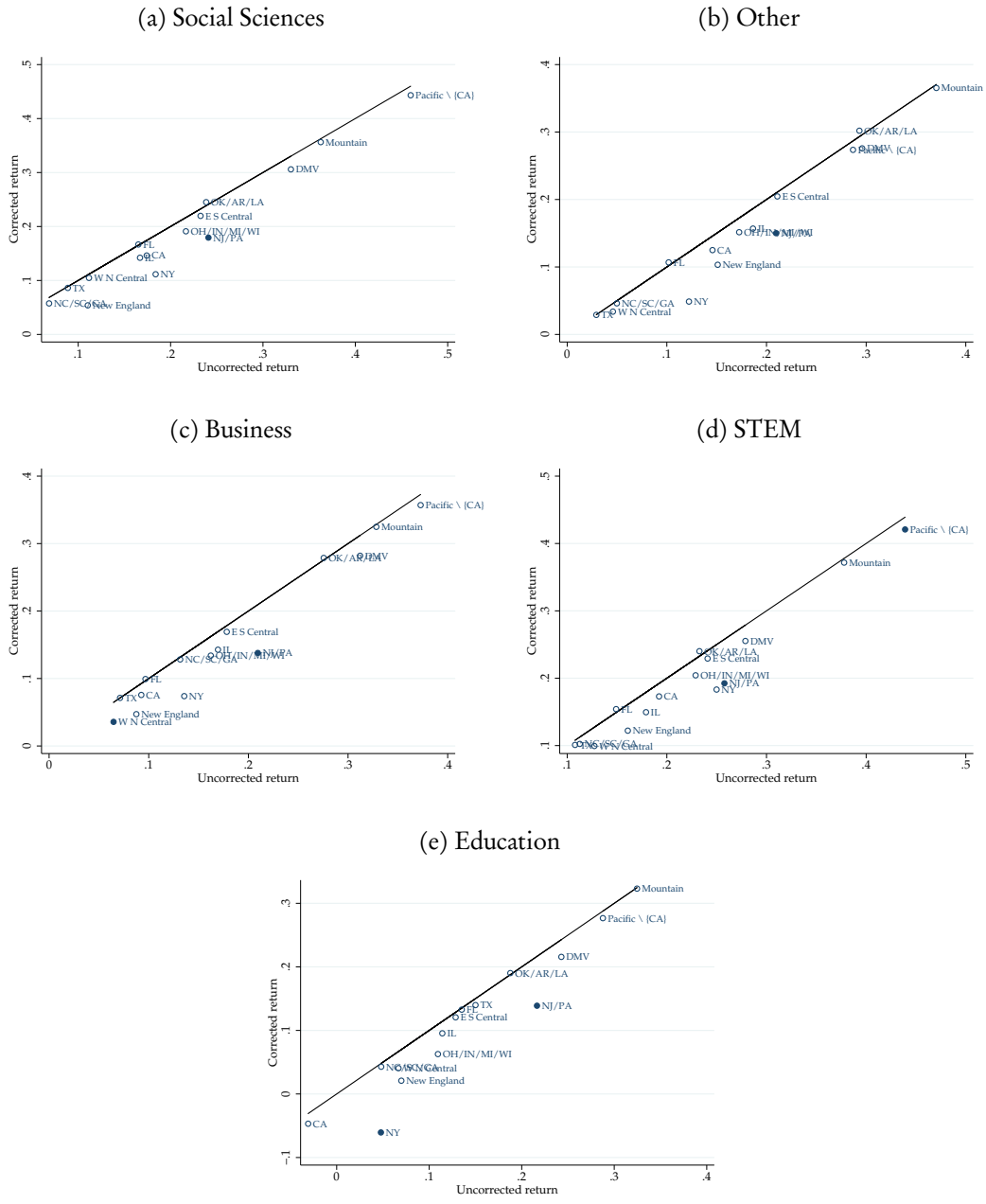
Figure 7: Scatter plots of uncorrected and corrected returns to major and working in a related occupation



Notes: Scatter plots of return to major for those working in an related occupation. Solid black lines are 45-degree lines. Blue dots are state-specific pairs marking the uncorrected and corrected returns.

Source: Author's calculations from American Community Survey, 2010-2015.

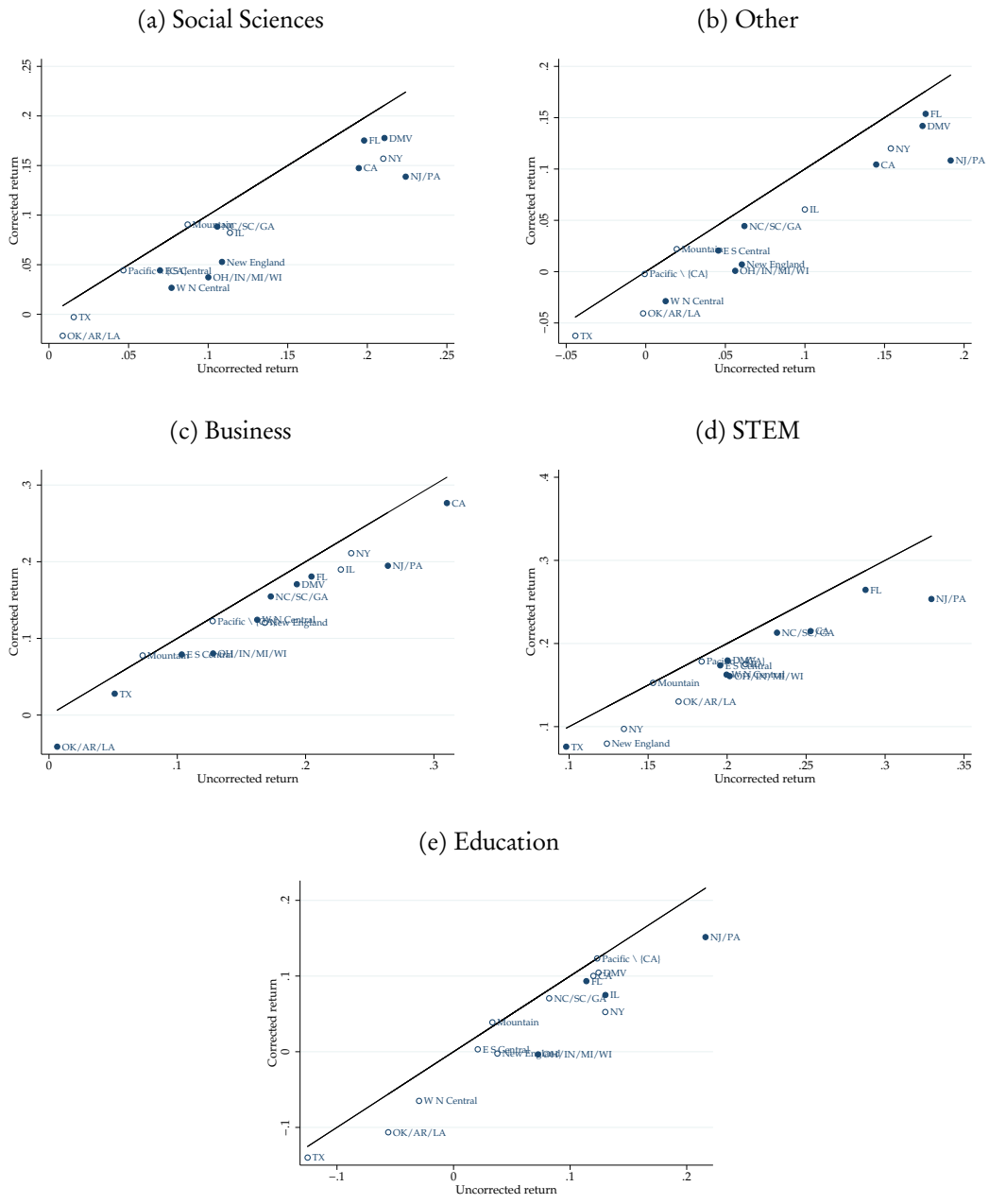
Figure 8: Scatter plots of uncorrected and corrected returns to major and working in an unrelated occupation, adv. degree holders



Notes: Scatter plots of return to major for those working in an unrelated occupation. Solid black lines are 45-degree lines. Blue circles or dots are state-specific pairs marking the uncorrected and corrected returns. Solid dots indicate statistically significant difference at the 90% level or higher.

Source: Author's calculations from American Community Survey, 2010-2015.

Figure 9: Scatter plots of uncorrected and corrected returns to major and working in a related occupation, adv. degree holders



Notes: Scatter plots of return to major for those working in an related occupation. Solid black lines are 45-degree lines. Blue circles or dots are state-specific pairs marking the uncorrected and corrected returns. Solid dots indicate statistically significant difference at the 90% level or higher.

Source: Author's calculations from American Community Survey, 2010-2015.

Table 4: Percent change in returns when correcting for selection

	Unrelated occupation			Related occupation		
Major	p10	Median	p90	p10	Median	p90
Bachelor's degrees						
Education	0	0	0	0	0	0
Social Sciences	-126.5	-18	36.1	-14.5	-2.3	6.2
Other	-143.5	-31.9	70.5	-20.9	-2.2	6.7
Business	-14.9	-5.1	1.7	-7.1	-.3	1.5
STEM	-13.4	-7.4	0.7	-14	-1.1	1.0
Advanced degrees						
Education	-70.1	-11.2	-0.5	-106.5	-29.9	0.1
Social Sciences	-39.3	-7.4	0.9	-118.4	-27.7	-5.0
Other	-31.6	-7.5	3.0	-330.5	-41.3	-12.6
Business	-45.5	-9.7	1.1	-45.5	-16.6	-4.0
STEM	-25.4	-9	3.2	-27.8	-17.3	-3.0

Note: Summary statistics of the 15-location distribution of the percent change between uncorrected and corrected returns to majors. Percentage changes are least informative for education, social science, and other majors because these majors have bases (i.e. uncorrected returns) that may be very close to zero.

Table 5: Determinants of cross-location migration flows among majors

	STEM		Business		Other		Social Science	
Dep. variable: $\ln\left(p^{m}_{aj\ell}\right) - \ln\left(p^{\text{Edu}}_{aj\ell}\right)$	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\Delta^m\Delta^\ell$ Corrected return, unrelated occ. × BA	0.004	0.063	0.744***	0.404**	0.793***	0.482***	1.224***	0.395***
	(0.190)	(0.180)	(0.146)	(0.165)	(0.175)	(0.152)	(0.119)	(0.136)
$\Delta^m\Delta^\ell$ Corrected return, related occ. × BA	0.004	0.063	0.744***	0.404**	0.793***	0.482***	1.224***	0.395***
	(0.190)	(0.180)	(0.146)	(0.165)	(0.175)	(0.152)	(0.119)	(0.136)
$\Delta^m\Delta^\ell$ Corrected return, unrelated occ. × Adv. deg.	0.004	0.063	0.744***	0.404**	0.793***	0.482***	1.224***	0.395***
	(0.190)	(0.180)	(0.146)	(0.165)	(0.175)	(0.152)	(0.119)	(0.136)
$\Delta^m\Delta^\ell$ Corrected return, related occ. × Adv. deg.	0.004	0.063	0.744***	0.404**	0.793***	0.482***	1.224***	0.395***
	(0.190)	(0.180)	(0.146)	(0.165)	(0.175)	(0.152)	(0.119)	(0.136)
$\ln\left(\text{distance}_{jk}\right)$ × BA	0.033	0.031	-0.019	-0.020	0.040	0.038*	0.024	0.024
	(0.023)	(0.020)	(0.018)	(0.016)	(0.027)	(0.020)	(0.024)	(0.022)
$\ln\left(\text{distance}_{jk}\right)$ × Advanced degree	0.130***	0.128***	0.037	0.042*	0.109**	0.107***	0.087**	0.087**
	(0.039)	(0.030)	(0.026)	(0.022)	(0.045)	(0.037)	(0.043)	(0.034)
$\Delta^m\Delta^\ell$ ln (share related occ.) × BA	0.265***	-0.131	0.457***	0.306**	1.582***	1.112***	0.823***	0.761***
	(0.097)	(0.136)	(0.057)	(0.121)	(0.083)	(0.169)	(0.064)	(0.131)
$\Delta^m\Delta^\ell$ ln (share related occ.) × Adv. deg.	1.108***	1.451***	1.868***	0.933***	-0.218	-0.516*	-0.434	-0.272
	(0.335)	(0.355)	(0.183)	(0.216)	(0.328)	(0.309)	(0.294)	(0.264)
Advanced degree	-0.488	-0.502**	-0.250	-0.291	-0.308	-0.328	-0.296	-0.302
	(0.299)	(0.247)	(0.209)	(0.187)	(0.350)	(0.287)	(0.331)	(0.280)
Constant	-0.097	-0.073	0.173	0.184*	-0.136	-0.106	-0.007	-0.002
	(0.149)	(0.136)	(0.118)	(0.109)	(0.181)	(0.137)	(0.159)	(0.152)
Climate measures		✓		✓		✓		✓
Quality of life measures		✓		✓		✓		✓
Local spending measures		✓		✓		✓		✓
Wald test for joint significance of amenity variables		24.60		20.17		38.37		23.93
		[0.000]		[0.000]		[0.000]		[0.000]
R^2	0.132	0.550	0.293	0.576	0.252	0.644	0.249	0.586
Observations	420	420	420	420	420	420	420	420

Note: Regression of cross-major log differences in migration flows among advanced degree group a from location j to ℓ on returns to major, distance, availability of related occupations, and local amenity measures. Huber-White standard errors are listed below coefficients in parentheses. $\Delta^m\Delta^\ell$ signifies the difference-in-differences operator, where differences are taken between majors m and Education, and between locations j and ℓ. Distance is measured in miles between population centroids using the Great Circle formula. Amenity variables are included as log differences, where the difference is taken between locations j and ℓ. Climate measures include cloudiness, average wind speed, heating degree days, cooling degree days, morning humidity, and precipitation. Quality of life variables include per-pupil schooling expenditures, population density, health care expenditures per capita, and violent crime rates. Local spending measures include state budget expenditures per capita and higher education expenditures per full-time equivalent student. State level variables are aggregated to regional level weighting by component state populations. Number of observations equals $2\left(L^2 - L\right)$. *** $p<0.01$; ** $p<0.05$; * $p<0.10$.

A Monte Carlo Simulation

In this section I detail the Monte Carlo simulation used to compare the performance of the conditional inference tree estimator with more traditional estimators.

A.1 Data generating process

Consider the following data generating process, structured similar to the model in Section 2.

$$w_{i\ell k} = x_i\gamma_{1\ell k} + s_i\gamma_{2\ell k} + \eta_{i\ell k} \tag{A.1}$$

$$u_{ij\ell k} = z_i\phi_{j\ell k} + \varepsilon_{ij\ell k} \tag{A.2}$$

$$V_{ij\ell k} = w_{i\ell k} + u_{ij\ell k} \tag{A.3}$$

$$w_{i\ell k} \text{ observed} \iff V_{ij\ell k} > V_{ij\ell' k'} \ \forall\left(\ell', k'\right) \tag{A.4}$$

In the baseline model, I consider the case where ℓ comes from a 15-dimensional set, and where k is two-dimensional. Thus, there are 30 sectors in the model. For simplicity, s_i is a binary variable while x_i contains a mixture of binary and continuous variables. z_i contains a number of binary variables as well as two continuous exclusion restrictions which measure preference intensity for staying in the birth location and for working in the related occupation. In addition, z_i contains a number of interactions among this set of variables. $\eta_{i\ell k}$ is assumed to be distributed iid $N(0,1)$ across all individuals, locations, and occupations. The same is true for $\varepsilon_{ij\ell k}$. In later simulations, I examine performance of the estimator when these error terms are correlated across locations and occupations.

The estimate of interest is $\hat{\gamma}_2$ in location 8 and occupation 2, which is chosen without loss of generality. The true value of this parameter is set to 2. I consider estimation of $\gamma_{2,8,2}$ in small samples ($N = 1{,}000$ per sector) and large samples ($N = 10{,}000$ per sector). Each simulation is repeated 100 times, and I report the resulting mean and standard deviation of the parameter estimates, along with the average root mean square error of each repetition.

I report the performance of nine different specifications under three different error structures. As a baseline, I include the naive OLS estimator that would be unbiased and consistent if no selection were present. I then consider four separate estimates of the selection probabilities in the polynomial selection terms. For each estimate, I consider including only the first-best probability, or the first-best and location probabilities as implemented in the empirical section of the paper. The four different probability estimators are as follows: (i) fully specified bin; (ii) conditional inference tree; (iii) logit; and (iv) coarse (misspecified) bin. I specifically include the coarse bin estimator to show the effect of the researcher being unable to include all relevant choice predictors, e.g. due to the curse of dimensionality. The three different error structures I consider are as follows: (i) the baseline described above; (ii) allowing the preference shocks to be

correlated across locations and occupations (i.e. $\varepsilon_{ij\ell k}$ distributed iid $N(0, \Sigma)$ across individuals, where Σ is a random covariance matrix); and (iii) allowing both preference shocks and earnings shocks to be multivariate normal distributions.

The results of the simulations are reported in Table A1. Each of the three error structures are reported respectively in Panels A, B, and C of the table. Within each panel are the nine different specifications used to estimate $\gamma_{2,8,2}$. The main takeaway from the simulations is that the tree algorithm performs very similarly to the fully specified bin estimator in large samples, but that the tree algorithm performs much better than all other estimators in small samples. The improved small-sample performance of the tree algorithm is consistent with Asher et al. (2016), who prove the consistency of tree classification and also show excellent small sample performance. For all specifications, the OLS estimate of the parameter of interest is severely downward biased, while the logit estimate is severely upward biased. The coarse bin estimator performs only slightly better than OLS and incurs a high efficiency cost.

The purpose of Panels A and B is to show that the nonparametric estimator used in this paper performs well when the distribution of preference shocks is either normal or multivariate normal. In both of these two scenarios, index sufficiency holds. In Panel C, however, index sufficiency is less likely to hold. In this case, none of the estimators is able to completely resolve the selection problem. However, the tree estimator performs best, again particularly in smaller samples.

Table A1: Monte Carlo simulation results (true parameter value equals 2)

	10,000 Observations per Sector				1,000 Observations per Sector			
	Mean	Std. Dev.	RMSE	Ave. Sample Size	Mean	Std. Dev.	RMSE	Ave. Sample Size
Panel A: 30 sectors, baseline								
OLS	1.6219	0.0308	0.8994	28502	1.6181	0.0984	0.8974	2857
1st Best Bin	1.9595	0.0304	0.8726		1.9281	0.0967	0.8724	
1st Best Tree	1.9597	0.0309	0.8701		2.0147	0.0991	0.8722	
1st Best Logit	2.1458	0.0336	0.8687		2.1401	0.1045	0.8675	
1st Best Coarse Bin	1.8505	0.0344	0.8959		1.8389	0.1071	0.8942	
1st+2nd Best Bin	1.9413	0.0332	0.8723		1.9085	0.1047	0.8720	
1st+2nd Best Tree	1.9468	0.0327	0.8700		2.0021	0.1066	0.8715	
1st+2nd Best Logit	2.1523	0.0409	0.8684		2.1734	0.1248	0.8673	
1st+2nd Best Coarse Bin	1.6186	0.1148	0.8944		1.7443	0.2826	0.8931	
Panel B: 30 sectors, $\varepsilon_{ij\ell k}$ correlated across (ℓ, k)								
OLS	1.6595	0.0311	0.9123	27504	1.6929	0.0875	0.9136	2755
1st Best Bin	1.9616	0.0340	0.8929		1.9576	0.0878	0.8949	
1st Best Tree	1.9592	0.0330	0.8914		2.0265	0.0902	0.8951	
1st Best Logit	2.1152	0.0361	0.8903		2.1405	0.0906	0.8912	
1st Best Coarse Bin	1.8819	0.0372	0.9095		1.8991	0.0974	0.9110	
1st+2nd Best Bin	1.9451	0.0376	0.8927		1.9468	0.1050	0.8947	
1st+2nd Best Tree	1.9422	0.0352	0.8913		2.0054	0.1106	0.8948	
1st+2nd Best Logit	2.1131	0.0431	0.8901		2.1627	0.1319	0.8909	
1st+2nd Best Coarse Bin	1.6737	0.1149	0.9085		1.8045	0.2774	0.9100	
Panel C: 30 sectors, both $\varepsilon_{ij\ell k}$ and $\eta_{i\ell k}$ correlated across (ℓ, k)								
OLS	1.5404	0.0943	1.0715	26613	1.5386	0.1265	1.0681	2676
1st Best Bin	1.9428	0.0513	1.0437		1.8828	0.1055	1.0429	
1st Best Tree	1.9394	0.0508	1.0417		1.9709	0.1080	1.0427	
1st Best Logit	2.1334	0.0554	1.0400		2.1160	0.1184	1.0376	
1st Best Coarse Bin	1.8703	0.0662	1.0670		1.8359	0.1362	1.0641	
1st+2nd Best Bin	1.9232	0.0534	1.0433		1.8804	0.1230	1.0423	
1st+2nd Best Tree	1.9207	0.0529	1.0414		1.9572	0.1285	1.0419	
1st+2nd Best Logit	2.1329	0.0598	1.0397		2.1474	0.1698	1.0372	
1st+2nd Best Coarse Bin	1.5805	0.1352	1.0658		1.7196	0.3794	1.0630	

Note: 100 replications used for all specifications. "OLS" indicates OLS estimation of the parameter of interest, ignoring potential selection bias. "1st Best Bin" indicates estimation of equation (3.4) using a cubic polynomial of the first-best probability from a simple bin estimator. "1st+2nd Best Bin" indicates the same, except that both the first-best and occupation probabilities are used, as described in Section 5.2. The polynomial is a full set of third-degree polynomial terms, including interactions. "Tree" refers to estimation using probabilities from the conditional inference tree algorithm described in Section 5.1.2. "Logit" indicates estimation using probabilities from a logit model. "Coarse Bin" refers to estimation using probabilities from a more coarsely defined bin estimator, as would be required in the empirical application of this paper.

B Data Appendix

This section describes additional details relating to the construction of the earnings and demographic variables used in the analysis.

Race and ethnicity I construct a measure of race and ethnicity by first assigning anyone of Hispanic origin to be Hispanic, and then assigning race based on whether the reported race is white, black, or other. Mixed-race individuals are classified as other.

Earnings and employment Earnings are measured as the individual's annual wage and salary income, expressed in constant 2010 dollars. I drop any nominal earnings measurements greater than $600,000 or less than $20,000. I classify a person as employed if they reported being employed at the time of the survey. I also create a variable indicating if the individual's spouse is employed.

Work experience I define work experience as potential experience in the usual way: age minus number of years of schooling minus 6.

Birth place I create separate variables indicating in which state the individual was born, and in which state the individual's spouse was born (if applicable).

Marital status and household composition Marital status is self-reported in the survey as one of six categories. I aggregate these categories into three: married (whether or not residing with spouse); divorced or separated; and single or widowed. Number of co-resident children is given in the survey and I distill this information into two dummies: one or more children under the age of 5; and one or more children under the age of 18. Family co-residence status is distilled into one dummy variable indicating whether the individual is in the same household as any relative. The relationship can be blood, or through marriage.

Dwelling characteristics Home ownership status is divided into "owned" or "rented."

Table B1: Sample selection details

Criterion	No. obs deleted	Remaining obs.
Respondents in 2010-2015 ACS	—	18,699,149
Drop those without a bachelor's degree or higher	14,689,233	4,009,916
Drop those outside of 22-54 age range	1,547,395	2,462,521
Drop those currently enrolled in school	269,606	2,192,915
Drop those currently residing in group quarters	13,752	2,179,163
Drop those not born in the US	386,866	1,792,297
Drop those with positive annual earnings below $20,000	196,246	1,596,051
Drop those with annual earnings above $600,000	1,015	1,595,036
Drop those with zero annual earnings	212,871	1,382,165
Drop females	698,912	683,253
Drop those with imputed earnings or occupations	97,994	585,259
Drop those with imputed labor force status	1,346	583,913
Final analysis sample	—	583,913

Table B2: Aggregation of the 51 detailed Department of Education majors

Education	*STEM*	*Other*
Primary Education	Agriculture and Agr. Science	Architecture
Secondary Education	All Other Engineering	Area, Ethnic, and Civ. Studies
	Biological Sciences	Art History and Fine Arts
Social Sciences	Chemical Engineering	Commercial Art and Design
Family and Consumer Science	Chemistry	Communications
International Relations	Civil Engineering	Film and Other Arts
Other Social Science	Computer Programming	Foreign Language
Philosophy and Religion	Computer and Info Tech	History
Political Science	Earth and Other Physical Sci	Journalism
Psychology	Electrical Engineering	Leisure Studies
Social Work and HR	Engineering Tech	Letters: Lit, Writing, Other
	Environmental Studies	Music and Speech/Drama
Business	Fitness and Nutrition	Prec. Prod. and Ind. Arts
Accounting	General Science	Protective Services
Business Mgt. and Admin.	Mathematics	Public Admin and Law
Economics	Mechanical Engineering	Public Health
Finance	Medical Tech	
Marketing	Nursing	
Misc. Bus. and Med. Support	Other Med/Health Services	
	Physics	

Note: Aggregation of the 51 detailed Department of Education majors analyzed in Altonji et al. (2016b).

Table B3: List of frequent occupations for select majors: Non-advanced degree holders

(a) Primary Education

Occupation	Share(%)
Primary school teachers	31.21
Secondary school teachers	13.26
Managers and administrators, n.e.c.	4.73
Salespersons, n.e.c.	3.12
Supervisors and proprietors of sales jobs	2.94
Teachers , n.e.c.	2.06
Police, detectives, and private investigators	1.75
Retail sales clerks	1.54
Computer systems analysts and computer scientists	1.33

(b) History

Occupation	Share(%)
Managers and administrators, n.e.c.	9.41
Supervisors and proprietors of sales jobs	5.41
Salespersons, n.e.c.	5.35
Primary school teachers	4.67
Military	3.62
Computer systems analysts and computer scientists	3.33
Police, detectives, and private investigators	3.11
Secondary school teachers	2.78
Managers and specialists in marketing, advertising, and public relations	2.76
Retail sales clerks	2.12
Other financial specialists	2.07
Customer service reps, investigators and adjusters, except insurance	1.98
Chief executives and public administrators	1.94
Financial managers	1.54

(c) Economics

Occupation	Share(%)
Managers and administrators, n.e.c.	11.32
Other financial specialists	8.53
Salespersons, n.e.c.	7.22
Supervisors and proprietors of sales jobs	5.42
Financial managers	4.87
Accountants and auditors	4.74
Computer systems analysts and computer scientists	4.6
Financial services sales occupations	4.02
Chief executives and public administrators	3.98
Managers and specialists in marketing, advertising, and public relations	2.98
Management analysts	2.71
Computer software developers	2.23
Retail sales clerks	1.8
Customer service reps, investigators and adjusters, except insurance	1.8
Insurance sales occupations	1.58

(d) Computer Programming

Occupation	Share(%)
Computer software developers	41.25
Computer systems analysts and computer scientists	18.22
Managers and administrators, n.e.c.	6.9
Managers and specialists in marketing, advertising, and public relations	1.81
Chief executives and public administrators	1.46
Supervisors and proprietors of sales jobs	1.39

Notes: Tables list occupations within the given major that are above the 2% cutoff defining relatedness, along with three additional occupations below the cutoff.

Table B4: List of frequent occupations for select majors: Advanced degree holders

(a) Primary Education

Occupation	Share(%)
Primary school teachers	39.73
Managers in education and related fields	15.54
Secondary school teachers	14.43
Subject instructors (HS/college)	3.85
Managers and administrators, n.e.c.	2.29
Special education teachers	1.65
Clergy and religious workers	1.65
Vocational and educational counselors	1.58

(b) History

Occupation	Share(%)
Lawyers	23.8
Primary school teachers	9.34
Subject instructors (HS/college)	6.38
Managers and administrators, n.e.c.	6.26
Secondary school teachers	5.25
Managers in education and related fields	4.2
Physicians	2.53
Military	2.5
Clergy and religious workers	2.27
Chief executives and public administrators	2.14
Computer systems analysts and computer scientists	1.63
Other financial specialists	1.62
Financial managers	1.51

(c) Economics

Occupation	Share(%)
Lawyers	18.77
Managers and administrators, n.e.c.	9.91
Financial managers	6.55
Other financial specialists	5.78
Accountants and auditors	5.11
Chief executives and public administrators	4.46
Management analysts	3.73
Subject instructors (HS/college)	3.52
Supervisors and proprietors of sales jobs	2.68
Computer systems analysts and computer scientists	2.67
Physicians	2.61
Salespersons, n.e.c.	2.47
Economists, market researchers, and survey researchers	2.43
Managers and specialists in marketing, advertising, and public relations	2.36
Financial services sales occupations	2.05
Primary school teachers	1.62
Managers in education and related fields	1.32
Computer software developers	1.1

(d) Computer Programming

Occupation	Share(%)
Computer software developers	44.97
Primary school teachers	9.38
Supervisors and proprietors of sales jobs	5.98
Computer systems analysts and computer scientists	5.71
Industrial engineers	5.1
Designers	3.86
Auto body repairers	3.48
Retail sales clerks	3.23
Managers and administrators, n.e.c.	3.15
Salespersons, n.e.c.	2.33
Customer service reps, investigators and adjusters, except insurance	2.1
Electrical engineer	1.82
Editors and reporters	1.8
Subject instructors (HS/college)	1.74

Notes: Tables list occupations within the given major that are above the 2% cutoff defining relatedness, along with three additional occupations below the cutoff.

Table B5: Complete list of related occupations by major: Non-advanced degree holders

Occupation	Edu.	Soc. Sci.	Other	Bus.	STEM
Chief executives and public administrators		✓	✓	✓	✓
Financial managers			✓	✓	✓
Human resources and labor relations managers		✓	✓		
Managers and specialists in marketing, advertising, and public relations		✓	✓	✓	✓
Managers of medicine and health occupations					✓
Managers of food-serving and lodging establishments			✓		
Funeral directors			✓		
Managers of service organizations, n.e.c.			✓		
Managers and administrators, n.e.c.	✓	✓	✓	✓	✓
Accountants and auditors				✓	✓
Other financial specialists		✓	✓	✓	✓
Management analysts		✓	✓	✓	✓
Personnel, HR, training, and labor relations specialists		✓	✓		
Inspectors and compliance officers, outside construction			✓		
Architects			✓		
Aerospace engineer					✓
Metallurgical and materials engineers, variously phrased					✓
Chemical engineers					✓
Civil engineers			✓		✓
Electrical engineer					✓
Industrial engineers					✓
Mechanical engineers					✓
Not-elsewhere-classified engineers					✓
Computer systems analysts and computer scientists		✓	✓	✓	✓
Actuaries					✓
Chemists					✓
Atmospheric and space scientists					✓
Geologists					✓
Physical scientists, n.e.c.					✓
Agricultural and food scientists					✓
Biological scientists					✓
Foresters and conservation scientists					✓
Registered nurses					✓
Pharmacists					✓
Respiratory therapists					✓
Occupational therapists					✓
Physical therapists					✓
Therapists, n.e.c.					✓
Primary school teachers	✓	✓	✓		✓
Secondary school teachers	✓		✓		✓
Teachers , n.e.c.	✓		✓		
Vocational and educational counselors		✓			
Economists, market researchers, and survey researchers				✓	
Social workers		✓			
Recreation workers					✓
Clergy and religious workers		✓	✓		
Writers and authors			✓		
Designers			✓		
Musician or composer			✓		
Actors, directors, producers			✓		
Art makers: painters, sculptors, craft-artists, and print-makers			✓		
Photographers			✓		
Editors and reporters			✓		
Athletes, sports instructors, and officials					✓
Clinical laboratory technologies and technicians					✓
Radiologic tech specialists					✓
Health technologists and technicians, n.e.c.			✓		✓
Engineering technicians, n.e.c.					✓
Drafters			✓		
Surveyors, cartographers, mapping scientists and technicians					✓
Chemical technicians					✓
Airplane pilots and navigators			✓		
Air traffic controllers			✓		
Computer software developers		✓	✓	✓	✓
Legal assistants, paralegals, legal support, etc			✓		
Supervisors and proprietors of sales jobs	✓	✓	✓	✓	✓
Insurance sales occupations				✓	
Financial services sales occupations				✓	
Salespersons, n.e.c.	✓	✓	✓	✓	✓
Retail sales clerks		✓	✓	✓	✓
Office supervisors			✓		
Customer service reps, investigators and adjusters, except insurance		✓	✓	✓	
Fire fighting, prevention, and inspection			✓		✓
Police, detectives, and private investigators		✓	✓		✓
Other law enforcement: sheriffs, bailiffs, correctional institution officers			✓		
Guards, watchmen, doorkeepers			✓		
Waiter/waitress			✓		
Cooks, variously defined		✓	✓		
Welfare service aides		✓	✓		
Farmers (owners and tenants)					✓
Farm workers					✓
Supervisors of agricultural occupations					✓
Gardeners and groundskeepers					✓
Supervisors of construction work			✓		
Production supervisors or foremen					✓
Military		✓	✓		✓

Note: Occupations not related to any college major are excluded from this table.

Table B6: Complete list of related occupations by major: Advanced degree holders

Occupation	Edu.	Soc. Sci.	Other	Bus.	STEM
Chief executives and public administrators	✓	✓	✓	✓	✓
Financial managers		✓	✓	✓	✓
Human resources and labor relations managers		✓			
Managers and specialists in marketing, advertising, and public relations		✓	✓	✓	✓
Managers in education and related fields	✓	✓	✓	✓	✓
Managers of medicine and health occupations		✓			✓
Managers of food-serving and lodging establishments			✓		
Managers of service organizations, n.e.c.		✓	✓		
Managers and administrators, n.e.c.	✓	✓	✓	✓	✓
Accountants and auditors			✓	✓	
Other financial specialists		✓		✓	✓
Management analysts		✓	✓	✓	✓
Personnel, HR, training, and labor relations specialists		✓	✓		
Architects			✓		
Aerospace engineer			✓		✓
Chemical engineers					✓
Civil engineers					✓
Electrical engineer					✓
Industrial engineers					✓
Mechanical engineers					✓
Not-elsewhere-classified engineers					✓
Computer systems analysts and computer scientists		✓	✓	✓	✓
Operations and systems researchers and analysts		✓			
Actuaries					✓
Mathematicians and mathematical scientists					✓
Physicists and astronomers					✓
Chemists					✓
Atmospheric and space scientists					✓
Geologists					✓
Physical scientists, n.e.c.					✓
Agricultural and food scientists					✓
Biological scientists					✓
Foresters and conservation scientists					✓
Medical scientists					✓
Physicians		✓	✓	✓	✓
Dentists		✓			✓
Veterinarians					✓
Other health and therapy					✓
Registered nurses					✓
Pharmacists					✓
Physical therapists					✓
Speech therapists					✓
Therapists, n.e.c.		✓			
Physicians assistants					✓
Subject instructors (HS/college)	✓	✓	✓	✓	✓
Primary school teachers	✓	✓	✓	✓	✓
Secondary school teachers	✓	✓	✓		✓
Teachers , n.e.c.			✓		✓
Vocational and educational counselors		✓	✓		✓
Archivists and curators			✓		
Economists, market researchers, and survey researchers				✓	
Psychologists		✓			
Urban and regional planners			✓		
Social workers		✓	✓		
Clergy and religious workers		✓	✓		
Lawyers		✓	✓	✓	✓
Writers and authors			✓		
Designers			✓		✓
Musician or composer			✓		
Art makers: painters, sculptors, craft-artists, and print-makers			✓		
Editors and reporters			✓		
Athletes, sports instructors, and officials					✓
Clinical laboratory technologies and technicians					✓
Radiologic tech specialists					✓
Health technologists and technicians, n.e.c.			✓		✓
Airplane pilots and navigators			✓		
Computer software developers			✓		✓
Supervisors and proprietors of sales jobs		✓	✓	✓	✓
Financial services sales occupations				✓	
Salespersons, n.e.c.		✓	✓	✓	✓
Retail sales clerks					✓
Office supervisors			✓		
Customer service reps, investigators and adjusters, except insurance					✓
Police, detectives, and private investigators		✓	✓		
Guards, watchmen, doorkeepers			✓		
Cooks, variously defined			✓		
Nursing aides, orderlies, and attendants					✓
Welfare service aides			✓		
Farmers (owners and tenants)					✓
Auto body repairers					✓
Production supervisors or foremen					✓
Military		✓	✓	✓	✓

Note: Occupations not related to any college major are excluded from this table.

Table B7: Aggregation of locations

Location	2010 Population
California	39,144,818
OH, IN, MI, WI	33,927,016
Texas	27,469,114
NC, SC, GA	25,153,808
Mountain Census Division	23,530,498
NJ, PA	21,760,516
West North Central Census Division	21,120,392
Florida	20,271,272
New York	19,795,791
East South Central Census Division	18,876,703
WV, VA, DC, MD, DE	17,851,684
New England Census Division	14,727,584
AK, HI, OR, WA	13,369,363
Illinois	12,859,995
OK, AR, LA	11,560,266

Notes: The Mountain Census Division includes the following states: AZ, NM, CO, UT, NV, ID, MT, WY. The West North Central Census Division includes the following states: ND, SD, NE, KS, MO, IA, and MN. The East South Central Census Division is comprised of AL, MS, TN, and KY. The New England Census Division is comprised of CT, RI, MA, VT, NH, and ME.

Table B8: Predictive performance of various algorithms

	Classification algorithm		
Performance Criterion	Logit	Bin	Tree
Training set performance:			
Accuracy	37.67%	36.43%	38.85%
Kappa	34.89%	33.62%	36.13%
Test set performance:			
Accuracy	37.32%	35.29%	37.68%
Kappa	34.54%	32.43%	34.92%

Note: "Logit" refers to a flexibly specified logit; "Bin" refers to a bin estimator; "Tree" refers to the conditional inference tree classification algorithm detailed in Section 5.1.2. I estimate each algorithm on a subset of the 2010-2015 ACS sample included in this paper and compute predictive performance out-of-sample using a holdout sample. To measure predictive performance, I compute the predicted alternative, defined as the alternative with the largest predicted probability. Predictive performance is measured via a multi-dimensional confusion matrix using two related but separate metrics: Accuracy and Kappa.

$\text{Accuracy} = \frac{\text{number of correctly classified predictions}}{\text{number of predictions}}$.

$\text{Kappa} = \frac{\text{Accuracy} - \text{Expected Accuracy}}{1 - \text{Expected Accuracy}}$.

Expected Accuracy is defined as $\text{Expected Accuracy} = \sum_{j=1}^{J} \left[\left(\sum_i d_{ij} \right) \left(\sum_i p_{ij} \right) \right] / N^J$, where d_{ij} represents the observed class for observation i in the data, p_{ij} represents the predicted class for observation i, and N represents the total number of observations. The Kappa statistic is meant to capture predictive performance net of guessing. For example, the Kappa statistic penalizes strategies that would predict that all observations belong to one class (for example, such strategies could yield high accuracy for classification problems where one class is extremely rare).

Table B9: Return to STEM majors in unrelated occupation, by location (uncorrected and corrected)

State	Uncorrected STEM Return	Corrected STEM Return	χ^2 Test for Difference	Wald Test for Correction Terms
California	0.161	0.162	0.015	1.937
	(0.036)	(0.039)	[0.901]	[0.051]
Texas	0.254	0.247	0.937	5.802
	(0.033)	(0.041)	[0.333]	[0.000]
Florida	0.173	0.165	2.887	4.789
	(0.040)	(0.044)	[0.089]	[0.000]
Illinois	0.210	0.191	1.561	2.649
	(0.037)	(0.035)	[0.212]	[0.010]
New York	0.180	0.139	2.166	4.874
	(0.039)	(0.040)	[0.141]	[0.000]
New England	0.213	0.185	3.365	5.033
	(0.038)	(0.039)	[0.067]	[0.000]
New Jersey & Penn.	0.308	0.279	6.257	5.400
	(0.030)	(0.036)	[0.012]	[0.000]
WV, VA, DC, MD, DE	0.221	0.215	0.883	6.509
	(0.032)	(0.043)	[0.347]	[0.000]
NC, SC, GA	0.241	0.235	1.305	3.193
	(0.029)	(0.035)	[0.253]	[0.002]
E S Central Div	0.218	0.202	4.337	2.730
	(0.031)	(0.035)	[0.037]	[0.010]
OH, IN, MI, WI	0.220	0.194	2.850	2.484
	(0.022)	(0.032)	[0.091]	[0.012]
W N Central Div	0.187	0.169	1.004	4.240
	(0.024)	(0.030)	[0.316]	[0.000]
OK, AR, LA	0.222	0.200	5.890	3.381
	(0.040)	(0.042)	[0.015]	[0.002]
Mountain Div	0.231	0.228	0.075	2.055
	(0.030)	(0.042)	[0.784]	[0.041]
OR, WA, AK, HI	0.167	0.175	0.631	1.970
	(0.041)	(0.033)	[0.427]	[0.051]

Table B10: Return to STEM majors in related occupation, by location (uncorrected and corrected)

State	Uncorrected STEM Return	Corrected STEM Return	χ^2 Test for Difference	Wald Test for Correction Terms
California	0.444	0.447	0.044	4.905
	(0.032)	(0.033)	[0.834]	[0.000]
Texas	0.344	0.349	0.294	6.917
	(0.023)	(0.033)	[0.587]	[0.000]
Florida	0.460	0.456	0.228	3.366
	(0.031)	(0.035)	[0.633]	[0.002]
Illinois	0.381	0.328	3.319	5.491
	(0.034)	(0.041)	[0.068]	[0.000]
New York	0.393	0.333	3.631	13.871
	(0.050)	(0.063)	[0.057]	[0.000]
New England	0.380	0.349	3.545	9.652
	(0.033)	(0.040)	[0.060]	[0.000]
New Jersey & Penn.	0.345	0.311	3.003	5.590
	(0.023)	(0.030)	[0.083]	[0.000]
WV, VA, DC, MD, DE	0.439	0.438	0.016	5.582
	(0.027)	(0.030)	[0.899]	[0.000]
NC, SC, GA	0.481	0.485	0.178	3.251
	(0.024)	(0.028)	[0.673]	[0.002]
E S Central Div	0.464	0.458	0.457	1.395
	(0.027)	(0.036)	[0.499]	[0.208]
OH, IN, MI, WI	0.400	0.364	2.966	5.115
	(0.018)	(0.023)	[0.085]	[0.000]
W N Central Div	0.401	0.376	2.008	4.081
	(0.021)	(0.029)	[0.157]	[0.000]
OK, AR, LA	0.482	0.477	0.325	4.775
	(0.033)	(0.040)	[0.568]	[0.000]
Mountain Div	0.481	0.479	0.032	1.439
	(0.025)	(0.018)	[0.857]	[0.178]
OR, WA, AK, HI	0.459	0.459	0.003	1.286
	(0.037)	(0.035)	[0.958]	[0.258]

Table B11: Return to Business majors in unrelated occupation, by location (uncorrected and corrected)

State	Uncorrected Business Return	Corrected Business Return	χ^2 Test for Difference	Wald Test for Correction Terms
California	0.137	0.140	0.121	1.937
	(0.036)	(0.041)	[0.728]	[0.051]
Texas	0.161	0.155	1.300	5.802
	(0.033)	(0.035)	[0.254]	[0.000]
Florida	0.165	0.158	1.811	4.789
	(0.040)	(0.051)	[0.178]	[0.000]
Illinois	0.174	0.156	2.276	2.649
	(0.037)	(0.029)	[0.131]	[0.010]
New York	0.172	0.134	1.891	4.874
	(0.039)	(0.045)	[0.169]	[0.000]
New England	0.178	0.160	1.029	5.033
	(0.038)	(0.039)	[0.310]	[0.000]
New Jersey & Penn.	0.259	0.241	3.599	5.400
	(0.029)	(0.038)	[0.058]	[0.000]
WV, VA, DC, MD, DE	0.176	0.175	0.048	6.509
	(0.032)	(0.047)	[0.826]	[0.000]
NC, SC, GA	0.151	0.143	1.719	3.193
	(0.029)	(0.033)	[0.190]	[0.002]
E S Central Div	0.151	0.138	2.484	2.730
	(0.031)	(0.031)	[0.115]	[0.010]
OH, IN, MI, WI	0.198	0.179	1.265	2.484
	(0.022)	(0.030)	[0.261]	[0.012]
W N Central Div	0.186	0.177	0.265	4.240
	(0.024)	(0.029)	[0.607]	[0.000]
OK, AR, LA	0.104	0.089	3.673	3.381
	(0.040)	(0.046)	[0.055]	[0.002]
Mountain Div	0.163	0.159	0.492	2.055
	(0.030)	(0.041)	[0.483]	[0.041]
OR, WA, AK, HI	0.177	0.180	0.129	1.970
	(0.041)	(0.039)	[0.720]	[0.051]

Table B12: Return to Business majors in related occupation, by location (uncorrected and corrected)

State	Uncorrected Business Return	Corrected Business Return	χ^2 Test for Difference	Wald Test for Correction Terms
California	0.460	0.467	0.286	4.905
	(0.032)	(0.037)	[0.593]	[0.000]
Texas	0.380	0.385	0.663	6.917
	(0.023)	(0.034)	[0.415]	[0.000]
Florida	0.500	0.500	0.000	3.366
	(0.030)	(0.037)	[0.988]	[0.002]
Illinois	0.435	0.404	3.825	5.491
	(0.033)	(0.036)	[0.050]	[0.000]
New York	0.467	0.413	3.556	13.871
	(0.049)	(0.065)	[0.059]	[0.000]
New England	0.432	0.410	1.424	9.652
	(0.033)	(0.041)	[0.233]	[0.000]
New Jersey & Penn.	0.411	0.382	2.829	5.590
	(0.023)	(0.030)	[0.093]	[0.000]
WV, VA, DC, MD, DE	0.449	0.451	0.110	5.582
	(0.027)	(0.038)	[0.741]	[0.000]
NC, SC, GA	0.502	0.509	0.917	3.251
	(0.024)	(0.030)	[0.338]	[0.002]
E S Central Div	0.463	0.463	0.007	1.395
	(0.027)	(0.037)	[0.933]	[0.208]
OH, IN, MI, WI	0.420	0.393	1.734	5.115
	(0.018)	(0.029)	[0.188]	[0.000]
W N Central Div	0.407	0.390	0.906	4.081
	(0.021)	(0.030)	[0.341]	[0.000]
OK, AR, LA	0.485	0.485	0.011	4.775
	(0.033)	(0.047)	[0.916]	[0.000]
Mountain Div	0.483	0.481	0.153	1.439
	(0.025)	(0.024)	[0.696]	[0.178]
OR, WA, AK, HI	0.443	0.442	0.032	1.286
	(0.037)	(0.034)	[0.857]	[0.258]

Table B13: Return to Soc. Sci. majors in unrelated occupation, by location (uncorrected and corrected)

State	Uncorrected Soc. Sci. Return	Corrected Soc. Sci. Return	χ^2 Test for Difference	Wald Test for Correction Terms
California	0.030	0.045	0.497	1.937
	(0.037)	(0.045)	[0.481]	[0.051]
Texas	0.066	0.054	1.417	5.802
	(0.036)	(0.043)	[0.234]	[0.000]
Florida	0.052	0.044	2.878	4.789
	(0.042)	(0.059)	[0.090]	[0.000]
Illinois	0.022	0.013	0.137	2.649
	(0.041)	(0.035)	[0.711]	[0.010]
New York	0.107	0.081	0.891	4.874
	(0.041)	(0.045)	[0.345]	[0.000]
New England	0.114	0.101	0.313	5.033
	(0.040)	(0.043)	[0.576]	[0.000]
New Jersey & Penn.	0.137	0.105	2.114	5.400
	(0.032)	(0.039)	[0.146]	[0.000]
WV, VA, DC, MD, DE	0.048	0.056	0.436	6.509
	(0.034)	(0.045)	[0.509]	[0.000]
NC, SC, GA	0.101	0.094	0.787	3.193
	(0.032)	(0.040)	[0.375]	[0.002]
E S Central Div	-0.006	-0.014	0.807	2.730
	(0.035)	(0.035)	[0.369]	[0.010]
OH, IN, MI, WI	0.030	0.021	0.762	2.484
	(0.024)	(0.030)	[0.383]	[0.012]
W N Central Div	0.017	-0.018	4.604	4.240
	(0.027)	(0.030)	[0.032]	[0.000]
OK, AR, LA	-0.086	-0.102	3.138	3.381
	(0.047)	(0.044)	[0.076]	[0.002]
Mountain Div	0.071	0.070	0.012	2.055
	(0.032)	(0.044)	[0.913]	[0.041]
OR, WA, AK, HI	0.006	0.008	0.065	1.970
	(0.042)	(0.037)	[0.798]	[0.051]

Table B14: Return to Soc. Sci. majors in related occupation, by location (uncorrected and corrected)

State	Uncorrected Soc. Sci. Return	Corrected Soc. Sci. Return	χ^2 Test for Difference	Wald Test for Correction Terms
California	0.312	0.331	2.963	4.905
	(0.034)	(0.036)	[0.085]	[0.000]
Texas	0.156	0.153	0.127	6.917
	(0.029)	(0.041)	[0.721]	[0.000]
Florida	0.313	0.312	0.019	3.366
	(0.037)	(0.043)	[0.891]	[0.002]
Illinois	0.244	0.199	2.512	5.491
	(0.040)	(0.055)	[0.113]	[0.000]
New York	0.212	0.181	0.950	13.871
	(0.053)	(0.064)	[0.330]	[0.000]
New England	0.231	0.220	0.263	9.652
	(0.037)	(0.042)	[0.608]	[0.000]
New Jersey & Penn.	0.155	0.146	0.146	5.590
	(0.028)	(0.040)	[0.702]	[0.000]
WV, VA, DC, MD, DE	0.263	0.280	3.683	5.582
	(0.031)	(0.038)	[0.055]	[0.000]
NC, SC, GA	0.282	0.285	0.054	3.251
	(0.028)	(0.030)	[0.816]	[0.002]
E S Central Div	0.187	0.184	0.069	1.395
	(0.034)	(0.052)	[0.792]	[0.208]
OH, IN, MI, WI	0.151	0.143	0.118	5.115
	(0.023)	(0.029)	[0.731]	[0.000]
W N Central Div	0.159	0.149	0.211	4.081
	(0.028)	(0.041)	[0.646]	[0.000]
OK, AR, LA	0.195	0.178	3.045	4.775
	(0.044)	(0.065)	[0.081]	[0.000]
Mountain Div	0.290	0.291	0.017	1.439
	(0.029)	(0.032)	[0.896]	[0.178]
OR, WA, AK, HI	0.243	0.242	0.020	1.286
	(0.041)	(0.036)	[0.888]	[0.258]

Table B15: Return to Adv. Deg. STEM majors in unrelated occupation, by location (uncorrected and corrected)

State	Uncorrected STEM Return	Corrected STEM Return	χ^2 Test for Difference	Wald Test for Correction Terms
California	0.192	0.173	1.287	1.937
	(0.058)	(0.062)	[0.257]	[0.051]
Texas	0.108	0.101	0.315	5.802
	(0.067)	(0.090)	[0.575]	[0.000]
Florida	0.149	0.154	0.142	4.789
	(0.085)	(0.070)	[0.706]	[0.000]
Illinois	0.179	0.150	0.830	2.649
	(0.079)	(0.086)	[0.362]	[0.010]
New York	0.250	0.183	1.577	4.874
	(0.060)	(0.081)	[0.209]	[0.000]
New England	0.161	0.122	1.203	5.033
	(0.065)	(0.091)	[0.273]	[0.000]
New Jersey & Penn.	0.258	0.192	6.158	5.400
	(0.061)	(0.067)	[0.013]	[0.000]
WV, VA, DC, MD, DE	0.279	0.255	1.408	6.509
	(0.055)	(0.076)	[0.235]	[0.000]
NC, SC, GA	0.113	0.102	0.485	3.193
	(0.062)	(0.066)	[0.486]	[0.002]
E S Central Div	0.241	0.229	0.589	2.730
	(0.079)	(0.090)	[0.443]	[0.010]
OH, IN, MI, WI	0.229	0.204	0.892	2.484
	(0.054)	(0.070)	[0.345]	[0.012]
W N Central Div	0.127	0.099	2.386	4.240
	(0.069)	(0.107)	[0.122]	[0.000]
OK, AR, LA	0.233	0.240	0.230	3.381
	(0.116)	(0.152)	[0.631]	[0.002]
Mountain Div	0.378	0.372	0.816	2.055
	(0.069)	(0.080)	[0.366]	[0.041]
OR, WA, AK, HI	0.439	0.421	2.737	1.970
	(0.093)	(0.094)	[0.098]	[0.051]

Table B16: Return to Adv. Deg. STEM majors in related occupation, by location (uncorrected and corrected)

State	Uncorrected STEM Return	Corrected STEM Return	χ^2 Test for Difference	Wald Test for Correction Terms
California	0.253	0.215	4.298	4.905
	(0.041)	(0.059)	[0.038]	[0.000]
Texas	0.098	0.076	4.203	6.917
	(0.046)	(0.057)	[0.040]	[0.000]
Florida	0.288	0.265	3.991	3.366
	(0.061)	(0.083)	[0.046]	[0.002]
Illinois	0.212	0.175	1.369	5.491
	(0.054)	(0.067)	[0.242]	[0.000]
New York	0.135	0.097	0.679	13.871
	(0.045)	(0.060)	[0.410]	[0.000]
New England	0.124	0.080	2.288	9.652
	(0.044)	(0.053)	[0.130]	[0.000]
New Jersey & Penn.	0.329	0.254	6.393	5.590
	(0.044)	(0.059)	[0.011]	[0.000]
WV, VA, DC, MD, DE	0.200	0.180	3.535	5.582
	(0.040)	(0.050)	[0.060]	[0.000]
NC, SC, GA	0.232	0.213	6.744	3.251
	(0.046)	(0.049)	[0.009]	[0.002]
E S Central Div	0.196	0.174	3.755	1.395
	(0.055)	(0.056)	[0.053]	[0.208]
OH, IN, MI, WI	0.202	0.161	2.731	5.115
	(0.036)	(0.056)	[0.098]	[0.000]
W N Central Div	0.200	0.163	4.223	4.081
	(0.046)	(0.076)	[0.040]	[0.000]
OK, AR, LA	0.169	0.130	2.625	4.775
	(0.079)	(0.132)	[0.105]	[0.000]
Mountain Div	0.153	0.153	0.004	1.439
	(0.051)	(0.054)	[0.949]	[0.178]
OR, WA, AK, HI	0.184	0.178	0.589	1.286
	(0.062)	(0.088)	[0.443]	[0.258]

Table B17: Return to Adv. Deg. Business majors in unrelated occupation, by location (uncorrected and corrected)

State	Uncorrected Business Return	Corrected Business Return	χ^2 Test for Difference	Wald Test for Correction Terms
California	0.093	0.076	0.931	1.937
	(0.060)	(0.069)	[0.334]	[0.051]
Texas	0.071	0.071	0.000	5.802
	(0.068)	(0.085)	[0.996]	[0.000]
Florida	0.097	0.099	0.034	4.789
	(0.086)	(0.073)	[0.854]	[0.000]
Illinois	0.169	0.143	0.523	2.649
	(0.080)	(0.096)	[0.470]	[0.010]
New York	0.136	0.074	1.345	4.874
	(0.061)	(0.085)	[0.246]	[0.000]
New England	0.088	0.047	1.117	5.033
	(0.067)	(0.092)	[0.290]	[0.000]
New Jersey & Penn.	0.209	0.138	7.567	5.400
	(0.062)	(0.066)	[0.006]	[0.000]
WV, VA, DC, MD, DE	0.312	0.282	2.017	6.509
	(0.056)	(0.074)	[0.156]	[0.000]
NC, SC, GA	0.132	0.128	0.059	3.193
	(0.063)	(0.066)	[0.808]	[0.002]
E S Central Div	0.178	0.169	0.391	2.730
	(0.081)	(0.087)	[0.532]	[0.010]
OH, IN, MI, WI	0.162	0.134	0.998	2.484
	(0.055)	(0.076)	[0.318]	[0.012]
W N Central Div	0.065	0.036	2.823	4.240
	(0.071)	(0.109)	[0.093]	[0.000]
OK, AR, LA	0.276	0.279	0.041	3.381
	(0.119)	(0.143)	[0.839]	[0.002]
Mountain Div	0.328	0.325	0.171	2.055
	(0.071)	(0.081)	[0.680]	[0.041]
OR, WA, AK, HI	0.373	0.357	1.365	1.970
	(0.096)	(0.100)	[0.243]	[0.051]

Table B18: Return to Adv. Deg. Business majors in related occupation, by location (uncorrected and corrected)

State	Uncorrected Business Return	Corrected Business Return	χ^2 Test for Difference	Wald Test for Correction Terms
California	0.310	0.277	3.719	4.905
	(0.042)	(0.058)	[0.054]	[0.000]
Texas	0.051	0.028	4.804	6.917
	(0.046)	(0.062)	[0.028]	[0.000]
Florida	0.205	0.181	3.449	3.366
	(0.061)	(0.074)	[0.063]	[0.002]
Illinois	0.227	0.190	1.687	5.491
	(0.054)	(0.067)	[0.194]	[0.000]
New York	0.236	0.211	0.330	13.871
	(0.045)	(0.071)	[0.566]	[0.000]
New England	0.168	0.120	2.345	9.652
	(0.045)	(0.053)	[0.126]	[0.000]
New Jersey & Penn.	0.264	0.195	4.765	5.590
	(0.045)	(0.060)	[0.029]	[0.000]
WV, VA, DC, MD, DE	0.193	0.171	3.756	5.582
	(0.041)	(0.047)	[0.053]	[0.000]
NC, SC, GA	0.173	0.155	6.920	3.251
	(0.047)	(0.048)	[0.009]	[0.002]
E S Central Div	0.103	0.079	3.812	1.395
	(0.055)	(0.059)	[0.051]	[0.208]
OH, IN, MI, WI	0.128	0.080	3.151	5.115
	(0.037)	(0.061)	[0.076]	[0.000]
W N Central Div	0.162	0.124	4.027	4.081
	(0.048)	(0.067)	[0.045]	[0.000]
OK, AR, LA	0.006	-0.041	3.532	4.775
	(0.081)	(0.105)	[0.060]	[0.000]
Mountain Div	0.073	0.078	0.459	1.439
	(0.052)	(0.062)	[0.498]	[0.178]
OR, WA, AK, HI	0.128	0.123	0.359	1.286
	(0.064)	(0.081)	[0.549]	[0.258]

Table B19: Return to Adv. Deg. Soc. Sci. majors in unrelated occupation, by location (uncorrected and corrected)

State	Uncorrected Soc. Sci. Return	Corrected Soc. Sci. Return	χ^2 Test for Difference	Wald Test for Correction Terms
California	0.174	0.146	0.511	1.937
	(0.060)	(0.074)	[0.475]	[0.051]
Texas	0.089	0.086	0.049	5.802
	(0.072)	(0.090)	[0.825]	[0.000]
Florida	0.165	0.167	0.016	4.789
	(0.088)	(0.073)	[0.898]	[0.000]
Illinois	0.167	0.142	0.891	2.649
	(0.082)	(0.093)	[0.345]	[0.010]
New York	0.184	0.111	1.782	4.874
	(0.063)	(0.081)	[0.182]	[0.000]
New England	0.110	0.054	2.206	5.033
	(0.068)	(0.089)	[0.137]	[0.000]
New Jersey & Penn.	0.241	0.179	6.539	5.400
	(0.064)	(0.071)	[0.011]	[0.000]
WV, VA, DC, MD, DE	0.330	0.306	1.313	6.509
	(0.056)	(0.073)	[0.252]	[0.000]
NC, SC, GA	0.068	0.057	0.495	3.193
	(0.066)	(0.075)	[0.482]	[0.002]
E S Central Div	0.232	0.220	0.574	2.730
	(0.085)	(0.083)	[0.449]	[0.010]
OH, IN, MI, WI	0.217	0.191	0.890	2.484
	(0.058)	(0.076)	[0.345]	[0.012]
W N Central Div	0.112	0.105	0.093	4.240
	(0.072)	(0.115)	[0.761]	[0.000]
OK, AR, LA	0.239	0.245	0.250	3.381
	(0.125)	(0.161)	[0.617]	[0.002]
Mountain Div	0.363	0.356	0.747	2.055
	(0.072)	(0.070)	[0.387]	[0.041]
OR, WA, AK, HI	0.460	0.443	0.902	1.970
	(0.095)	(0.086)	[0.342]	[0.051]

Table B20: Return to Adv. Deg. Soc. Sci. majors in related occupation, by location (uncorrected and corrected)

State	Uncorrected Soc. Sci. Return	Corrected Soc. Sci. Return	χ^2 Test for Difference	Wald Test for Correction Terms
California	0.195	0.147	3.590	4.905
	(0.045)	(0.058)	[0.058]	[0.000]
Texas	0.016	-0.003	2.452	6.917
	(0.052)	(0.073)	[0.117]	[0.000]
Florida	0.198	0.175	3.659	3.366
	(0.066)	(0.092)	[0.056]	[0.002]
Illinois	0.114	0.082	1.115	5.491
	(0.062)	(0.081)	[0.291]	[0.000]
New York	0.210	0.157	0.873	13.871
	(0.051)	(0.086)	[0.350]	[0.000]
New England	0.109	0.053	3.010	9.652
	(0.050)	(0.056)	[0.083]	[0.000]
New Jersey & Penn.	0.224	0.139	7.469	5.590
	(0.050)	(0.072)	[0.006]	[0.000]
WV, VA, DC, MD, DE	0.211	0.178	6.762	5.582
	(0.043)	(0.056)	[0.009]	[0.000]
NC, SC, GA	0.106	0.088	5.874	3.251
	(0.051)	(0.052)	[0.015]	[0.002]
E S Central Div	0.070	0.044	3.042	1.395
	(0.061)	(0.071)	[0.081]	[0.208]
OH, IN, MI, WI	0.100	0.037	5.494	5.115
	(0.041)	(0.062)	[0.019]	[0.000]
W N Central Div	0.077	0.027	6.133	4.081
	(0.052)	(0.078)	[0.013]	[0.000]
OK, AR, LA	0.009	-0.022	1.338	4.775
	(0.089)	(0.154)	[0.247]	[0.000]
Mountain Div	0.087	0.091	0.262	1.439
	(0.055)	(0.064)	[0.609]	[0.178]
OR, WA, AK, HI	0.047	0.044	0.055	1.286
	(0.068)	(0.083)	[0.815]	[0.258]