Mismatch Unemployment and the Geography of Job Search†

Ioana Marinescu   Roland Rathelot

University of Chicago and NBER   University of Warwick and CEPR

Abstract

Could we significantly reduce U.S. unemployment by helping job seekers move closer to jobs? Using data from the leading employment board CareerBuilder.com, we show that, indeed, workers dislike applying to distant jobs: job seekers are 35% less likely to apply to a job 10 miles away from their ZIP code of residence. However, because job seekers are close enough to vacancies on average, this distaste for distance is fairly inconsequential: our search and matching model predicts that relocating job seekers to minimize unemployment would decrease unemployment by only 5.3%. Geographic mismatch is thus a minor driver of aggregate unemployment.

Keywords: local labor markets, job search, misallocation, mismatch, applications, vacancies, unemployment.

JEL: E24, J21, J61, J62, J64.

†Marinescu: ioanamarinescu@uchicago.edu; Rathelot: r.rathelot@warwick.ac.uk. We would like to thank Pierre Cahuc, Raj Chetty, Bruno Crépon, Xavier D'Haultfoeuille, François Fontaine, Florence Goffette-Nagot, Larry Katz, Philip Kircher, Francis Kramarz, Thomas Le Barbanchon, Nathan Hendren, Etienne Lehmann, Manasa Patnam, Barbara Petrongolo, Julien Prat, Craig Riddell, Thijs van Rens, Etienne Wasmer and participants to seminars at CREST, Cachan/Paris-Sud, GATE (Lyon), Mannheim, Uppsala, Louvain-la-Neuve, Bonn, EIEF (Rome), FGV (Rio), PUC (Rio), Sciences Po, Stockholm, Vienna, Warwick, Aarhus, Copenhagen, Montreal, Pompeu Fabra, UIC, Harvard, Harris School (Chicago), Booth (Chicago), Queen Mary, Helsinki, George Washington University, as well as the SaM Meeting at Edinburgh, the ESEM at Toulouse, the workshop Topics in Labor Economics, the NBER LS Spring Meeting at San Francisco 2015, SOLE/EALE world meetings Montreal 2015 for useful comments. We are grateful to CareerBuilder.com for letting us use their data. Any opinions expressed here are those of the authors and not of any institution.

1 Introduction

In the aftermath of the Great Recession, the recovery of the U.S. labor market has been sluggish. The unemployment rate doubled from about 5% to 10% during the Great Recession. In 2012, three years after the official end of the recession, the unemployment rate remained as high as 8%. One potential explanation for this sluggish recovery is geographic mismatch: job seekers in depressed areas may not be able or willing to relocate to areas with better job prospects. Yet, recent work (Şahin et al., 2014) has shown that the contribution of geographic mismatch to the increase in unemployment during the Great Recession was limited.[1]

Even if geographic mismatch did not contribute much to *increasing* unemployment during the Great Recession, this does not imply that geographic mismatch plays no role in the *level* of unemployment. Determining the level of mismatch unemployment is important because it allows us to predict the effects of policies that aim at bringing job seekers and vacancies closer to each other. If geographic mismatch is high, such policies could have a large impact on aggregate unemployment.

Because existing mismatch indices make restrictive assumptions, they cannot determine the level of geographic mismatch. First, existing mismatch indices assume that job seekers do not search for jobs across geographic boundaries, and this assumption is consequential. Indeed, the level of mismatch is sensitive to the assumption made about the geography of job search: if we

[1]Relatedly, evidence does not support the "house lock" hypothesis during the Great Recession. Indeed, homeowners' lower mobility did not contribute to increasing unemployment (Farber, 2012; Valletta, 2013).

assume that workers search within a large area like the state, measured mismatch is systematically lower than if we assume that workers search within a smaller area like the county (Şahin et al., 2014). Therefore, choosing too small a search area could lead us to overestimate mismatch, while choosing too big a search area could lead us to underestimate mismatch. Second, existing indices assume that job seekers are equally likely to match with any job in their area. Such an assumption ignores within area frictions due to distance (spatial mismatch, as in e.g. Ihlanfeldt and Sjoquist (1998)), and could lead us to underestimate mismatch. Therefore, existing mismatch indices cannot determine the level of geographic mismatch because they do not account for the geography of job search both across and within geographic areas.

In this paper, we exploit a large and rich dataset to determine the level of geographic mismatch in the U.S. We use ZIP code level data on the geography of job search for close to 500,000 job seekers sending more than 5 million applications in 2012. The data is from CareerBuilder.com, arguably the largest job board in the U.S., and is broadly representative of the U.S. labor market.[2] Using this data to document the geography of job search, we find that job seekers are more likely to apply to jobs closer to home: a job seeker is 35% less likely to apply to a vacancy that is 10 miles away than to a vacancy that is in the job seekers' ZIP code of residence. Still, we find that, on average, a job seeker sends 11% of their applications to out-of-state vacancies. Overall, American job seekers' distaste for distant jobs is much smaller than that of British job seekers (Manning and Petrongolo, 2011).

[2]Monster.com is the other leading job board and is comparable in size. Which of CareerBuilder or Monster is larger depends on the exact size metric used.

Then, to determine the level of geographic mismatch, we use a directed search model where workers strategically choose where to send their applications given that vacancies closer to home yield higher utility. The model takes as an input the geographic distribution of job seekers and vacancies, as well as our estimate of job seekers' distaste for distance. It predicts where a job seeker applies based on her distance to jobs and the expected probability of getting an offer given the locations of employers and all other job seekers. Using these ingredients, the model predicts the aggregate number of hires. Leaving the vacancies where they are, we compute the maximum number of hires that can be obtained by reallocating job seekers across ZIP codes. Interestingly, this is equivalent to computing the number of hires if job seekers were equally willing to apply anywhere. Finally, geographic mismatch is measured as the difference between the number of hires with the hires-maximizing geographic distribution of job seekers and the number of hires with the existing geographic distribution of job seekers. We find that 5.3% of hires are lost due to job seekers not being close enough to jobs.

Our results are robust to a number of alternative ways of calculating mismatch. In particular, we show that mismatch only slightly increases when we also allow for heterogeneity by occupation. Furthermore, we show that mismatch is not very sensitive to the estimated distaste for distance. Multiplying by two the distaste for distance parameter increases mismatch to 6.0% (from 5.3%).

Given that geographic mismatch is low, policies that reduce barriers to worker mobility (Fan, 2012), or place-based policies that encourage job creation in local areas (Neumark and Kolko, 2010; Kline and Moretti, 2013; Busso et al., 2013; Neumark and Simpson, 2014) are likely to have a limited impact on *aggregate* unemployment[3] in contemporary U.S.

Our paper makes two key contributions to the literature. On the theoretical side, we develop a new model of geographic mismatch that fully takes into account the geography of job search, in contrast to prior measures of geographic mismatch that assume job seekers only search in their own location (e.g. Lazear and Spletzer, 2012; Şahin et al., 2014; Herz and Van Rens, 2015). Our model can also straightforwardly take into account job search across occupations. While Manning and Petrongolo (2011) used a model that takes into account the geography of job search, they did address mismatch. Our model fully takes into account the geography of job search and thus allows us to pin down the level of geographic mismatch.

Our second key contribution is empirical. We use detailed ZIP code level data on applications in the U.S., while Manning and Petrongolo (2011) indirectly infer distaste for distance in the UK from their model and the location of job seekers and jobs. Our data allows us to take into account applications across geographic units such as MSAs, *and* has sufficiently high geographic resolution (ZIP code level) that we can take into account within MSA frictions as well. Furthermore, our model and data allow us to combine mismatch by geography and by occupation to show that mismatch remains relatively low even when we take into account heterogeneity by 2-digit occupations. Ar-

[3]To the extent that such policies create jobs on net, this would change our conclusion, which only pertains to moving jobs or job seekers while keeping their numbers fixed. Furthermore, such policies can have important distributional impacts, which we do not address here.

guably, prior literature on geographic mismatch did not attempt to develop models that fully take into account the geography of job search or applications across occupations because of a lack of adequate data to estimate such models. We are thus in the privileged position to have the necessary data to estimate a model of geographic (and occupational) mismatch that is significantly more realistic.

Our paper is related to the literature on mismatch (e.g. Lazear and Spletzer, 2012; Şahin et al., 2014; Herz and Van Rens, 2015) and the efficiency of the matching function (e.g. Barlevy, 2011; Veracierto, 2011; Davis et al., 2012; Barnichon and Figura, 2013) during and after the Great Recession. Compared to this literature, we focus on precisely measuring one specific type of mismatch: geographic mismatch.

Since we document how likely job seekers are to apply to jobs far away from home, our paper is related to the literature on geographic mobility in the U.S. This literature typically measures moves across states (Molloy et al., 2011). We complement this work by showing which locations (both within and across states) job seekers consider during their job search. A strand of the literature on geographic mobility shows that people move to places with better economic conditions (Greenwood et al., 1986; Bound and Holzer, 2000; Wozniak, 2010). Our results complement this literature by investigating the macro effect of such mobility: mobility is high enough that there is little geographic mismatch.

Our work is also related to the urban economics literature that investigates the distance between the place of residence and the place of employment, and the spatial mismatch hypothesis (Ihlanfeldt and Sjoquist, 1998). This literature uses unemployment insurance data, matched employer-employee datasets or commuting surveys (Hellerstein et al., 2008; Rupert and Wasmer, 2012; McKenzie, 2013; Guglielminetti et al., 2015). We complement this research with evidence on the job search process.

Finally, the evidence we provide about the geography of job search is relevant to the literature on the impact evaluation of many types of local labor market shocks: shocks to labor demand such as a plant opening/closure, place-based policies, etc., or shocks to labor supply such as immigration, training, job search assistance programs, etc.[4]

The next section presents the data. In the third section, we present our theoretical framework. In the fourth section, we provide results about the geography of job search and the level of geographic mismatch. Section five provides robustness tests and extensions. Section six concludes.

[4]This issue is relevant to measure the impact of immigrants on natives' wages or employment rates (Card, 1990; Altonji and Card, 1991; Friedberg and Hunt, 1995; Borjas et al., 1996, 1997; DiNardo and Card, 2000; Card, 2001, 2005; Borjas, 2003; Ottaviano and Peri, 2006), the impact of local shocks on labor demand and supply (Blanchard and Katz, 1992; Bound and Holzer, 2000; Notowidigdo, 2011; Yagan, 2016), the impact of trade and FDI on labor market outcomes (Autor et al., 2013a,b), the equilibrium effects of active labor market policies (Davidson and Woodbury, 1993; Blundell et al., 2004; Gautier et al., 2012; Crépon et al., 2013; Ferracci et al., 2014), the heterogeneity of the negative duration dependence with local conditions (Kroft et al., 2013), or spatial mismatch (Patacchini and Zenou, 2005; Hellerstein et al., 2008; Boustan and Margo, 2009; Åslund et al., 2010).

2 The Geography of Job Search

2.1 Data

We use proprietary data provided by CareerBuilder.com, the largest U.S. employment website. We merge three data sets extracted from CareerBuilder's database. The first one is a random sample of registered users whose accounts were active between April and June 2012. For each job seeker, we have the residence location at the ZIP code level. In order for our results to be comparable with prior literature on job search, we restrict the data to unemployed users. After dropping those who do not reside in the U.S., who live in Alaska and Puerto Rico, and those whose location is unknown, we end up with a data set of 451,783 users.

The second data set is a sample of vacancies published on the website between April and June 2012, and therefore available to the job seekers to apply to. For each job, we know its location at the ZIP code level. Removing non-consistent observations, duplicates and vacancies not located in the U.S. (or located in Alaska or Puerto Rico), and vacancies without ZIP code information leaves 696,975 observations. 37% of the vacancy sample is lost due to the ZIP code availability restriction. We check whether these vacancies without ZIP code are different in terms of location or occupation compared to the vacancies with a ZIP code. The correlation between the city counts of vacancies with ZIP code and without ZIP code is 0.97. The correlation between the SOC-6 level count of vacancies with and without a ZIP code is 0.91. We conclude that vacancies without a ZIP code have the same distribution across cities and occupations as vacancies with a ZIP code, and thus omitting these vacancies should not bias our results[5]. Finally, the third data set connects the two previous data sets by showing which jobs each job seeker applied to. An application is defined as a click on the "Apply now" button that can be found on the full job listing webpage. On average, job seekers sent around 12.8 applications, and vacancies receive 15.8 applications from job seekers in this sample.

We now address the representativity of the data. Background work (Marinescu and Wolthoff, 2015) was done to compare the industry distribution of job vacancies in CareerBuilder.com with the distribution in Job Openings and Labor Turnover Survey (JOLTS). Compared to the distribution of vacancies across industries in JOLTS, some industries are overrepresented in CareerBuilder data, in particular information technology, finance and insurance, and real estate, rental and leasing. The most underrepresented industries are state and local government, accommodation and food services, other services, and construction. While the vacancies on CareerBuilder are not perfectly representative of the ones in the U.S. economy as a whole, they form a substantial fraction of the market. Indeed, the number of vacancies on CareerBuilder.com represented 35% of the total number of vacancies in the U.S. in January 2011 as counted in JOLTS.

In terms of occupation (2-digit SOC codes), the distribution of unemployed job seekers' occupations in CareerBuilder data is very similar to the

[5]In a robustness test (footnote 26), we include these vacancies in our calculation of mismatch at the MSA and commuting-zone levels, and find that doing doing so yields almost the same level of mismatch as using the full sample.

CPS (correlation of 0.71 between the shares of job seekers in each occupation in the two datasets), and the distribution of vacancies' occupations in the CareerBuilder data is essentially identical to the distribution of vacancies in all online jobs (correlation of 0.95 with Help Wanted Online data).

Since the geographic aspect is very important for the purpose of this paper, we verified that the location of vacancies and job seekers in this data is representative of the location of vacancies and job seekers in the U.S. in general. Across U.S. regions, vacancies in our dataset are distributed very similarly to vacancies in the nationally representative Job Openings and Labor Turnover Survey (JOLTS) in April-June 2012 (96% correlation between the shares of vacancies in each region in the two datasets). Across U.S. states, job seekers in this data are also distributed very similarly to the unemployed in the Current Population Survey in April-June 2012, with a correlation of 88%.

In our data, job seekers send 11% of their applications out of state. Of these, some will commute to the other state, and some will move. Using the American Community Survey (ACS) from 2006 to 2010, we find that 4% of employed people commute across state lines for work. Using the 2008 SIPP panel covering years 2008-2013, we find that 5.1% of unemployed people who become employed move across states in the six months before and after the event (this number accounts for slight differences in the composition of the SIPP sample in age and education compared to the CareeBuilder sample). These figures added up together are not far from the 11% of cross-state applications, which suggests that the vast majority of applications can be considered as "serious".

We also compare the destinations of Americans who move across states in the ACS in 2012 with the destinations of out of state applications in our data. We find a very high correlation between the destinations of moves and applications (matrices with the share of moves from each state to each other state), at 0.82. We perform a similar exercise at the county level, comparing the destinations of within-state cross-county applications with within-state cross-county commuting destinations observed in the ACS: we find a high correlation of 0.78.

In conclusion, our data is broadly representative of the U.S. distribution of vacancies and job seekers, and the distribution of applications across geographic units is consistent with the moving and commuting behavior of Americans.

2.2 Estimating the distaste for distance

To understand the geography of job search, we must understand how important distance is in job seekers' application behavior. We first use a descriptive approach and show, for each commonly used geographic unit, the share of applications that are sent to jobs within this unit on average across job seekers (Figure 1).[6] The average share of within state applications is 89%. At the other extreme, the average share of applications within ZIP code is only 4%. Overall, this descriptive approach suggests that job seekers are willing to apply away from their ZIP code but that this willingness declines with distance.

To get a more systematic picture of the impact of geographic distance on job seekers' application behavior, we use a Poisson regression to estimate

[6] All figures have been made with ggplot2 in R (Wickham, 2009).

the probability p_{ij} that a job seeker in ZIP code i applies to a vacancy in ZIP code j as a function of distance between i and j. p_{ij} pins down job seekers' distaste for distance, which will be used to calculate the degree of geographic mismatch.

We model the number of applications from job seekers in ZIP i to vacancies in ZIP j as a Poisson[7] with parameter μ_{ij}:

$$\mu_{ij} = U_i V_j \exp[\alpha_i + \lambda_j + s(d_{ij})] \tag{1}$$

where U_i and V_j are the number of job seekers in i and vacancies in j, α_i and λ_j are fixed effects[8] for job seekers' and vacancies' ZIP codes respectively, and $s(.)$ is a spline function whose parameters are estimated. We use a piecewise-linear spline function, defined by its slopes. With n nodes $\{\bar{d}_i\}_{i=1\ldots n}$, the spline is parametrized by $n+1$ parameters $\{\gamma_i\}_{i=1\ldots(n+1)}$. It is defined so that the derivative of the spline with respect to distance is $s'(d) = \gamma_1$ when distance is below the first node, i.e. when $d < \bar{d}_1$; then $s'(d) = \sum_{i=1}^{j} \gamma_i$ when $d \in (\bar{d}_{j-1}, \bar{d}_j)$ and $j = 2 \ldots n$; $s'(d) = \sum_{i=1}^{n+1} \gamma_i$ when $d > \bar{d}_n$. In words, for $d < \bar{d}_1$, a 1 mile increase in distance multiplies the probability of application by $\exp(\gamma_1)$. This implies that the probability of application changes by approximately $\gamma_1\%$ for a one mile increase in distance.

We chose 10 nodes $\{\bar{d}_i\}_{i=1\ldots10}$ for the spline that parametrizes workers'

[7] The data on applications is collapsed by job seeker ZIP code and vacancy ZIP code to obtain the total count of applications from i to j.

[8] We estimate conditional fixed-effect models, to deal with the incidental parameter problem (Hausman et al., 1984). For the model with two-way fixed effects, we follow the estimation procedure proposed by Guimarães and Portugal (2010) and are only able to perform the estimation on a 10% random subsample, given the computational burden.

willingness to apply as a function of distance: at 10, 20, 30, 50, 75, 100, 200, 500, 1000 and 2000 miles.[9]

The estimated spline function that captures how far away job seekers apply is displayed graphically in Figure 2, based on the regression coefficients $\{\gamma_i\}_{i=1...11}$ in Table 1. Overall, applications clearly decrease with distance. One potential concern is that job seekers in different locations may send different numbers of applications. Similarly, vacancies in some locations may be more attractive to all job seekers, which could bias our estimates of the distaste for distance. Reassuringly, the estimate of the spline is not sensitive to the presence of job seeker ZIP code and vacancy ZIP code fixed effects (Figure 2).

Substantively, job seekers are 38% less likely to apply to a vacancy 10 miles away than to one in their ZIP code of residence (estimates without fixed effects in the first column of Table 1). At larger distances, the distaste for distance is much smaller: job seekers are 9% less likely to apply to a vacancy 110 miles away from their ZIP code of residence than to a vacancy 100 miles away. If we take estimates with one-way fixed effects for the job seekers ZIP codes (the estimates we use to calculate mismatch below), we find very similar effects, with a 35% and a 9% decline in applications at 10 and 110 miles respectively.

Is there any systematic difference in the distaste for distance by education or job type? More educated workers are less likely to apply far away from

[9] Allowing for a flexible function of distance at smaller distances is important to accurately identify job seekers' distaste for distance. Indeed, we have also experimented with a linear specification in distance and found that it does a worse job than the spline in explaining the data (Pseudo $R^2 = 0.53$ vs. 0.72 for the spline specification). The linear specification strongly overestimates job seekers' willingness to apply at short distances away from their ZIP code (under 75 miles) compared to the spline specification.

home for short distances (below 30 miles) but more likely to apply far away for long distances (Figure 3). The result for long distances is consistent with the higher mobility of college educated workers across states (Wozniak, 2010). In Figure 4, we compare the distaste for distance for the most common 8-digit SOC among job seekers (customer service representatives) and the most common 8-digit SOC among vacancies (registered nurse). Customer service representatives, a relatively low skill occupation, exhibit a higher distaste for distance than the overall sample. On the other hand, registered nurses have a higher willingness to apply far away from their ZIP code of residence than the overall sample.

Overall, we find that job seekers are less likely to apply to vacancies further away from their ZIP code of residence, and these results are robust to controls for job seeker and vacancy ZIP code fixed effects. What is yet to be determined is whether job seekers' preference for jobs close to home is high enough to generate substantial geographic mismatch. This is the topic of the next sections.

3 Mismatch unemployment with distinct labor markets

Geographic mismatch occurs when there are too many job seekers (relative to jobs) in some places and too few in other places. Therefore, a greater geographic dispersion in labor market tightness (vacancies/unemployment) implies that there is more geographic mismatch. But how can we quantify the impact of a given level of dispersion in tightness on aggregate unemployment? To pin down this impact, we need to make assumptions about how the geographic distribution of job seekers and vacancies affects hires.

Assume that the location of vacancies is exogenous and fixed. Define mismatch as the percent shortfall in hires resulting from the misallocation of job seekers, i.e. 1-(Total number of hires given observed geographic allocation of job seekers)/(Maximum number of hires across all allocations of job seekers). To calculate the number of hires, we need a matching function, i.e. a mapping from the geographic distributions of U and V to the total number of hires (matches).

The standard approach (Nickell, 1982; Jackman and Roper, 1987) assumes that job seekers are equally likely to match with any job within their home labor market, and will never match with a job outside their home labor market. A Cobb-Douglas matching function is assumed for each market. In this case the mismatch index[10] is:

$$\mathcal{M}_{CD} = 1 - \sum_i \left(\frac{V_i}{\sum_i V_i} \right)^{\gamma} \left(\frac{U_i}{\sum_i U_i} \right)^{1-\gamma} \tag{2}$$

where V_i, U_i are the number of vacancies and unemployed workers in geographic area (labor market) i respectively.

Şahin et al. (2014) show that the Cobb-Douglas mismatch index $\mathcal{M}_{CD}$ represents the percentage shortfall in hires obtained with the actual allocation of job seekers relative to the hires-maximizing allocation of job seekers.In what

[10]See Jackman et al. (1989); Lazear and Spletzer (2012) for a dissimilarity index, which provides a measure of the proportion of the unemployed who are in the "wrong" market. Using the dissimilarity measure yields qualitative results very similar to Figure 5.

follows, we take $\gamma = .5$, as in Şahin et al. (2014).

In order to calculate this mismatch index, one must choose a geographic unit for the location of job seekers, such as the MSA. Working with too broad areas is likely to create a downward bias on the index. If there is only one area (e.g. United States), all applications from job seekers residing in this area are obviously sent within the same area. In this case, the index will obviously be equal to zero but will understate the actual geographic mismatch. Conversely, if we use ZIP codes as the unit of observation, we have the opposite problem. Many applications are directed to vacancies that are not located in the area where the job seeker resides, and we run the risk of overestimating geographic mismatch. As demonstrated by Şahin et al. (2014), choosing a larger area to define the location of job seekers mechanically yields lower mismatch according to the Cobb-Douglas index $\mathcal{M}_{CD}$.

The standard approach assumes that job seekers are as likely to apply to any job within a geographic unit, regardless of how far jobs may be from job seekers' homes. When choosing small units such as ZIP codes, this assumption seems reasonable. But for larger units such as MSAs, this may no longer be the case and job seekers may greatly prefer those jobs within the MSA that are closer to home. Therefore, choosing larger search areas will tend to make us underestimate the amount of friction within each geographic unit, and this is a further reason why $\mathcal{M}_{CD}$ is sensitive to the choice of a geographic unit.

Figure 5 shows how mismatch $\mathcal{M}_{CD}$ varies with the size of the geographic area where job seekers are assumed to look for jobs. When job seekers' search area is defined as the state, 1.6% of hires are lost due to the misallocation of job seekers. If we define the search area as the MSA or the commuting zone (CZ), mismatch is about 2.5%. When search areas are counties, this figure doubles, to 4.9%. At the ZIP code level, the fraction of hires lost due to misallocation of job seekers is a very large 22.9%. When using the Cobb-Douglas mismatch index $\mathcal{M}_{CD}$, the magnitude of geographic mismatch thus strongly depends on the size of the geographic area where job seekers are assumed to look for jobs, with smaller areas yielding larger mismatch values.

4 Mismatch unemployment with interconnected labor markets

4.1 A search and matching model with interconnected markets

Our approach to mismatch seeks to overcome the limitations of the standard approach, which assumes that job seekers only apply to jobs in their own labor market and are equally likely to apply to any job within their labor market. We modify the standard mismatch index in two ways. First, we allow job seekers to apply to jobs in all locations (ZIP codes). Second, in order to be able to model applications across locations, we replace the Cobb-Douglas matching function with a standard urn-ball matching function.

Before developing the details of the model, it is worth noting that the model applies to any setting with heterogeneous labor markets cells. For example, while our baseline empirical application focuses on heterogeneity by geography, we also apply the model to a case where jobs and job seekers differ by geography *and* occupation.

Our objective is to obtain an expression for the total number of matches as a function of the number of job seekers and vacancies in each location, and structural parameters. We use a directed search model where workers choose where to send their applications based on the *location* of the vacancies. Vacancies closer to job seekers' home yield higher utility. Our theoretical model is similar to the one by Manning and Petrongolo (2011), where agents must choose a set of places to apply to, and borrows elements from Albrecht et al. (2006), and Galenianos and Kircher (2009).

Each firm has one vacancy. The location of vacancies is exogenous and fixed. While this assumption may seem restrictive, geographic mismatch is going to be even smaller if both vacancies and workers are allowed to relocate in order to improve their chances of matching. All workers and all firms are identical, risk neutral, and they produce one unit of output when matched and zero otherwise. The utility of an employed worker is defined below, and an unmatched worker has a utility of zero. Workers observe all vacancies. Workers and vacancies are spread across S locations. $i(u)$ and $j(v)$ denote the geographic units where unemployed worker u and vacancy v are respectively located. Each location k has V_k vacancies and U_k unemployed workers. A worker's strategy is a set of $\bar{a}$ vacancies that s/he applies to. The timing of the game is the following.

1. Job seekers apply to vacancies: each job seeker sends $\bar{a}$ applications.
2. Firms gather the applications they receive: each application has a prob-

ability q to be valid in the sense that the applicant will produce positive output if hired. q is a scale parameter: it helps us calibrate the model by capturing the fact that the matching rate in the labor market is lower than what can be predicted on the basis of the number of applications that firms receive.

3. Firms can only make one offer. If a vacancy has more than one valid application, the firm randomly picks the job seeker to whom it makes an offer.

4. Offers are sent to job seekers.

5. Job seekers can only accept one job offer. If a job seeker has received more than one offer, he accepts the offer that generates the highest utility.

6. Matches are realized. If a firm's chosen applicant rejects the job offer, the firm remains unmatched.

The application of a worker u to a vacancy v provides the worker utility $w_{uv} = f(d_{i(u)j(v)})\varepsilon_{uv}$, the product of a deterministic decreasing function f of the geographic distance $d_{i(u)j(v)}$ between the job seeker and the vacancy, and an idiosyncratic term ε_{uv} that is job-worker pair specific.

ε is assumed to be uncorrelated across job seekers and vacancies. This idiosyncratic term ε explains why job seekers do not only apply within their own ZIP code: some very desirable vacancies (high ε) are located in other ZIP codes and job seekers have to trade off distance with ε. Second, ε allows for workers in a given location i to have different preferences over vacancy locations j. Finally, ε allows for unobserved job heterogeneity within a location from the point of view of each specific job seeker.

We assume that the probability π_{uv} that a worker u gets an offer for vacancy v conditional on applying only depends on the location of the vacancy: $\pi_{uv} = \pi_{j(v)}$. This assumption, which is crucial for the tractability of the model, is also present in Manning and Petrongolo (2011).

We now discuss job seekers' optimal strategy. If vacancies closer from a worker's residence do not have a systematically lower probability of yielding an offer, a worker's optimal strategy is to apply to the $\bar{a}$ vacancies with the highest expected utility.[11] This assumption seems like a reasonable approximation: distance to workers' residence and the probability of getting an offer from a vacancy cannot be systematically negatively correlated because workers are geographically dispersed.[12]

Given job seekers' optimal strategy, we derive p_{ij} the probability for a job seeker in i to apply to a vacancy in j. A job seeker u applies to the $\bar{a}$ vacancies with the highest expected utilities $\pi_{j(v)} f(d_{i(u),j(v)})\varepsilon_{uv}$. Assuming that ε has a Pareto distribution[13] of parameter α, p_{ij} is proportional to $\pi_j^\alpha f^\alpha(d_{ij})$. Given that the total number of applications per job seeker is equal to $\bar{a}$, and denoting $g(d_{ij}) = f^\alpha(d_{ij})$, we obtain:

$$p_{ij} = \bar{a}\frac{\pi_j^\alpha g(d_{ij})}{\sum_\ell \pi_\ell^\alpha g(d_{i\ell})V_\ell}, \forall i, j \tag{3}$$

The probability of applying p_{ij} increases in the probability of getting an offer

[11] Assuming that a worker can receive at most one offer, as in Manning and Petrongolo (2011), leads to the same optimal strategy.

[12] For a more in depth discussion of these assumptions, see Appendix A.

[13] This assumption is also present in Manning and Petrongolo (2011).

from a vacancy in j, π_j, and decreases with distance d_{ij} between the job seeker and the vacancy, according to the distaste for distance function g.

The probability that job seekers match depends on the probability of getting an offer from vacancies in each j where they applied, π_j. To derive π_j, we first need to determine how many valid applications a vacancy receives. The total number of applications received by a vacancy located in j from job seekers located in i is distributed as a Poisson$(p_{ij}U_i)$. Summing applications coming from all origins and keeping only the valid ones (probability q), the distribution of the number of valid applications received by a vacancy in j is a Poisson(qr_j), where $r_j = \sum_k p_{kj}U_k$ is the expected number of applications received by a vacancy in j.

From the point of view of job seekers, the probability π_j that an application generates an offer is the probability that their application is valid (q), and that it is picked out by the firm among all other valid applications that the vacancy has received.[14]

$$\pi_j = q\mathcal{R}\left(qr_j\right) \tag{4}$$

where $\mathcal{R}(x) = [1 - \exp(-x)]/x$. Combining equations (3) and (4) and eliminating p and q, we obtain:

$$\pi_j = q\mathcal{R}\left(\pi_j^{\alpha} q\bar{a} \sum_k \frac{g(d_{kj})U_k}{\sum_\ell \pi_\ell^{\alpha} g(d_{k\ell})V_\ell}\right) \tag{5}$$

The total number of matches can be expressed as the number of job seekers multiplied by the probability that each job seeker forms a match. The

[14] π_j is equal to q multiplied by the expectation of $1/(X_j + 1)$, where X_j is the expected number of valid applications made by other job seekers to the job, with $X_j \sim$ Poisson$\left(qr_j\right)$.

probability that a job seeker matches depends on the number of offers received by a job seeker in i from vacancies in each location j, which is distributed as Poisson $(\pi_j p_{ij} V_j)$. The total number of offers received by this job seeker in i from all locations is thus distributed as $Y_i \sim$ Poisson $(\sum_\ell \pi_\ell p_{i\ell} V_\ell)$. A job seeker in k will match if and only if he receives at least one offer, which is one minus the probability of getting zero offers, i.e. $1 - \exp\left(-\sum_\ell p_{k\ell}\pi_\ell V_\ell\right)$. Using equation (3) to substitute $p_{k\ell}$ by its expression, the total number of matches M is:

$$M = \sum_k U_k \left[1 - \exp\left(-\bar{a}\frac{\sum_\ell \pi_\ell^{1+\alpha} g(d_{k\ell})V_\ell}{\sum_\ell \pi_\ell^{\alpha} g(d_{k\ell})V_\ell}\right)\right] \tag{6}$$

In a nutshell, the total number of matches M is equal to the sum of job seekers weighted by the probability that each job seeker gets at least one offer. In turn, the probability of getting at least one offer (in brackets) depends on the number of vacancies weighted by the probability that a vacancy yields an offer (π_ℓ) and a decreasing function of the distance from the job seeker ($g(d_{k\ell})$).

How can we determine the total number of hires M given the parameters $\bar{a}$, α, q, $g(.)$, as well as vectors U and V? Once π is known, it is straightforward to find the total number of hires using equation (6). However, π is difficult to pin down because π is a non-linear function of itself: equation (5) defines a system of S equations where the S π_j are the unknowns, and $\bar{a}$, α, q, $g(.)$, U and V are the parameters. We do not have a proof for the existence or unicity of a solution vector π. However, we find numerically that the expression for π_j in equation (5) defines a contraction mapping that reaches an equilibrium very fast, for a large range of parameters. Trying several starting points always leads to the same solution, which argues in favor of a unique equilibrium.[15]

4.2 Estimation of the structural parameters

We now turn to the issue of estimating the structural parameters of the model: g, the distaste for distance, α, the Pareto parameter for the match-specific utility component ε, and q, the probability of a valid application. We start with describing the parameters we set, and then turn to the estimation of the structural parameters.

First, we set $\bar{a}$ as the average number of applications by job seekers observed in the data.[16] Second, because the number of matches in the model depends on labor market tightness, we must make sure that labor market tightness in our data is representative of the U.S. economy. To do so, we apply a proportionality factor to our vacancies so that the aggregate labor market tightness in our data is the same as in the U.S. economy[17].

Turning to the estimation of the structural parameters, we need to deter-

[15]We can analytically derive π and a closed for form the mismatch index if we assume that $\alpha = 0$, i.e. job seekers do not take into account other job seekers' applications when deciding where to apply for jobs (Appendix D). This closed form mismatch index yields results that are very similar to our preferred mismatch index and can straightforwardly be used to compute mismatch with other datasets. Another way to get a closed form solution is provided by Manning and Petrongolo (2011), who can prove the existence and unicity of the solution in a similar case by assuming that q is small enough that job offers made by employers are always accepted.

[16]As the total number of hires depends on the product $q\bar{a}$ and q is estimated, the mismatch index is actually not sensitive to the value chosen for $\bar{a}$.

[17]For each month of April to June 2012, we compute the monthly tightness by dividing the total number of vacancies (from JOLTS) by the total number of unemployed job seekers (as reported by the Bureau of Labor Statistics based on the Current Population Survey) and take the average as our measure of national labor market tightness. Keeping the geographic distribution fixed, we then inflate the number of vacancies in our data such that the global tightness is equal to the national labor market tightness.

mine the values of g, α and q given $\bar{a}$ and vectors U and V. In order to estimate the distaste for distance g, we choose the parametrization $g = \exp(\eta s(d))$, where $s(.)$ is the spline already estimated in the reduced-form equation (1), and η is a scalar to be estimated now. Under this parametrization of g, the estimation of g and α amounts to the estimation of two parameters, η and α.

We now explain how the estimation proceeds in order to determine η, α and q, given U, V and $\bar{a}$. For each value of (α, η), q is set so that the average job finding rate predicted by the model matches the national job finding rate computed using the CPS.[18] We estimate α and η by maximum likelihood. For a given value of (α, η), we can use our model to compute p_{ij}, the probability that an individual in i applies to a job in j. p_{ij} is directly related to an observed quantity, the number of applications from i to j, A_{ij}. According to our model, A_{ij} is drawn from a Poisson distribution of parameter $\lambda_{ij} = U_i V_j p_{ij}$. We need to find the values of α and η such that $\lambda_{ij}(\alpha, \eta)$ is the most likely parameter of the Poisson underlying A_{ij}. Formally, we find α and η by maximizing the quasi-log-likelihood $\mathcal{L}(\alpha, \eta) = \sum_{i,j} A_{ij} \log \lambda_{ij}(\alpha, \eta) - \lambda_{ij}(\alpha, \eta)$.[19]

We now give an intuition for the identification of η and α. We start with the parameter of the distaste for distance η. Consider two distinct ZIP codes 1 and 2. Intuitively, the further away 1 and 2 are from each other, the fewer across ZIP applications there will be relative to within ZIP applications; so

[18] The national job finding rate is computed with the CPS as the number of unemployment to employment transitions in a given month divided by the number of unemployed workers in the previous month. We then compute a target number of hires $\hat{M}$, equal to the national job finding rate times the number of job seekers in our sample. q is estimated as the quantity minimizing the squared difference between the number of hires predicted by the model M and the target $\hat{M}$.

[19] As can be seen in appendix Figure 9, the log likelihood has a local maximum in the neighborhood of the optimal values of η and α.

this comparison of across vs. within ZIP applications allows us to track down the distaste for distance. Based on our model, the number of applications is $\lambda_{ij} = U_i V_j p_{ij}$, so this comparison can be written as:

$$\frac{\lambda_{12}\lambda_{21}}{\lambda_{11}\lambda_{22}} = \frac{p_{12}p_{21}}{p_{11}p_{22}} = \frac{g(d_{12})g(d_{21})}{g(d_{11})g(d_{22})} = \exp(2\eta s(d_{12}))$$

Note that the right-hand side does not depend on the parameter α and thus we conjecture that η can be identified separately from α. The transformation on λ is equivalent to introducing fixed effects for origin and destination ZIP codes in a reduced-form context. If $s(.)$ was estimated using a two-way fixed-effect on the full sample, we would expect to have $\eta = 1$. Using the $s(.)$ estimated with fixed effects on the job seeker ZIP code[20], the maximum likelihood estimate of η is 1.0020, so the reduced-form estimate of the distaste for distance is essentially unbiased. Thus, we consider in what follows that $g(d) = \exp(s(d))$.

Now, given η, what is the source of identification for α? We can show that the expected number of applications received by a vacancy in j is:

$$r_j = q\bar{a}\mathcal{R}^{\alpha}(qr_j)\sum_k U_k g(d_{kj})$$

In this expression, $q\bar{a}\sum_k U_k g(d_{kj})$ is the number of valid applications that a given vacancy in j would receive if applicants did not factor in the probability of getting an offer π_j. The term $\mathcal{R}^{\alpha}(qr_j)$ plays the role of a moderating force:

[20] The one-way job seeker ZIP code fixed effect gives essentially the same estimate for the distaste for distance as the two-way fixed effects. However, the two-way fixed effect estimate is only based on 10% of the sample, which is why we use the one-way fixed effect. Either way, the estimates are so close that mismatch is not sensitive to using the one or the other.

if a place attracts more applications r_j, $\mathcal{R}^{\alpha}(qr_j)$ decreases and moderates the increase in r_j. The higher α, the more this force is at play. A higher α lowers the dispersion in r_j, so α can be identified by matching the r_j predicted by the model to the number of applications observed in the data.[21] Table 2 lists the parameters of the model and the values of the estimated parameters. Note that the probability of a valid application q is quite low because the job finding rate in the CPS is only 18.2%, despite the fact that job seekers are sending multiple applications. In order to match the CPS job finding rate, we must assume that most applicants are not qualified for the job.

4.3 Mismatch index

In this section, we assume that a social planner can move job seekers at no cost to maximize the number of hires. Just as in the standard approach (section 3 above), we define mismatch as the difference between the maximum number of hires obtained by the planner (M^*) and the number of hires obtained with the actual allocation of job seekers (M): mismatch is then $1 - M/M^*$. Note that this concept of mismatch has no implications for social welfare defined in a general way. Mismatch only represents a deviation from the objective of maximizing the number of matches and thus minimizing aggregate unemployment.

We want to find the allocation of job seekers that maximizes M. If distance to jobs did not matter to job seekers, we would have a single integrated

[21] Appendix D shows that mismatch stays of the same order of magnitude for a plausible range of values of α.

labor market. The social planner would not have to move job seekers: matches M would be maximized, regardless of the location of job seekers. We thus use the case of no distaste for distance to infer how the social planner can maximize hires.

In appendix B, we show that, if job seekers have no distaste for distance, the probability of getting an offer π is equalized across locations. This makes sense because, with no distaste for distance, any job is as good as any other (up to ε, which is i.i.d.). Based on this observation, we conjecture that, to maximize hires, the social planner should reallocate job seekers in order to equalize π across locations.

If π is equal across locations j, the average number of applications r_j received by a vacancy does not depend on its location, as there is a bijective relationship between π_j and r_j. Thus, r_j will be equal to the total number of applications divided by the number of vacancies: $\sum_k p_{kj} U_k = \bar{a}\bar{U}/\bar{V}$, where $\bar{U}$ is the total number of unemployed workers in the economy and $\bar{V}$ is the total number of vacancies. Thus, we can rewrite π as:

$$\pi = q\mathcal{R}\left(q\bar{a}\frac{\bar{U}}{\bar{V}}\right)$$

and the total number of matches is:

$$M^* = \bar{U}\left[1 - \exp\left(-q\bar{a}\mathcal{R}\left(q\bar{a}\frac{\bar{U}}{\bar{V}}\right)\right)\right] \tag{7}$$

Interestingly, the number of matches obtained with the allocation that equalizes π across ZIPs is identical to the one obtained with any allocation of job seekers in the case where there is no distaste for distance (Appendix B), which supports our initial conjecture that reallocating job seekers to equalize π maximizes matches.[22]

Our interconnected-markets mismatch index is then defined as one minus the ratio between the number of matches with the actual allocation of job seekers and the maximum number of matches:

$$\mathcal{M}_i = 1 - \sum_k \frac{U_k}{\bar{U}} \frac{1 - \exp\left(-\bar{a}\frac{\sum_\ell \pi_\ell^{1+\alpha} g(d_{k\ell})V_\ell}{\sum_\ell \pi_\ell^{\alpha} g(d_{k\ell})V_\ell}\right)}{1 - \exp\left(-q\bar{a}\mathcal{R}\left(q\bar{a}\bar{U}/\bar{V}\right)\right)} \tag{8}$$

In contrast with our approach, most of the existing literature makes the simplifying assumption that markets are distinct, that is: (i) job seekers can only apply to vacancies within their own unit, (ii) job seekers are equally likely to apply to all vacancies within their own unit. If we assume $g(d_{ii}) = 1$ within the unit and $g(d_{ij}) = 0$ if $i \neq j$, we can build a distinct-market mismatch index equal to:

$$\mathcal{M}_d = 1 - \sum_k \frac{U_k}{\bar{U}} \frac{1 - \exp\left(-q\bar{a}\mathcal{R}\left(q\bar{a}U_k/V_k\right)\right)}{1 - \exp\left(-q\bar{a}\mathcal{R}\left(q\bar{a}\bar{U}/\bar{V}\right)\right)} \tag{9}$$

4.4 Geographic mismatch: results

Mismatch is most accurately captured by our mismatch index $\mathcal{M}_i$ (equation 8) at the ZIP code level, because it allows for detailed geography and for interconnected labor markets. Using our preferred measure of mismatch, we

[22]One can also define the allocation of job seekers that maximizes hires. Denote $\tilde{V}_k = \sum_\ell g(d_{k\ell})V_\ell$, X the matrix of term $[g(d_{ij})\tilde{V}_i]_{ij}$ and b a vector of ones (of dimension the number of ZIP codes). The allocation of job seekers such that π is constant across ZIP codes is equal to $U^* = \frac{\bar{U}}{b'X^{-1}b}X^{-1}b$

find that geographic mismatch is very small: 5.3% of hires are lost due to the misallocation of job seekers (Figure 6, interconnected markets).

Our mismatch estimate of 5.3% implies that we could reduce U.S. aggregate unemployment by approximately 5.3% if we reallocated job seekers to maximize hires.[23] Aggregating the data to the MSA, CZ, county[24] or ZIP code level consistently yields a mismatch close to 5%.[25] It is only if we aggregate the data at the state level that mismatch is markedly smaller. Thus, for a broad range of aggregation levels, mismatch is stable around 5%.[26]

The fact that geographic mismatch is low may seem surprising given the differences in unemployment rates across U.S. states. In our model, dispersion in unemployment across states is due to dispersion in labor market tightness. Crucially, our matching function implies a small impact of dispersion in tightness on aggregate unemployment, a feature that is shared by the Cobb-Douglas mismatch index at the state level. Therefore, high dispersion in unemployment

[23]At the steady state, the unemployment rate is given by $u = \mu_u/(\mu_e + \mu_u)$ with μ_u the entry rate and μ_e the exit rate to/from unemployment. If μ_u is fixed and μ_e increases to μ_e^* when we go to the hires-maximizing allocation, $\mathcal{M} = 1 - \mu_e/\mu_e^*$. The decrease in the steady-state unemployment rate is equal to $(u - u^*)/u = (\mu_e^* - \mu_e)/(\mu_e^* + \mu_u)$. Given that $\mu_e >> \mu_u$, $(u - u^*)/u \simeq \mathcal{M}$.

[24]The level of mismatch at the county level is slightly higher than at the ZIP code level. Indeed, unlike the Cobb-Douglas mismatch index, our mismatch index does not monotonically decline when data is more aggregated. When we aggregate the data at the county level, we place all job seekers and all jobs in the middle of the county. Such an aggregation procedure puts job seekers closer to jobs in their own county but further away from jobs in other counties, leading to a negative net effect on the number of matches.

[25]When the model is estimated at the ZIP or county level, we consider that the internal distance (within the same ZIP or same county) is 0. When the model is estimated at coarser levels (CZ, MSA, state), we follow the trade literature dealing with the estimation of gravity models and introduce the internal distance defined as two-third of the square root of the area of the unit divided by π (Head and Mayer, 2004).

[26]In a robustness test, we recalculated mismatch at the MSA and CZ levels including vacancies for which we have the city but not the ZIP code (i.e. essentially all vacancies). The resulting mismatch do not change much: 5.28% at the MSA level and 5.47% at the CZ level.

rates across states is compatible with low geographic mismatch because such dispersion does not result in large losses in the aggregate number of matches.

A recent literature has attempted to isolate the determinants of workers' location decisions (e.g. Diamond, 2015) and to explore the sources of differences in unemployment rates across states (e.g. Herz and Van Rens, 2015; Amior and Manning, 2015). Our results suggest that these differences in unemployment rates across locations do not matter much for aggregate unemployment: whatever its sources, unemployment dispersion accounts for a very limited amount of aggregate unemployment.

The Cobb-Douglas index $\mathcal{M}_{CD}$ (Figure 5) using county level data yields a level of mismatch that is similar to our preferred measure based on ZIP code data. However, $\mathcal{M}_{CD}$ grossly overestimates mismatch based on ZIP code data, and underestimates mismatch based on CZ or MSA data. These differences arise both because our model uses a different matching function, and because we allow for applications across geographic units.

To understand the independent role of allowing for applications across geographic areas, we recalculate mismatch with our model, but preventing job seekers from applying across geographic areas ($\mathcal{M}_d$, equation 9). The distinct markets mismatch index $\mathcal{M}_d$ (Figure 6) yields similar results to those arising from the Cobb-Douglas mismatch index $\mathcal{M}_{CD}$. Just like $\mathcal{M}_{CD}$, $\mathcal{M}_d$ is sensitive to the size of the area where job seekers are assumed to look for jobs, with larger areas yielding smaller levels of mismatch. This similarity suggests that the discrepancy between $\mathcal{M}_i$ and the Cobb-Douglas index is mostly due to the fact that our index accounts for across markets applications rather than to the functional form of the matching function.

Overall, using a search and matching model that fully takes into account the geography of job search and data at the ZIP code level, we find that eliminating geographic mismatch would reduce U.S. aggregate unemployment by at most 5.3%.

5 Robustness and Extensions

5.1 Geographic and occupational mismatch

Mismatch unemployment can be the result of a different geographic distribution of job seekers and job vacancies, but it can also result from a different distribution of job seekers and job vacancies across occupations. Moreover, the occupation and spatial dimensions may interact to further increase mismatch. In this sub-section, we move beyond purely geographical mismatch, and compute mismatch combining geographic and occupational heterogeneity.

We define a labor market as a location *and* an occupation and calculate mismatch using these two dimensions at the same time. In order to keep computations tractable, we define labor markets as the intersection of SOC-2 occupations and Commuting Zones, obtaining around 10,000 CZ*SOC-2 labor markets. For job seekers, their occupation is defined as the occupation of their last job on their resume. Just as we do not assume that job seekers only apply in their home location, we do not assume that job seekers whose last job was in a given occupation will restrict their applications to the same occupation.

5.1.1 Distaste for geographic and occupational distance

Restricting applications to be within CZ*SOC2 would be a bad approximation to reality since only 26% of applications are within CZ*SOC2. Therefore, we need to define an application function that depends both on the geographic distance between CZs and on the occupational distance between SOC2s.

To estimate distance between two SOC, we use factor analysis, an approach common to the existing literature (Poletaev and Robinson, 2008). For each 8-digit SOC, there is a vector (defined by ONet) of about 200 elements that represents the knowledge, skills and abilities associated with the jobs in this occupation. We perform a factor analysis on these vectors to extract the major dimensions of heterogeneity across occupations. We consider the first two factors: the first one corresponds roughly to the level of intellectual knowledge and abilities required for an occupation (high for executives, for instance), while the second corresponds to physical and technical skills (high for construction workers or electricians, for instance). Then, for each 2-digit SOC, we take the mean of each factor across the 8-digit SOCs that together constitute the 2-digit SOC.

We estimate a model similar to the one described in equation (1), and we add a dummy for applying to an occupation that is different from the one held in the last job, as well as functions of the two factors estimated above. Specifically, the probability for a job seeker in labor market i to apply for a job in labor market j is:

$$\begin{aligned}\mu_{ij} = U_i V_j \exp[\alpha_i + s(d_{ij}) + \alpha_1 1\{SOC2_i \neq SOC2_j\} \\ +\alpha_2 \left[(\phi_i - \phi_j)^2 + (\psi_i - \psi_j)^2\right]^{1/2} + \alpha_3(\phi_i - \phi_j) + \alpha_4(\psi_i - \psi_j)]\end{aligned} \tag{10}$$

where d_{ij} is the geographic distance between the centroids of the CZs corresponding to i and j, α_i is a job seeker location fixed effect[27]. α_1 estimates a discontinuous preference for one's own SOC2, so we expect $\alpha_1 < 0$. α_2 is the coefficient on the distance between two SOC2 using the two factors ϕ and ψ: we expect $\alpha_2 < 0$. α_3 and α_4 capture the fact that, in each skill dimension, it might be easier to apply to jobs that are less skilled than one's own occupation, so we expect $\alpha_3, \alpha_4 > 0$. In appendix E Table 3, we show the results of these estimates, and confirm the predictions about the sign of the coefficients. For example, the estimates imply that job seekers are 2.8 times less likely to apply to a SOC2 different from their own, even after accounting for levels of each SOC2 in the two main factors, as well as the differences in factors between the two SOC2. Clearly, job seekers prefer their own SOC2, but this preference is not overwhelming and hence it is necessary to model across-SOC2 applications.

5.1.2 Geographic and occupational mismatch: results

Plugging the estimates of the distaste for geographic and occupational distance into our mismatch index $\mathcal{M}_{i\cdot}$, we find that 6.9% of hires are lost due to a

[27] We also estimated the model with job location fixed effects, and estimates are very similar as shown in appendix E Table 3.

combination of geographic and occupational mismatch (Figure 7). Thus, the mismatch index corresponding to the CZ*SOC-2 labor markets is higher than the one corresponding to interconnected CZ labor markets (4.25%), or even distinct CZ labor markets (5.5%) (Figure 7).

How does heterogeneity by occupation contribute to mismatch compared to heterogeneity across geography? To shed light on this question, we shut down applications across CZ, across SOC2, or both (Figure 7). If we shut down applications across CZ (distinct CZ) but still allow job seekers to apply across SOC2, mismatch is 8.0%. If we instead shut down applications across SOC2 (distinct SOC2) but allow applications across CZ, mismatch is 14.6%, which is twice as high as the level of geographic mismatch estimated when allowing applications across occupations. Finally, in the case where we shut down applications both across CZs and across SOCs, thus assuming distinct CZ*SOC2 markets, mismatch is 17.2%.[28] We conclude that mismatch is more severely overestimated by not allowing for applications across SOC-2 than by not allowing for applications across CZ.

5.1.3 Mismatch by detailed SOC code and by education

Since in the previous section we used a relatively coarse grouping of occupations, this could underestimate mismatch. Here we compute mismatch for the most common 8-digit SOC among job seekers (customer service representatives) and the most common 8-digit SOC among vacancies (registered nurse).

[28]There is an interaction effect whereby not allowing for geographically interconnected markets yields a larger increase in mismatch if we also do not allow for interconnected occupations: 14.6% to 17.2% vs. 6.9% to 8% in the case of interconnected occupations.

We estimate mismatch with occupation-specific distaste for distance (Figure 4) and leaving all other parameters as in the baseline.

For registered nurses, 5.1% of hires are lost due to mismatch. For customer service representatives, mismatch is much higher at 9.3%. The higher mismatch for customer service representatives is mostly due to their worse distribution across the territory (further away from jobs) rather than to their greater distaste for distance. Indeed, if we assume that customer service representatives have the same distaste for distance as registered nurses, mismatch for customer service representatives is still 8.6%. Note that these mismatch indices are overestimated because we computed them under the assumption that job seekers only apply to jobs in their past 6-digit occupation. The overall conclusion is that mismatch can vary considerably across occupations but it stays relatively small even for occupations that are more prone to mismatch.

We also compute mismatch unemployment by level of education (high school, associates, BA and above), assuming that job seekers only apply to jobs in their own occupation. We find that mismatch unemployment decreases with the level of education consistent with more educated workers being more willing to apply far away from home at long distances (Figure 3). Yet, mismatch is never much higher than 5% (appendix C).

Overall, we find that occupational mismatch is considerably overestimated if we ignore applications across occupations. Eliminating *both* geographical (CZ) and occupational (SOC-2) mismatch would reduce U.S. aggregate unemployment by only 6.9%, and not 17.2% as would be the case if job seekers did not apply across occupations and CZs.

5.2 Mismatch for various distastes for distance

People who use CareerBuilder for job search may have lower distaste for distance. So, would mismatch increase a lot if distaste for distance were greater? We compute mismatch at the ZIP code level and for different distastes for distance. We rely on our baseline estimate of the distaste for distance and parametrize it with ξ: $g = \exp(\xi s(d))$, where $s(.)$ is the spline estimated in the reduced-form equation (1). We let the parameter ξ vary between 0 and 10 in increments of 0.5.

An increase in the distaste for distance starting from our baseline of $\xi = 1$ barely increases geographic mismatch (Figure 8, actual allocation). If we multiply the distaste for distance by two, mismatch increases to 6%, and even if we multiply the distaste for distance by 5, mismatch is still only 7.8%. Thus, even if our data underestimates the distaste for distance compared to a representative sample, this barely affects the level of geographic mismatch.

How high would mismatch be if American job seekers had the same distaste for distance as the British job seekers? American job seekers are eight times[29] more willing to apply to vacancies far away from home than the British job seekers[30] studied by Manning and Petrongolo (2011). When we plug the British distaste for distance in our model, we find that the U.S. mismatch

[29] Job 0 is preferred to job 1 to job 1 iff $g(d_1)\varepsilon_1 > g(d_0)\varepsilon_0$. Because ε_0 and ε_1 are Pareto, the probability to prefer a job at distance d rather than the one at distance 0 is after some algebra $\exp(s(d_1))/2$. In our case, this amounts to $\exp(-0.0471397 * 6.2)/2 = 0.37$; with the estimates found in (Manning and Petrongolo, 2011), $\exp(-0.3 * 10) = 5\%$.

[30] This difference may reflect differences between the U.S. and UK labor markets. It might also be influenced by the methodology used: Manning and Petrongolo (2011) infer the distaste for distance parameter from the estimation of a search-and-matching model but cannot directly observe job seekers' application behavior.

almost doubles, at 10.8%..

Despite the fact that British job seekers have a greater distaste for distance than American job seekers, Manning and Petrongolo (2011) found that place-based policies are rather ineffective. Indeed, job seekers from other areas apply to newly created jobs in target areas, so the positive employment effect for target area residents is muted. Because job seekers in the US apply much further away from their home area than in the UK, the impact of place-based policies in contemporary U.S. is likely to be even more muted.

We have just seen that the level of mismatch is not very sensitive to the distaste for distance, which suggests that job seekers are fairly close to jobs already. On the other hand, if job seekers were allocated uniformly across space (i.e. the number of job seekers in a ZIP code depends on the ZIP code area), greater distaste for distance would dramatically increase geographic mismatch. For most values of the distaste for distance, mismatch is much higher with the uniform than with the actual allocation (Figure 8, compare blue and red bars).[31] With a uniform allocation and the British distaste for distance, mismatch would be as high as 63.7%! These results suggest that, with the actual allocation of job seekers, increasing the distaste for distance has little impact on mismatch because job seekers already live pretty close to vacancies on average.

[31] When the distaste for distance is very low, the uniform allocation of job seekers yields a lower level of geographic mismatch than the actual allocation of job seekers. This is most likely due to the fact that job seekers are overly concentrated close to big job centers, which reduces their job finding probability and makes them miss out on vacancies that are a further afield. Thus, the job finding rate predicted by our model is lower close to business centers (defined as the ZIP code in each state with the highest number of vacancies). For low distaste for distance, the uniform allocation fixes this issue by placing job seekers further away from business centers.

Based on this analysis, geographic mismatch is low because distaste for distance is low enough, and job seekers are already fairly close to vacancies. In a dynamic framework, low distaste for distance can explain why job seekers are relatively well allocated across space: over time, job seekers relocate to follow vacancies so that, at any given point in time, job seekers live close to vacancies on average.

6 Conclusion

In this paper, we have used a novel dataset from CareerBuilder.com to document how far job seekers are willing to apply to jobs and, based on this evidence, we have measured the degree of geographic mismatch. Our measure of geographic mismatch is based on a search and matching model of the labor market in which job seekers strategically choose where to send their applications. Quantitatively, we find that U.S. aggregate unemployment would be reduced by at most 5.3% if job seekers were reallocated so as to maximize hires. Therefore, geographic mismatch is a minor driver of U.S. unemployment.

Low mismatch can be explained by job seekers' high enough willingness to apply far away from home combined with the fact that the typical job seeker does not live very far away from jobs. We also extend our model to measure geographic and occupational mismatch taken together. Adding the occupational dimension (2-digit SOC codes) naturally increases mismatch. Yet, geographic and occupational mismatch remains low (6.9%) as long as job seekers are allowed to apply across occupations.

Overall, we find that geographic mismatch is a minor cause of unemployment at the macro level. Thus, policies that attempt to combat geographic mismatch by reducing barriers to worker mobility or moving job seekers and jobs closer to each other are likely to have a limited effect on aggregate unemployment.

References

Albrecht, J. W., P. A. Gautier, and S. B. Vroman (2006): "Equilibrium Directed Search with Multiple Applications," *Review of Economic Studies*, 73, 869–891.

Altonji, J. G. and D. Card (1991): "The Effects of Immigration on the Labor Market Outcomes of Less-skilled Natives," in *Immigration, Trade and the Labor Market*, NBER Chapters, 201–234.

Amior, M. and A. Manning (2015): "The persistence of local joblessness," Mimeo LSE.

Åslund, O., J. Østh, and Y. Zenou (2010): "How Important Is Access to Jobs? Old Question - Improved Answer," *Journal of Economic Geography*, 10, 389–422.

Autor, D. H., D. Dorn, and G. H. Hanson (2013a): "The Geography of Trade and Technology Shocks in the United States," *American Economic Review*, 103, 220–25.

——— (2013b): "Untangling Trade and Technology: Evidence from Local Labor Markets," NBER Working Paper 18938.

Barlevy, G. (2011): "Evaluating the role of labor market mismatch in rising unemployment," *Economic Perspectives*, 82–96.

Barnichon, R. and A. Figura (2013): "Labor Market Heterogeneity and the Aggregate Matching Function," Mimeo CREi.

Blanchard, O. J. and L. F. Katz (1992): "Regional Evolutions," *Brookings Papers on Economic Activity*, 23, 1–76.

Blundell, R., M. C. Dias, C. Meghir, and J. V. Reenen (2004): "Evaluating the Employment Impact of a Mandatory Job Search Program," *Journal of the European Economic Association*, 2, 569–606.

Borjas, G. J. (2003): "The Labor Demand Curve Is Downward Sloping: Reexamining The Impact Of Immigration On The Labor Market," *The Quarterly Journal of Economics*, 118, 1335–1374.

Borjas, G. J., R. B. Freeman, and L. F. Katz (1996): "Searching for the Effect of Immigration on the Labor Market," *American Economic Review*, 86, 246–51.

——— (1997): "How Much Do Immigration and Trade Affect Labor Market Outcomes?" *Brookings Papers on Economic Activity*, 28, 1–90.

Bound, J. and H. J. Holzer (2000): "Demand Shifts, Population Adjustments, and Labor Market Outcomes during the 1980s," *Journal of Labor Economics*, 18, 20–54.

Boustan, L. P. and R. A. Margo (2009): "Race, Segregation, and Postal Employment: New Evidence on Spatial Mismatch," *Journal of Urban Economics*, 65, 11–23.

Busso, M., J. Gregory, and P. Kline (2013): "Assessing the Incidence and Efficiency of a Prominent Place Based Policy," *American Economic Review*, 103, 897–947.

Card, D. (1990): "The Impact of the Mariel Boatlift on the Miami Labor Market," *Industrial and Labor Relations Review*, 43, 245–257.

——— (2001): "Immigrant Inflows, Native Outflows, and the Local Labor Market Impacts of Higher Immigration," *Journal of Labor Economics*, 19, 22–64.

——— (2005): "Is the New Immigration Really so Bad?" *Economic Journal*, 115, F300–F323.

Chade, H. and L. Smith (2006): "Simultaneous Search," *Econometrica*, 74, 1293–1307.

Crépon, B., E. Duflo, M. Gurgand, R. Rathelot, and P. Zamora (2013): "Do Labor Market Policies have Displacement Effects? Evidence from a Clustered Randomized Experiment," *The Quarterly Journal of Economics*, 128, 531–580.

Davidson, C. and S. A. Woodbury (1993): "The Displacement Effect of Reemployment Bonus Programs," *Journal of Labor Economics*, 11, 575–605.

Davis, S. J., R. J. Faberman, and J. C. Haltiwanger (2012): "Recruiting Intensity during and after the Great Recession: National and Industry Evidence," *American Economic Review*, 102, 584–588.

Diamond, R. (2015): "The Determinants and Welfare Implications of US Workers' Diverging Location Choices by Skill: 1980-2000," Mimeo.

DiNardo, J. and D. Card (2000): "Do Immigrant Inflows Lead to Native Outflows?" *American Economic Review*, 90, 360–367.

Fan, Y. (2012): "The Planners' War against Spatial Mismatch Lessons Learned and Ways Forward," *Journal of Planning Literature*, 27, 153–169.

Farber, H. S. (2012): "Unemployment in the Great Recession: Did the Housing Market Crisis Prevent the Unemployed from Moving to Take Jobs?" *American Economic Review*, 102, 520–25.

Ferracci, M., G. Jolivet, and G. J. van den Berg (2014): "Evidence of Treatment Spillovers Within Markets," *The Review of Economics and Statistics*, 96, 812–823.

Friedberg, R. M. and J. Hunt (1995): "The Impact of Immigrants on Host Country Wages, Employment and Growth," *Journal of Economic Perspectives*, 9, 23–44.

Galenianos, M. and P. Kircher (2009): "Directed Search with Multiple Job Applications," *Journal of Economic Theory*, 144, 445–471.

Gautier, P., P. Muller, B. van der Klaauw, M. Rosholm, and M. Svarer (2012): "Estimating Equilibrium Effects of Job Search Assistance," Tinbergen Institute Discussion Paper 12-071/3.

Greenwood, M., G. Hunt, and J. McDowell (1986): "Migration and Employment Change: Empirical Evidence on the Spatial and Temporal Dimensions of the Linkage," *Journal of Regional Science*, 26, 223–234.

Guglielminetti, E., R. Lalive, P. Ruh, and E. Wasmer (2015): "Spatial search strategies of job seekers and the role of unemployment insurance," Mimeo.

Guimarães, P. and P. Portugal (2010): "A simple feasible procedure to fit models with high-dimensional fixed effects," *Stata Journal*, 10, 628–649(22).

Hausman, J., B. H. Hall, and Z. Griliches (1984): "Econometric Models for Count Data with an Application to the Patents-R&D Relationship," *Econometrica*, 52, 909–38.

Head, K. and T. Mayer (2004): "The Empirics of Agglomeration and Trade," *Handbook of Regional and Urban Economics*, 4, 2609–2669.

Hellerstein, J. K., D. Neumark, and M. McInerney (2008): "Spatial mismatch or racial mismatch?" *Journal of Urban Economics*, 64, 464–479.

Herz, B. and T. Van Rens (2015): "Accounting for mismatch unemployment," Mimeo, University of Warwick.

Ihlanfeldt, K. R. and D. L. Sjoquist (1998): "The spatial mismatch hypothesis: a review of recent studies and their implications for welfare reform," *Housing policy debate*, 9, 849–892.

Jackman, R., R. Layard, and C. Pissarides (1989): "On Vacancies," *Oxford Bulletin of Economics and Statistics*, 51, 377–394.

Jackman, R. and S. Roper (1987): "Structural Unemployment," *Oxford Bulletin of Economics and Statistics*, 49, 9–36.

Kline, P. and E. Moretti (2013): "People, Places and Public Policy: Some Simple Welfare Economics of Local Economic Development Programs," NBER Working Paper 19659.

Kroft, K., F. Lange, and M. J. Notowidigdo (2013): "Duration Dependence and Labor Market Conditions: Evidence from a Field Experiment*," *The Quarterly Journal of Economics*, 128, 1123–1167.

Lazear, E. P. and J. R. Spletzer (2012): "Hiring, Churn and the Business Cycle," *American Economic Review: Papers and Proceedings*, 102, 575–579.

Manning, A. and B. Petrongolo (2011): "How Local are Labor Markets? Evidence from a Spatial Job Search Model," IZA Discussion Paper 6178.

Marinescu, I. E. and R. P. Wolthoff (2015): "Opening the Black Box of the Matching Function: the Power of Words," Mimeo.

McKenzie, B. (2013): "County-to-County Commuting Flows: 2006-10," U.S. Census Bureau.

Molloy, R., C. L. Smith, and A. K. Wozniak (2011): "Internal Migration in the United States," *Journal of Economic Perspectives*, 25, 173–196.

Neumark, D. and J. Kolko (2010): "Do Enterprise Zones Create Jobs? Evidence from California's Enterprise Zone Program," *Journal of Urban Economics*.

Neumark, D. and H. Simpson (2014): "Place-Based Policies," NBER Working Paper 20049.

Nickell, S. (1982): "The Determinants of Equilibrium Unemployment in Britain," *The Economic Journal*, 92, 555–575.

Notowidigdo, M. J. (2011): "The Incidence of Local Labor Demand Shocks," NBER Working Paper 17167.

Ottaviano, G. and G. Peri (2006): "The economic value of cultural diversity: evidence from US cities," *Journal of Economic Geography*, 6, 9–44.

Patacchini, E. and Y. Zenou (2005): "Spatial mismatch, transport mode and search decisions in England," *Journal of Urban Economics*, 58, 62–90.

Poletaev, M. and C. Robinson (2008): "Human Capital Specificity: Evidence from the Dictionary of Occupational Titles and Displaced Worker Surveys, 1984-2000," *Journal of Labor Economics*, 26, 387–420.

Rupert, P. and E. Wasmer (2012): "Housing and the labor market: Time to move and aggregate unemployment," *Journal of Monetary Economics*, 59, 24–36.

Şahin, A., J. Song, G. Topa, and G. L. Violante (2014): "Mismatch Unemployment," *American Economic Review*, 104, 3529–64.

Valletta, R. G. (2013): "House lock and structural unemployment," *Labour Economics*, 25, 86–97.

Veracierto, M. (2011): "Worker flows and matching efficiency," *Economic Perspectives*, 147–169.

Wickham, H. (2009): *ggplot2: elegant graphics for data analysis*, Springer.

Wozniak, A. (2010): "Are College Graduates More Responsive to Distant Labor Market Opportunities?" *Journal of Human Resources*, 45, 944–970.

Yagan, D. (2016): "Is the Great Recession Really Over? Longitudinal Evidence of Enduring Employment Impacts," *Unpublished Manuscript*.

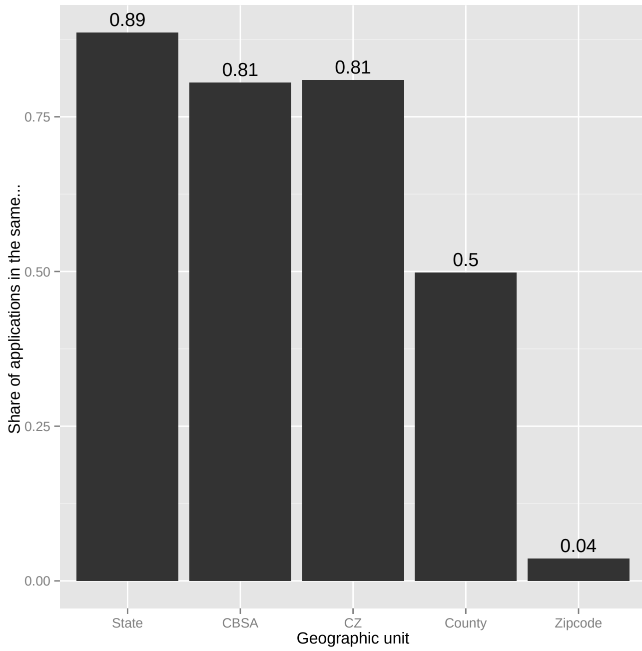

Figure 1: Average share of applications sent within the same geographic area
Source: CareerBuilder database.

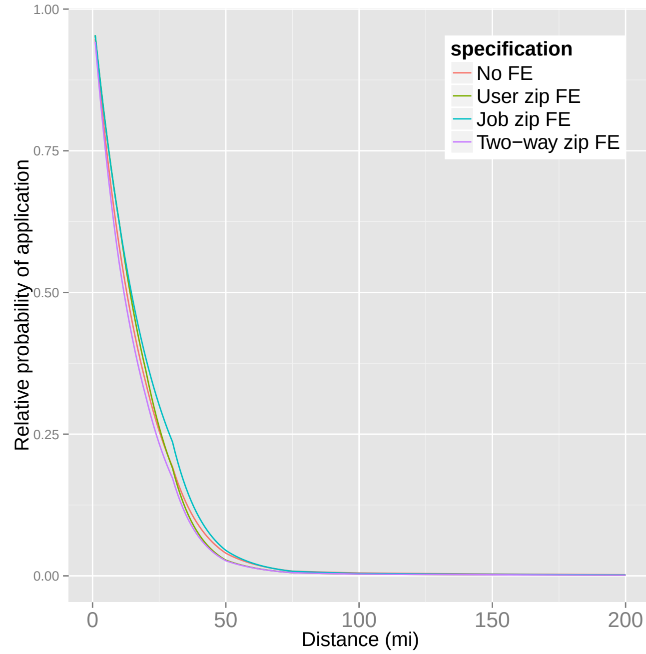

Figure 2: Relative probability of application as a function of geographic distance: predictions from Poisson model with or without fixed effects
Source: CareerBuilder database.

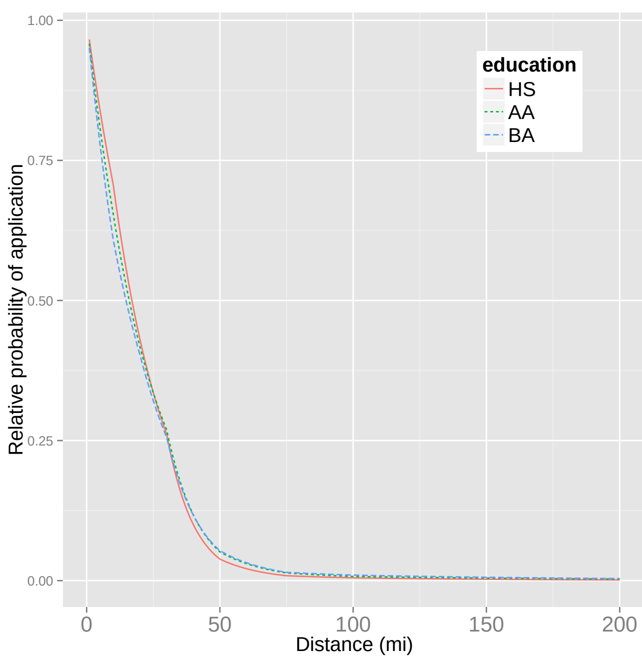

Figure 3: Relative probability of application as a function of geographic distance: predictions from Poisson model by education with job seeker ZIP code fixed effects
Source: CareerBuilder database.

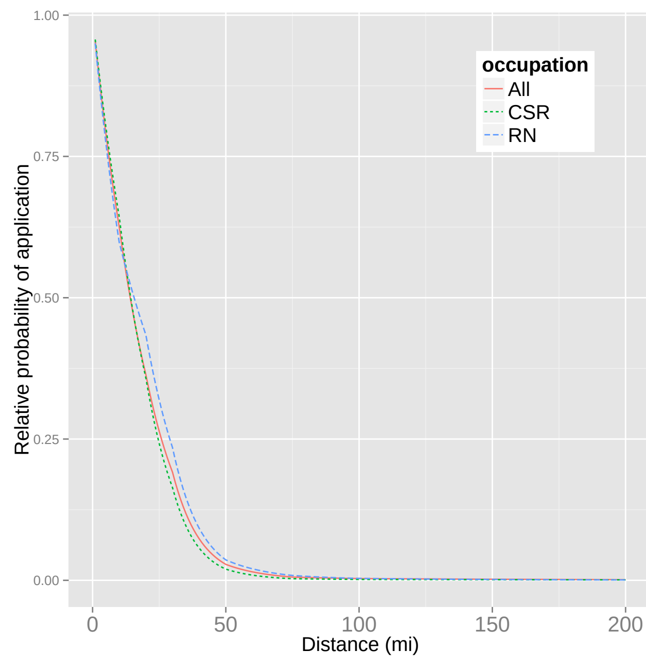

Figure 4: Relative probability of application as a function of geographic distance: predictions from Poisson model with job seeker ZIP code fixed effects for two common occupations
Source: CareerBuilder database.

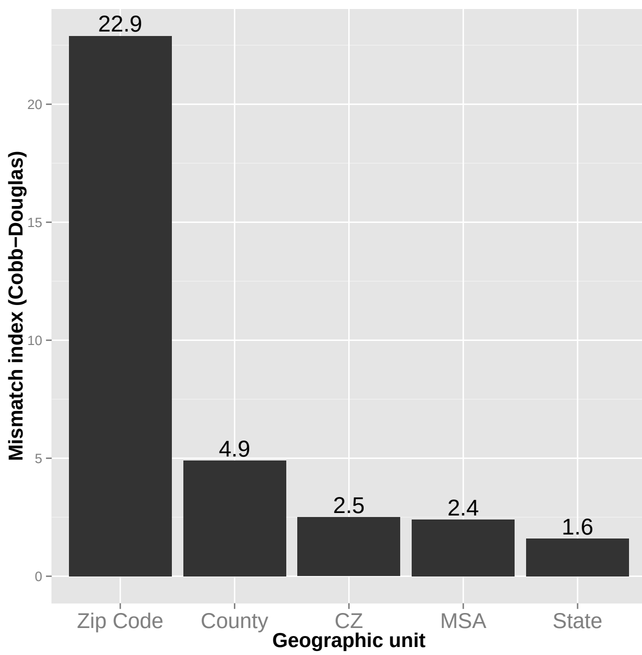

Figure 5: Mismatch unemployment: distinct markets and Cobb-Douglas matching function
Source: CareerBuilder database.

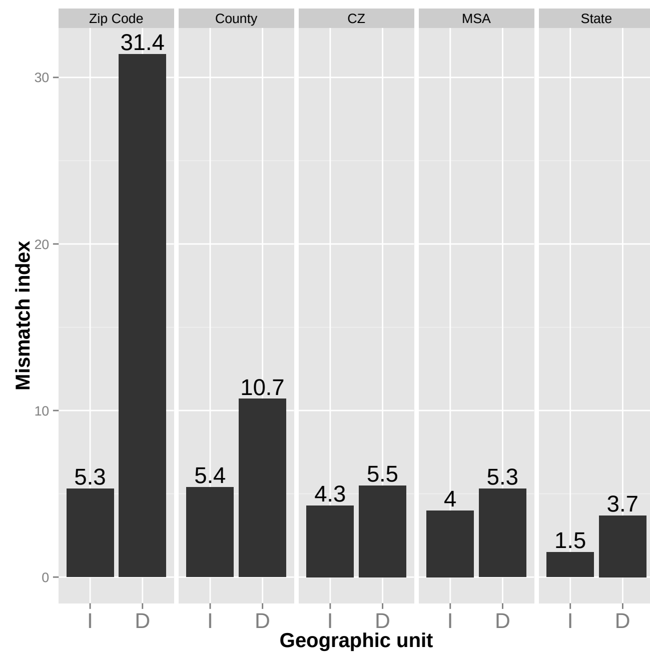

Figure 6: Mismatch unemployment: interconnected and distinct markets
Source: CareerBuilder database.

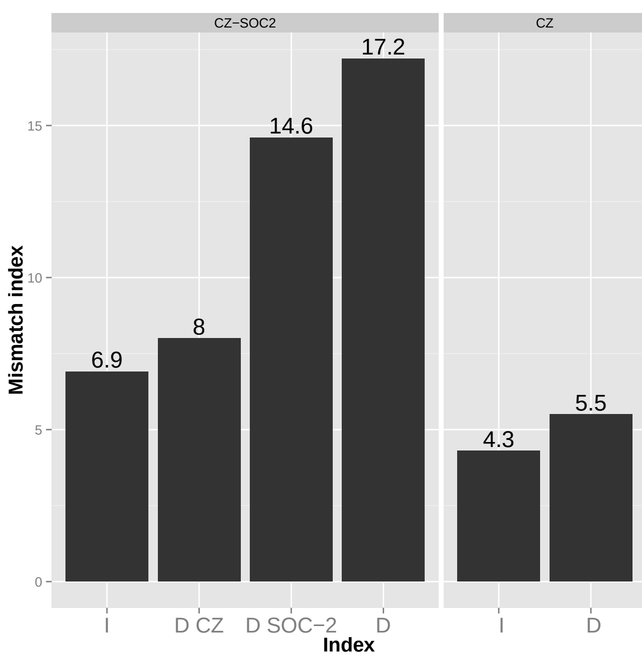

Figure 7: Mismatch when markets are defined as interaction of geographic and occupational units (CZ× SOC-2); and as CZ
Source: CareerBuilder database.

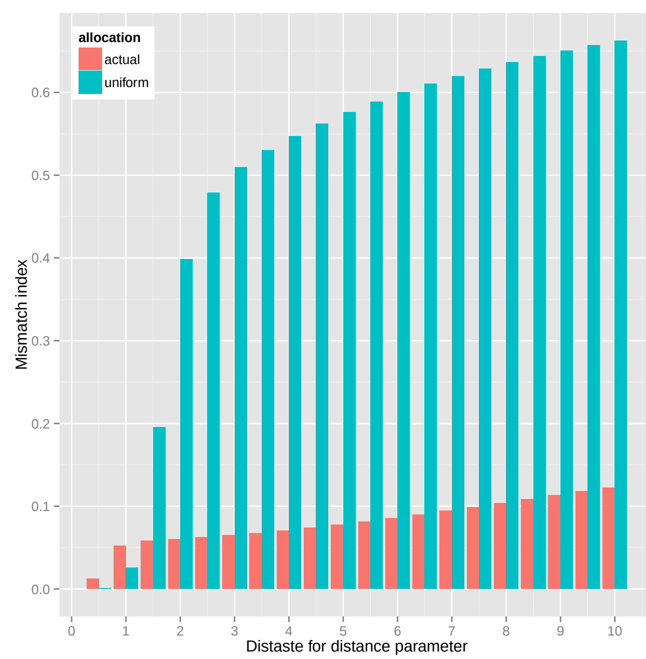

Figure 8: Robustness to various distaste-for-distance parameters, actual and uniform allocations
Source: CareerBuilder database.

Table 1: Probability of application as a function of distance: Poisson regression

	No fixed effect	User fixed effect	Job fixed effect	Two-way fixed effect (10 % sample)
γ_1	-0.054***	-0.047***	-0.047***	-0.059***
	(0.001)	(0.002)	(0.001)	(0.001)
γ_2	0.0005	-0.007***	-0.001	0.004***
	(0.002)	(0.002)	(0.001)	(0.001)
γ_3	-0.003**	-0.010***	-0.0001	-0.005***
	(.00158)	(0.002)	(0.002)	(0.001)
γ_4	-0.021***	-0.032***	-0.034***	-0.033***
	(0.001)	(0.002)	(0.001)	(0.001)
γ_5	0.016***	0.032***	0.014***	0.030***
	(0.001)	(0.002)	(0.002)	(0.001)
γ_6	0.043***	0.041***	0.044***	0.039***
	(0.001)	(0.002)	(0.002)	(0.002)
γ_7	0.011***	0.014***	0.016***	0.014***
	(0.001)	(0.002)	(0.001)	(0.001)
γ_8	0.005***	0.005***	0.005***	0.005***
	(0.000)	(0.000)	(0.000)	(0.000)
γ_9	0.004***	0.005***	0.005***	0.005***
	(0.000)	(0.000)	(0.000)	(0.000)
γ_{10}	-0.0002***	-0.0002**	-0.0003***	-0.0004***
	(0.000)	(0.000)	(0.000)	(0.000)
γ_{11}	0.001***	0.001***	0.001***	0.001***
	(0.000)	(0.000)	(0.000)	(0.000)
Number of Observations	3.37×10^8	3.37×10^8	3.37×10^8	1.84×10^7
Log-Pseudolikelihood	-7961547.5	-6419687.7	-6191151.4	-487220.13
Pseudo-R^2	0.7218			

Notes: Poisson model (column 1) or conditional Fixed-Effect Poisson model with user ZIP code fixed effects (column 2), job ZIP code fixed effects (column 3) or twoway fixed effects (column 4). Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

The 10 nodes for the spline that parametrizes workers' willingness to apply as a function of distance are at 10, 20, 30, 50, 75, 100, 200, 500, 1000 and 2000 miles. The piecewise-linear spline function is defined by its slopes. With 10 nodes $\{\bar{d}_i\}_{i=1...10}$, the spline is parameterized by 11 parameters $\{\gamma_i\}_{i=1...(11)}$. It is defined so that the derivative of the spline with respect to distance is $s'(d) = \gamma_1$ when distance is below the first node, i.e. when $d < \bar{d}_1$; $s'(d) = \sum_{i=1}^{j} \gamma_i$ when $d \in (\bar{d}_{j-1}, \bar{d}_j)$ and $j = 2 \dots 10$; $s'(d) = \sum_{i=1}^{11} \gamma_i$ when $d > \bar{d}_{10}$.

Table 2: Model parameters

Parameter	Notation	Value	Setting procedure
Job seekers in each ZIP code	U_i		From data
Vacancies in each ZIP code	V_j		From data, adjusted so that aggregate tightness matches V_{JOLTS}/U_{CPS}
Average number of applications	$\bar{a}$	12.6	From data
Probability of a valid application	q	0.0290	Estimated to match the national CPS job finding rate
Scaling parameter for the desutility of distance	η	1.0020	Estimated to match the geographic distribution of applications
Pareto parameter for match-specific utility shock	α	0.4629	Estimated to match the geographic distribution of applications

FOR ONLINE PUBLICATION: Appendix

Appendix A Job seekers' optimal application strategy

Here, we derive job seekers' optimal strategies. Let $\mathbf{v} = \{v_1, \dots v_{\bar{a}}\}$ be the $\bar{a}$-tuple of vacancies worker u applies to. We use the convention that utilities are ranked as: $w_{uv_1} \geq w_{uv_2} \geq \dots w_{uv_{\bar{a}}}$. The expected utility associated with strategy $\mathbf{v}$ is:

$$\mathcal{U}(\mathbf{v}) = \pi_{j(v_1)} w_{uv_1} + \sum_{k=2}^{\bar{a}} \left[\prod_{\ell=1}^{k-1} (1 - \pi_{j(v_\ell)}) \right] \pi_{j(v_k)} w_{uv_k} \tag{11}$$

With probability $\pi_{j(v_1)}$, the job seeker u gets an offer from the highest utility vacancy v_1, which is located in j. Whatever other offers he might get, he takes v_1 and his utility is w_{uv_1}. He only takes an offer from vacancy v_k if he does not get any offer from higher utility vacancies $v_{k'}, k' < k$, which happens with probability $\prod_{\ell=1}^{k-1}(1 - \pi_{j(v_\ell)})$.

Determining which strategy maximizes expected utility in equation 11 is complex: an algorithm such as the one described in Chade and Smith (2006) should be used. In the general case, it is *not* an optimal strategy to apply to the $\bar{a}$ highest expected utility jobs. Instead, workers should first apply to the highest expected utility job, and then gamble upwards by applying to jobs that have lower probability of yielding an offer but higher utility. Computing the optimal strategy using the Chade and Smith (2006) algorithm would make our model computationally intractable. We must therefore find some reasonable simplifying assumption to restore tractability.

One way to simplify the problem is to assume that the probability of a worker getting more than one offer is zero. Manning and Petrongolo (2011) assume that the probability of getting an offer from any given job is so low that the probability to receive two offers or more is negligible. In this case, the expected utility simplifies to $\mathcal{U}(\mathbf{v}) = \sum_k \pi_{j(v_k)} w_{uv_k}$, implying that the optimal strategy is to apply to the vacancies with the highest expected utility.

Another way of simplifying the problem is to assume that the probability π_v of getting an offer and the utility w_{uv} associated with a vacancy v are not negatively correlated. In this particular case, applying to the $\bar{a}$ vacancies with the highest expected utility is optimal, and the model becomes computationally tractable. The intuition is this: if the probability of getting the job and the reward are not negatively correlated, there is no trade-off between risk and reward (utility), and there is therefore no opportunity for gambling upwards. Therefore, if there is no negative correlation between the probability of getting an offer from a job in a location j and the utility derived from a job in location j, it is optimal to apply to the highest expected utility vacancies.

How likely is it that there is no negative correlation between the probability of getting an offer from a job in a location j and the utility derived from a job in location j? Utility is the product of two terms: $f(d)$ is strictly decreasing with geographic distance and ε is an idiosyncratic shock. By assumption, ε is a random draw across vacancies, and thus will not generate any correlation between the probability of getting an offer π and the utility w for a given vacancy.

Then, only a positive correlation between the probability of getting an offer π and the distance d may generate a negative (remember that $f(d)$ is strictly decreasing in d) correlation between the probability of getting an offer and utility. Unfortunately, it is hard to directly measure the correlation between π and the distance d because we don't observe the probability of getting an offer but instead infer it on the basis of applicants' behavior. Therefore, the inferred probabilities of getting an offer π_j in different locations j depend precisely on the assumption about the strategy pursued by job seekers. To make the case that the correlation is unlikely to be negative, we use two arguments. First, we show that, based on the structure of the problem and the data, the correlation between a job's utility and the probability of getting an offer is unlikely to be strongly negative. Second, we use the fact that, in the hires-maximizing allocation of job seekers, the correlation between the probability of getting an offer π and the distance d is zero and therefore non-negative.

Using the first line of argument, we can say that, in general, if job seekers are geographically dispersed as is the case in our data, π and d cannot be highly correlated either positively or negatively. To see this, suppose that there are only two places A and B, and two job seekers X and Y who live respectively in A and B. Jobs in place A have a higher probability π of generating an offer than jobs in place B. Therefore, for job seekers like X, there is a negative correlation between distance and the probability of getting an offer. For job seekers like Y, there is a positive correlation between distance and the probability of getting an offer. So, depending on the job seekers' location, the correlation between distance and the probability of getting an offer from a job could be positive or negative, implying that overall the correlation cannot be strongly positive or negative.

The question then becomes: how frequent are job seekers like Y and how often do opportunities for gambling upwards arise? In the simple example above, the opportunity for gambling upwards only arises if, for job seeker Y, jobs in A have a higher expected utility than jobs in B. In this case, job seeker Y would not only apply to jobs with the highest expected utility in A, but would want to gamble upwards by applying to jobs in their own location B that have a higher utility but a lower probability of yielding an offer. For jobs in A to have a higher expected utility than jobs in B for Y, it must be that the distance from B to A is not too large and/or that the probability of getting an offer from a job in A is large enough. More generally, this suggests that applying to the highest expected utility jobs is not optimal for job seekers in places where the probability of getting an offer increases more steeply with distance than the disutility of distance.

The conclusion of this first line of argument based on the structure of the problem and the data is this: as long as there are few job seekers for whom labor market conditions (as measured by the probability of generating an offer π) improve drastically within 60 miles or so of their place of residence (remember than 90% of application are sent within 60 miles), the assumption that the probability of getting an offer π and utility are not negatively correlated will be generally correct.

The second line of argument relies on the hires-maximizing allocation of job seekers. In this allocation, π_j is equal across all locations j (see equation 13): therefore, there is no correlation between the probability of getting an offer π and distance d, and so applying to the highest expected utility vacancies is indeed optimal. Since it turns out that the actual allocation of job seekers is fairly close to the hires-maximizing allocation of job seekers (there is little mismatch), the π_j tend to be very similar across locations, and there is therefore not much correlation between the probability of getting an offer π and distance d. In conclusion, the assumption that there is no negative correlation between the probability of getting an offer π and distaste for distance $f(d)$ seems reasonable given the structure of the problem and the fact that the allocation of job seekers is close to the hires-maximizing allocation.

Appendix B Number of matches when job seekers have no distaste for distance

Starting from equation (3), we examine the case in which job seekers have no distaste for distance, i.e. $g(d_{ij}) = 1, \forall i, j$. We derive the probability for a job seeker in i to apply to a vacancy in j p_{ij} as:

$$p_{ij} = \bar{a} \frac{\pi_j^\alpha}{\sum_\ell \pi_\ell^\alpha V_\ell}, \forall i, j \tag{12}$$

In this case, p_{ij} does not depend on i. Let $\bar{U}$ and $\bar{V}$ be the total number of job seekers and vacancies in the economy. We now derive the probability of getting an offer π. We have, for all j:

$$\pi_j = q\mathcal{R}\left(q\bar{a}\pi_j^\alpha \sum_k \frac{U_k}{\sum_\ell \pi_\ell^\alpha V_\ell}\right)$$
$$= q\mathcal{R}\left(\bar{U}\frac{q\bar{a}\pi_j^\alpha}{\sum_\ell \pi_\ell^\alpha V_\ell}\right)$$

The only term that depends on j on the right-hand side is π_j itself. Therefore, solving for π_j is the same for any ZIP code j. Hence π is equal across ZIP codes in the case of no distaste for distance. Since π is equal across ZIP codes, we can rewrite π as a function of parameters, i.e.:

$$\pi = q\mathcal{R}\left(q\bar{a}\frac{\bar{U}}{\bar{V}}\right) \tag{13}$$

If $g(d_{ij}) = 1$, the total number of matches is:

$$M = \sum_k U_k \left[1 - \exp\left(-\bar{a}\frac{\sum_\ell \pi_\ell^{1+\alpha} V_\ell}{\sum_\ell \pi_\ell^\alpha V_\ell}\right)\right] \tag{14}$$

Since π is equal across ZIP codes, the total number of matches when there is no distaste for distance is:

$$M = U\left[1 - \exp\left(-\bar{a}\pi\right)\right]$$

Replacing π by its expression in equation (13),

$$M = \bar{U}\left[1 - \exp\left(-q\bar{a}\mathcal{R}\left(q\bar{a}\frac{\bar{U}}{\bar{V}}\right)\right)\right] \tag{15}$$

Thus, the number of matches obtained with no distaste for distance depends on the aggregate number of job seekers $\bar{U}$ and the inverse of aggregate labor market tightness $(\bar{U}/\bar{V})$. Since there is no distaste for distance, only the aggregates matter: the location of jobs and job seekers is irrelevant. The total number of matches also depends on $q\bar{a}$, i.e. the product between the probability of a valid application and the average number of applications sent by a job seeker, which is equal to the average number of valid applications per job seeker. This makes sense since, intuitively, a larger number of valid applications leads to more matches.

Appendix C Mismatch unemployment by education

Our main results assume that job seekers are homogeneous: here we estimate mismatch while allowing for worker heterogeneity by education. Specifically, we divide job seekers in three educational groups: high school graduates, associate degrees (AA), and bachelor degrees (BA) and more.[32] We also compute the number of vacancies for each education category based on the SOC code of each vacancy and O*NET's determination of the level of education needed in each SOC code.

We compute mismatch by education assuming that job seekers only apply to jobs in their own educational category, so that each education level is a

[32] In our data, we cannot separate high school dropouts from individuals with missing information on education. While mismatch is likely to be higher for high-school dropouts than for high-school graduates, we cannot estimate a mismatch index for this category.

completely separate market. In a first version, we keep all parameters as in the baseline case (i.e. Table 2), except for the geographic distribution of job seekers and vacancies. Mismatch decreases with education (Figure 10). Yet, even for high school graduates, mismatch is only 6.9%. In a second version, we adjust all parameters for each education category, and we find that mismatch for high school graduates and AA is only about 4%, while mismatch for BA and above is only 1.8% (Figure 10).[33]

Overall, since mismatch remains low even for less educated workers, these results reinforce our main conclusion that geographic mismatch is a minor driver of U.S. aggregate unemployment.

[33] See Tables 4 and 5 in appendix for the parameters used and for the estimated distaste for distance parameters by education.

Appendix D A simpler mismatch index

In this section of the appendix, we investigate how mismatch varies with the Pareto parameter for the match-specific utility component α, and we show that a simpler mismatch index can be derived when $\alpha = 0$. Specifically, we vary α between 0 and 2 in increments of 0.2 (remember that our baseline estimate is $\alpha = 0.4629$). Mismatch is maximum at 6% when $\alpha = 0$ and decreases for larger values of α (appendix Figure 11). This makes sense because α can be interpreted as the weight put by applicants on the probability of getting an offer from a given vacancy relative to the distance to that vacancy (see equation 3). A smaller α increases mismatch because it hinders job seekers from directing applications to vacancies with higher probability of yielding an offer. Since we estimate a value of α that is close to 0, our geographic mismatch is close to the maximum that it could be as a function of α.

When $\alpha = 0$, job seekers only care about distance and do not take into account the probability of getting an offer when they apply, i.e. they are not strategic. In this case, the mismatch index simplifies considerably because we do not need to ensure that the probability of getting an offer π is consistent with the behavior of job seekers as was the case in equation 5. The mismatch takes a closed form that depends only on where job seekers and vacancies are located and job seekers' distaste for distance:

$$\mathcal{M}_{ns} = 1 - \sum_k \frac{U_k}{M^* \bar{U}} \left[1 - \exp\left(-q\bar{a} \frac{\sum_\ell g(d_{k\ell}) V_\ell \mathcal{R}(q\bar{a}\nu_\ell)}{\sum_\ell g(d_{k\ell}) V_\ell} \right) \right] \tag{16}$$

where $\mathcal{R}(x) = [1 - \exp(-x)]/x$, M^* is defined in equation (7) and ν_j is a generalized inverse tightness[34] in the no-strategy case defined as:

$$\nu_j = \sum_k \frac{g(d_{kj})U_k}{\sum_\ell g(d_{k\ell})V_\ell} \tag{17}$$

Mismatch with non-strategic job seekers is very similar but slightly higher than our baseline estimates (compare appendix Figure 12 and Figure 6 interconnected). This is not surprising since job seekers do not behave optimally: they apply to vacancies only as a function of distance, and do not take into account the probability of getting an offer. Overall, we conclude that, in the case of the U.S. in 2012, this mismatch index with non-strategic job seekers is a fair approximation of our more comprehensive approach.

Because it is much simpler to compute, this non-strategic mismatch index could be straightforwardly used to calculate mismatch with other datasets that contain the geographic distribution of job seekers and vacancies, U_i, V_j. Apart from the distribution of job seekers and vacancies, only two other ingredients are needed:

- The distaste for distance g, which we provide in Table 1. Alternatively, users can specify any other distaste for distance.
- $q\bar{a}$, the scale parameter, which should be calibrated using a target job finding rate.

[34] If we are interested in measuring the number of job seekers who compete for a job in a ZIP code j, we don't want to use the simple inverse tightness U_j/V_j because job seekers apply to jobs beyond their own ZIP code. Since labor markets are interconnected, the generalized inverse tightness at a place j will depend on the number of job seekers and job vacancies around j. To illustrate how the generalized inverse tightness ν_j varies with j, we plot it for each ZIP code j in the U.S. (appendix Figure 13).

To conclude, mismatch is maximum when $\alpha = 0$, but it is still only 6%. Furthermore, the assumption that $\alpha = 0$ yields a simpler mismatch index that can be used in other applications.

Appendix E Additional figures and tables

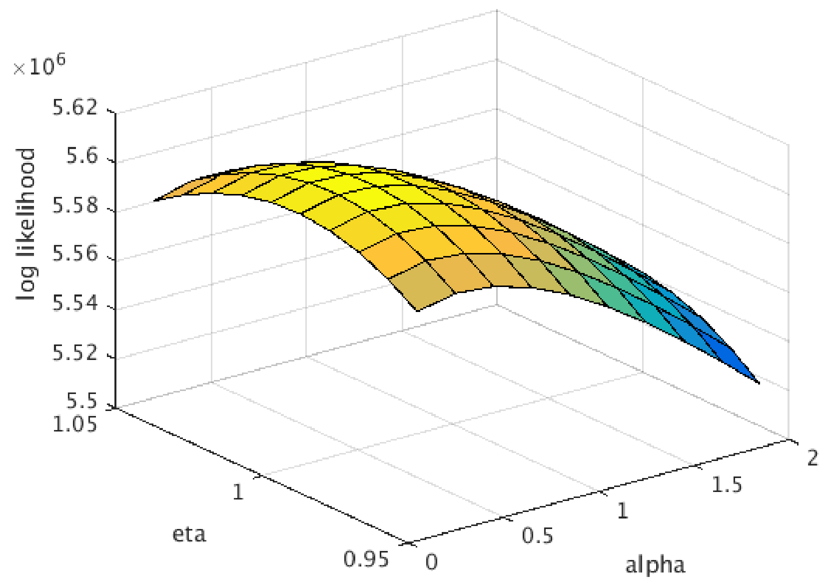

Figure 9: Log likelihood as a function of η, the scaling parameter for the distaste for distance, and α the Pareto parameter for the match-specific utility shock

Source: CareerBuilder database.

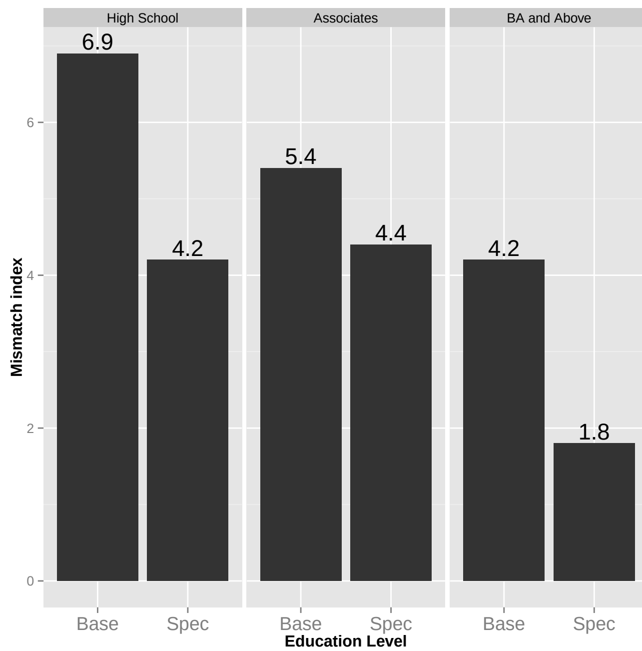

Figure 10: Mismatch unemployment by education: baseline parameters ("Base") and each education category's own specific parameters ("Spec")

Source: CareerBuilder database.

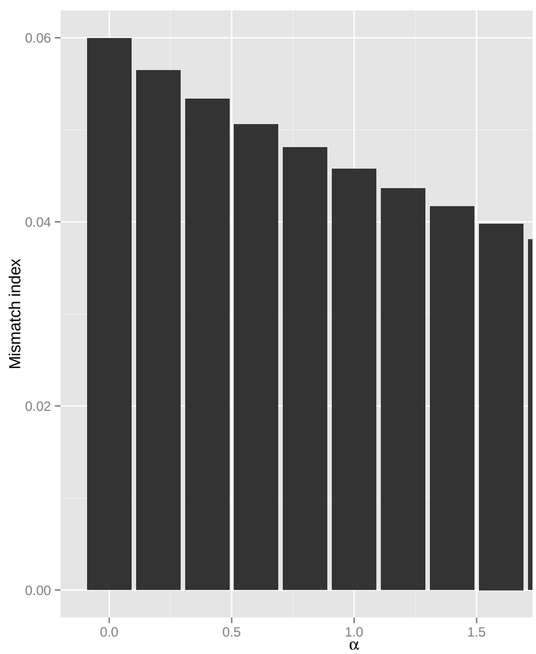

Figure 11: Robustness to various values of the Pareto parameter for the match-specific utility component α
Source: CareerBuilder database.

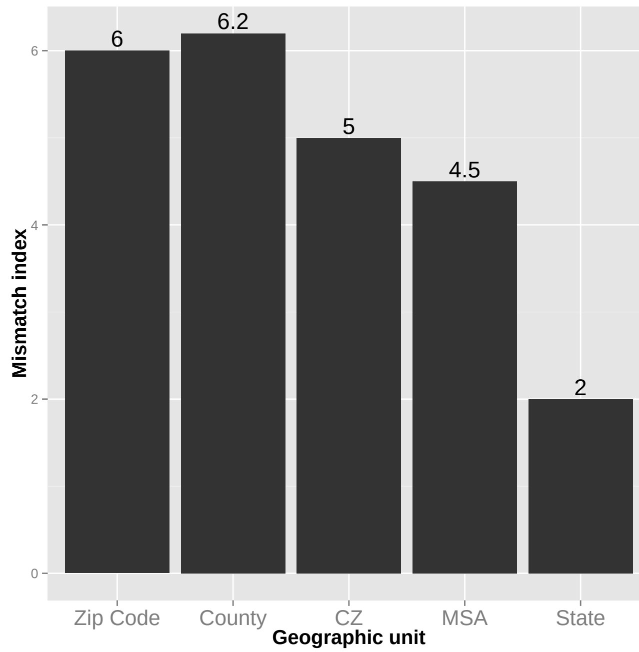

Figure 12: Mismatch unemployment with interconnected markets and non-strategic job seekers
Source: CareerBuilder database.

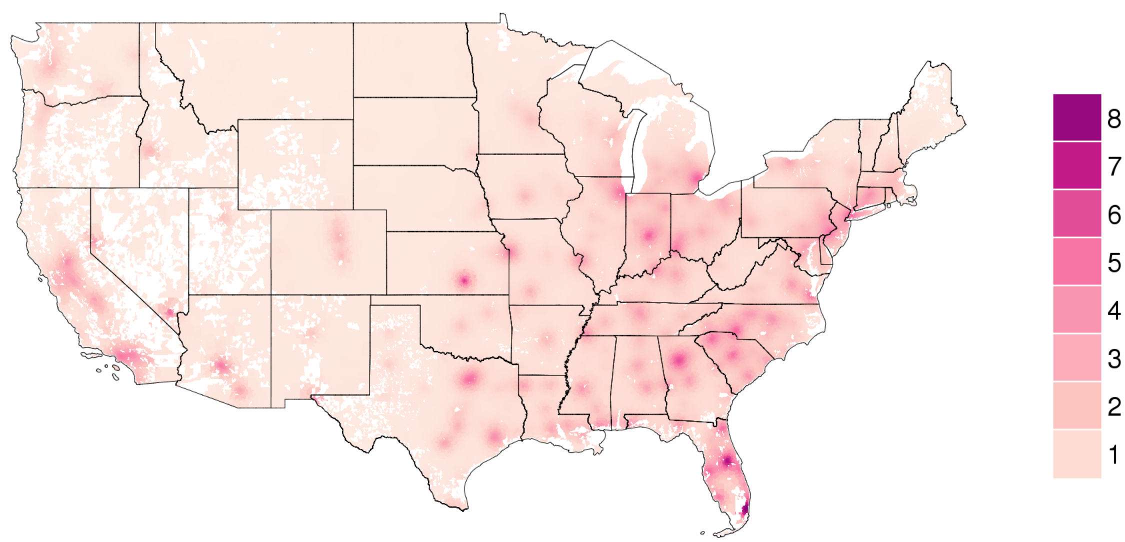

Figure 13: Generalized inverse tightness: number of unemployed workers per job, taking into account the geography of job search

Source: CareerBuilder database.

Table 3: Estimation of the $CZ \times SOC$ model

	(1)	(2)	(3)
Geographic distance			
< 50 miles	-0.0405	-0.0711	-0.0671
	0.0027	0.0021	0.0025
< 75 miles	0.0190	0.0586	0.0582
	0.0075	0.0068	0.0080
< 100 miles	-0.0573	-0.0564	-0.0603
	0.0115	0.0098	0.0117
< 200 miles	0.0598	0.0459	0.0463
	0.0072	0.0061	0.0073
< 500 miles	0.0155	0.0193	0.0194
	0.0011	0.0014	0.0016
$< 1,000$ miles	0.0031	0.0032	0.0030
	0.0002	0.0003	0.0003
$< 2,000$ miles	0.0001	-0.0003	-0.0004
	0.0001	0.0002	0.0002
$> 2,000$ miles	0.0004	0.0022	0.0021
	0.0001	0.0003	0.0003
SOC2			
Different SOC2	-1.2922	-1.0231	-0.7224
	0.0696	0.0539	0.0437
Distance SOC2	-0.2271	-0.3734	-0.4532
	0.0261	0.0223	0.0113
Difference Factor 1	0.2957	0.6082	0.0061
	0.0112	0.0130	0.0119
Difference Factor 2	0.2598	0.3551	0.1916
	0.0121	0.0116	0.0100
N	83,533,150	80,833,282	67,653,650
Fixed-effects	No	User CZSOC	Job CZSOC

Notes: Poisson model (column 1) or Conditional Fixed-Effect Poisson model (columns 2 and 3). Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.
The 10 nodes for the spline that parametrizes workers' willingness to apply as a function of distance are at 10, 20, 30, 50, 75, 100, 200, 500, 1000 and 2000 miles. The piecewise-linear spline function is defined by its slopes. With 10 nodes $\{\bar{d_i}\}_{i=1...10}$, the spline is parameterized by 11 parameters $\{\gamma_i\}_{i=1...(11)}$. It is defined so that the derivative of the spline with respect to distance is $s'(d) = \gamma_1$ when distance is below the first node, i.e. when $d < \bar{d}_1$; $s'(d) = \sum_{i=1}^{j} \gamma_i$ when $d \in (\bar{d}_{j-1}, \bar{d}_j)$ and $j = 2 \dots 10$; $s'(d) = \sum_{i=1}^{11} \gamma_i$ when $d > \bar{d}_{10}$. Different SOC2 is a dummy for the SOC2 of the applicant's last job differing from the SOC2 of the vacancy. Distance SOC2 is the distance between the applicant's SOC2 and the vacancy's SOC2. Difference Factor 1 is the difference between the first factor of the applicant's SOC2 and the first factor of the vacancy's SOC2; the same definition holds for Difference Factor 2.

Table 4: Parameters for each education category

Parameter	High School	Associates	BA and above
Number of applications	13.8	14.0	13.6
Tightness	0.20	0.33	0.80
Job Finding Rate	0.17	0.17	0.20

Table 5: Probability of application as a function of distance by education: Poisson regression

	(1) High School	(2) AA	(3) BA and Above
γ_1	-0.0348***	-0.0425***	-0.0498***
	(0.00476)	(0.00222)	(0.00531)
γ_2	-0.0146**	-0.00239	0.00867
	(0.00695)	(0.00338)	(0.00799)
γ_3	-0.000455	0.00112	-0.00425
	(0.00698)	(0.00299)	(0.00736)
γ_4	-0.0461***	-0.0391***	-0.0329***
	(0.00578)	(0.00259)	(0.00705)
γ_5	0.0367***	0.0302***	0.0268***
	(0.00635)	(0.00306)	(0.00756)
γ_6	0.0368***	0.0278***	0.0353***
	(0.00787)	(0.00407)	(0.00827)
γ_7	0.00855	0.0152***	0.00627
	(0.00546)	(0.00287)	(0.00513)
γ_8	0.00974***	0.00500***	0.00631***
	(0.00166)	(0.000822)	(0.00153)
γ_9	0.00401***	0.00431***	0.00346***
	(0.000873)	(0.000334)	(0.000626)
γ_{10}	-0.000335	4.36e-05	-0.000238
	(0.000458)	(0.000230)	(0.000325)
γ_{11}	3.07e-05	0.000253	0.000698***
	(0.000367)	(0.000214)	(0.000237)
Observations	57,997,472	178,134,756	29,997,033
Log-PseudoLikelihood	-122959.6	-1256988.9	-100679.32

Notes: Conditional Fixed-Effect Poisson model with user ZIP code fixed effects. Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.
The 10 nodes for the spline that parametrizes workers' willingness to apply as a function of distance are at 10, 20, 30, 50, 75, 100, 200, 500, 1000 and 2000 miles. The piecewise-linear spline function is defined by its slopes. With 10 nodes $\{\bar{d}_i\}_{i=1...10}$, the spline is parameterized by 11 parameters $\{\gamma_i\}_{i=1...(11)}$. It is defined so that the derivative of the spline with respect to distance is $s'(d) = \gamma_1$ when distance is below the first node, i.e. when $d < \bar{d}_1$; $s'(d) = \sum_{i=1}^{j} \gamma_i$ when $d \in (\bar{d}_{j-1}, \bar{d}_j)$ and $j = 2 \ldots 10$; $s'(d) = \sum_{i=1}^{11} \gamma_i$ when $d > \bar{d}_{10}$.