Forecasting U.S. Domestic Migration Using Internet Search Queries

Allen Yilun Lin
Northwestern University
Evanston, Illinois
allen.lin@eecs.northwestern.edu

Justin Cranshaw
Microsoft Research
Redmond, Washington
justincr@microsoft.com

Scott Counts
Microsoft Research
Redmond, Washington
counts@microsoft.com

ABSTRACT

Roughly one in ten Americans move every year, bringing significant social and economic impact to both the places they move from and places they move to. We show that migration intent mined from internet search queries can forecast domestic migration and provide new insights beyond government data. We extract from a major search engine (Bing.com) 120 million raw queries with migration intent from 2014 to 2016, including origin and destination geographies, and the specific intent for migration such as whether the potential migration is housing or employment related. Using these queries, we map U.S. state level migration flows, validate them against government data, and demonstrate that adding search query-based metrics explains variance in migration prediction above robust baseline models. In addition, we show that the specific migration intent extracted from these queries unpack the differential demands of migrants with different demographic backgrounds and geographic interests. Examples include interactions between age, education, and income, and migration attributes such as buying versus renting housing and employment in technology versus manual labor job sectors. We discuss how local government, policy makers, and computational social scientists can benefit from this information.

KEYWORDS

internet search; migration; big data; housing; employment

ACM Reference Format:
Allen Yilun Lin, Justin Cranshaw, and Scott Counts. 2019. Forecasting U.S. Domestic Migration Using Internet Search Queries. In *Proceedings of the 2019 World Wide Web Conference (WWW'19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3308558.3313667

1 INTRODUCTION

Every year over 32 million Americans, about 10% of the total population, migrate from one location in the United States to another [34]. The inflow and outflow of domestic migration can have substantial economic, political, and social impact. Because of this, many cities and states track and analyze migration data closely, looking to inform policies designed to retain existing residents and attract domestic migrants from other regions. For example, the Vermont General Assembly recently passed legislation that pays remote workers $10,000 if they move to Vermont from other states [50]. Even among states such as Texas and Arizona that have no difficulty attracting migrants, improved migration measurement and forecasts can aid development and planning related to local housing and job markets.

To track domestic migration in the United States, state governments and other interested parties rely almost exclusively on two migration datasets from the federal government [48], one from the American Community Survey (ACS) and another from the Internal Revenue Service (IRS). Both have limitations that might be mitigated by incorporating signals from large scale, naturalistic sources such as internet search queries. Most importantly, ACS and IRS data typically take two years to collect, prepare, and publish, forcing the consumers of these data to base current decisions on past, rather than fresh, measurements. Also, because ACS utilizes survey methods, the resulting data are based on a comparatively small samples, resulting in larger margins of error for smaller regions. IRS migration data, on the other hand, are derived from tax return records. People filing tax returns from a different state than they did the year prior are assumed to have migrated. While this method has a considerably larger sample size than ACS, it is subject to population biases, as it underestimates the migration of poorer, wealthier, and elderly citizens, all of whom may file taxes less frequently [29].

In terms of why people move, in over a century of research on migration, scholars have identified *housing* and *employment* as the two key motivators for moving [4, 30, 35, 69]. Prior research has demonstrated the predictive power of internet search in forecasting housing and jobs related measures. For instance, Wu and Brynjolfsson enhanced prediction of housing demand using Google searches [68]. Likewise, search queries have been used to "nowcast" unemployment rates [3] and job search interests [13]. Thus given how central the internet is in helping people find both jobs and housing, we explore whether internet search data can be used as a signal for forecasting and explaining domestic migration, with a focus on housing and employment as primary drivers of intent to migrate.

Using data from Microsoft's Bing search engine, we extracted and assessed the value of search queries from people researching potential moves. Drawing on theories from migration studies and the web search literature, we designed a highly accurate migration query filter that performs two tasks: it retrieves queries expressing the intent to migrate with F_1 score of 0.880, and it categorizes each query into specific types of migration intent, such as whether it indicates housing or employment needs. Using 120 million raw queries containing migration intent from 2014 to 2016, we show that these internet search data provide a meaningful measurement of migration when validated against core government migration

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313667

statistics. We further show that adding search-based variables to a model can improve its predictive power at forecasting future net migration flow at the U.S. state level, compared to two well-established and strong baseline models from migration literature. Finally, we examine the explanatory power of migration search data through the lenses of demography and geography, showing how an analysis of the different categories of migration queries (e.g., housing versus job, renting versus buying, technology versus manufacturing jobs) offers new insights beyond what government data provide.

In summary, this paper contributes a novel approach to domestic migration measurement and forecasting that utilizes search query data that are larger in scale, more timely, and more nuanced in explanation than current government migration data. Although we focus our analysis on the United States, we expect the approach would generalize to any region with reasonable internet penetration.

2 RELATED WORK

2.1 Motivation for Domestic Migration

Theories from migration studies have suggested that the needs for housing and employment are two forces that drive domestic migration. Researchers have conceptualized migration as a search and matching process [19, 48, 59] in that prior to actual moving, migrants are constantly searching for places that maximize the net benefit of moving. In particular, housing and employment are two of the most important factors when considering the benefits of moving [4, 30, 35, 69]. That is, on the one hand, people are willing to move to places with better job opportunities [15, 38], while on the other hand, the costs of housing could motivate or limit migration to certain areas [9, 36]. Although there has been debate on whether housing leads employment in motivating migration or vice versa [45, 49, 62], the consensus is that migrants almost inevitably need to search for both of them in their migration activities [48].

Empirical results from census surveys reaffirm that migrants are motivated by housing and jobs. According to the latest *Reason for Moving* report published by Census Bureau [33], employment-related reasons and housing-related reasons motivate nearly 68% domestic migrants. Although the report also shows that family-related reasons motivate 30% migrants, closer examination [32] reveals that some of the family-related reasons directly involve housing activities, including "To Establish Own Household."

As such, given how housing and employment are central in domestic migration, we consider housing and job as the most relevant topics when defining search queries with migration intent.

2.2 Macroeconomic Forecasting Using Search

With the internet becoming the default channel to search for information, economists and computer scientists have used search queries to model macroeconomic measurements that previously were available only through government and financial agencies, including phenomena such as car sales [16] and the stock market volumes [5, 54, 55]. Housing and employment demand, again two macroeconomic factors highly relevant to migration, have also been successfully modeled with internet searches. Most prior work relies on Google Trends, which are normalized, aggregated query volumes for particular keywords and entities. Google Trends data have also been used, for example, to forecast unemployment rate for both U.S. [16, 20, 22] and other countries [52].

Similar work has been done to predict housing trends. Wu and Brynjolfsson [68] have shown that search frequencies from Google could enhance the prediction of the Housing Price Index (HPI) on both national scale and local scale. McLaren and Shanbhogue [46] have used a similar approach to nowcast housing market dynamics in UK. Finally, other projects have sought to go beyond predicting macroeconomic measures to understanding socioeconomic drivers and effects, such as the relationship between searches for different types of jobs and searchers' demographics [13].

Taken together, this body of work provides us confidence about the applicability of search query based data to accurately model demand for housing and employment, the two key motivations related to migration. Methodologically, we were also able to leverage keyword dictionaries established in this prior work to distinguish employment and job related migration search queries.

2.3 Identifying Migrants with Internet Data

Most relevant to our work, a number of studies from computational social science have used internet data to identify migrants. The first type of work does so by mining personal digital traces on social media platforms. Using Twitter data, Zagheni et. al. [70] identified migrants as those whose geolocated tweets exhibit several distinct geographic clusters chronologically. Later work by Fiorio et. al. [24] also used Twitter data to identify and distinguish between short-term movers and long-term migrants. The location field in online profiles from other social network platforms, including Google+ [47], Facebook [51] and LinkedIn [2], have also been exploited to identify migrants. Finally, by using IP address data from Yahoo!'s 100 million email users, researchers identified international migrants and studied their migration patterns across many demographic attributes [67, 71].

However, as the data collected from this prior work only identified migrants *after* they moved, they only shed light on existing migration trends and will be less useful in forecasting migration. Decision makers who rely on this information can not prepare themselves beforehand. Moreover, previous methods that use open-access data from Twitter can only collect data for a small percentage of migrants, which limits their ability at accurately predicting national level migration statistics. [70].

By mining large scale search queries, our approach has two advantages over prior work. First, search queries usually precede actual migration, offering a unique advantage for prediction. Second, the content of the search queries offers additional insights into the specific types of demand (e.g., housing or job related) that can indicate economic impact to local markets.

3 MIGRATION QUERIES FILTER

In this section, we define *migration queries* in the context of this paper. We then describe how we design a filter to accurately extract migration queries from a larger pool of search queries.

3.1 Defining Migration Queries

Formally, we define migration queries as *housing or employment queries that explicitly target a destination geographic unit different*

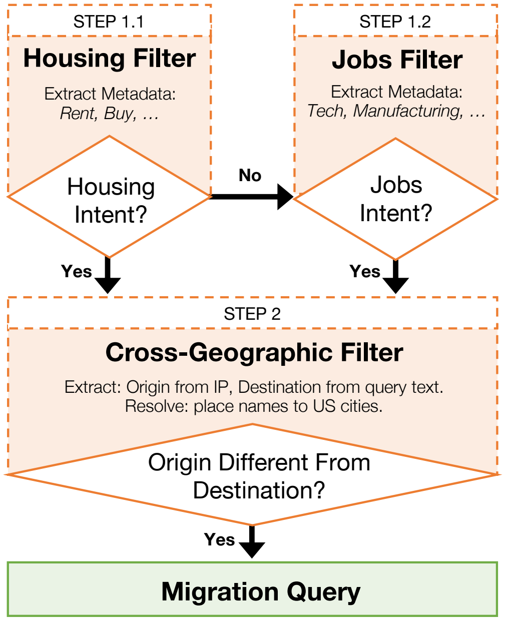

Figure 1: Simplified flowchart of our migration query filter.

from the origin of the query. Our emphasis on housing and employment queries is driven by the theories from migration literature (see Related Work) in which housing and jobs are two major reasons for moving. In the context of search queries, housing and employment keywords make the intent to migrate more apparent. For instance, the housing related term “apartments” in a query such as “apartments in Boston” that originated from Phoenix raises the likelihood of an intended move from Phoenix to Boston.

An important component of our definition is the notion of “geographic unit”. Migration researchers usually select relatively large geographic units, including county, metropolitan statistical area (MSA), core-based statistical area (CBSA) and state [48]. However, empirically we observed that housing and employment queries often mention the more specific geographic units (e.g. “apartment to rent in DC”, “software engineer job in San Francisco”) than those that are used in migration studies. This trend is particularly prominent for housing queries that target a big metropolitan area where certain cities or even neighborhoods within that MSA function as residential areas. As such, we decide to focus on *county*, *city*, and *zip code* in the migration query retrieval process (detailed below) and then later aggregated them to larger geographic units (states) used in the analyses. We also note that our definition requires an explicit mention of the geographic unit in the query. We do so because housing or job queries without mentioning geography, such as “data scientists jobs”, “apartment near me” are more likely to be local rather than migration related.

3.2 Design of the Filter

Before we detail the implementation of the migration query filter, we discuss two high-level challenges that influence many specific design choices. First, our task requires identifying *both* the topic and the destination geography of the query, a problem not seen in previous work. As such, we can’t leverage a readily available search trend product such as Google Trends since its statistics are not segmented by query destination, and instead must filter raw queries. This presents the second challenge: processing an enormous amount of data.

To address the first challenge, we divide the task into extracting housing and jobs topics and extracting place names. We leverage prior studies that use housing and jobs related search queries (e.g. [13, 16, 68]) to guide our selection of keywords. For extracting geography, we turn to the literature of toponym disambiguation [7, 8, 60]. While different approaches exist, an external knowledge source, such as gazetteers or Wikipedia, usually serves as the foundation to more advanced methods [18, 66], and often times provide a competitive baseline accuracy [60]. As such, we adopt a straightforward gazetteer approach to build our geography extractor which will be detailed below. To address the second challenge of large data volume, our design prioritizes *run time* over complexity and *precision* over recall. As such, we employ rule-based filters and avoid machine learning approaches whenever possible. However, we show that our accuracy is comparable to prior work that implements machine learning based filters.

Figure 1 presents the detailed implementation of our migration query filter. After filtering out abnormally long queries, each raw query is first passed through a housing query filter (Step 1.1). To implement this filter, we leverage the keywords from prior work that uses housing search queries to forecast housing sales [68], which includes “real estate”, “apartment”, “property”, “house” etc. Notably, we expanded these keywords by adding the names of popular housing websites such as “zillow” and “redfin.” Known as “navigational queries” in the web search literature [6], search queries containing these websites also indicate users’ intention for housing information. Specifically, we added the names for the top ten housing websites according to rankings from Alexa.com. If a query is identified as a housing query, our filter identifies whether the user’s intent is to *buy* or *rent* housing by looking for the existence of terms such as “prices”, “sale” (for buying) and “rent”, “rental” (for renting). These housing queries are saved for location extraction, which will be described below.

Non-housing queries are subsequently processed by a jobs query filter (Step 1.2). We took a similar approach to construct a set of keywords to identify job related migration searches. First, we obtained the specific keywords for 14 job categories that were used in [13]. In a later analysis, we focus our attention on four of these categories: Science, Technology, Manufacturing, and Transportation. Similar to the housing query filter, these topical keywords were expanded with navigational search terms that refer to top 10 job search websites from Alexa.com, including “indeed”, “monster” etc.

If a query is either a housing query or job query, it’s transferred to Step 2 to extract and compare the origin and the destination geography. While the origin of the query is contained in the reverse-IP look up results, which was provided to us as part of the query data sample, extracting the destination of the query is non-trivial. As we discussed earlier, we leveraged a gazetteer to identify the reference of cities, counties and zip codes in the query. While the zip code and county names are straightforward to obtain, cities are ambiguously defined in U.S. geography. We adopt the Census definition of “place”, which is most similar to scale of a colloquial “city” [26] and includes

	#	%	Example queries
Housing	100,123,000	83	"condo in san francisco", "home for sale in maricopa county", "apartment in santa clara, ca"
↪ **Buy**	39,095,000	39	"home for sale in maricopa county", "house price 20500", "duplex for sale in chicago"
↪ **Rent**	32,144,000	32	"apartment for rent in seattle", "house for rent in texas city, texas", "apartment for rent 60231"
Jobs	20,890,000	17	"jobs in st paul", "software engineer jobs in mountain view", "accounting jobs in cook county"
↪ **Sci/Tech**	204,000	1	"software engineer jobs in mountain view", "jobs in r&d in boston", "sql dba jobs in tampa, fl"
↪ **Mfg/Transp**	719,000	3	"van driver job in santa clara ca", "schoolbus driver jobs in king county", "cdl jobs in philadelphia"
Total	121,014,000		

Table 1: Distributions and examples of migration queries. Number are rounded and content is very slightly altered for privacy reasons. Percentages for sub-type queries are normalized by all housing or all jobs queries accordingly.

both Incorporated Places and Census Designated Places. In total, we extracted over 20,000 such names.

When matching these place names to the query content, an additional challenge is the ambiguity problem. For example, there are over 30 cities named "Franklin" in the United States. To increase the precision, we took a two step approach. First, for the largest 50 cities in the U.S. we assume the city name refers to that city (e.g., "chicago" in the query "biotech jobs Chicago" refers to Chicago, IL). For smaller cities, we disambiguate with a state name or abbreviation (e.g., "biotech jobs Portland ME") if one is present in the query and adjacent to the city name. In the second step, queries containing names of smaller cities that lacked an accompanying state name or abbreviation were post-processed by Bing's own location disambiguation service. This disambiguation process breaks the query into n-grams, identifies location names, and then provides a ranked list of the most likely matching locations using features such as the popularity of the location and proximity to the origin of the query (as determined by the reverse-IP look up). Using this, we chose the query destination to be the the top ranked result returned by the disambiguator. Generally, this process conforms to how search engines interpret location queries, and how users behave after they understand search engines' interpretations. For instance, users typically know to specify the state of a city when that city is small or is not in their immediate locale. Finally, we compare the origin and destination on *state* level for the purpose of analyses (detailed later) and extract queries whose origin states are different from destination states as migration queries.

3.3 Evaluation and Filtering Results

We conducted a lightweight evaluation of our migration query filter before deploying it on all queries. We first developed a test set by collecting 500 housing and job queries as identified by Step 1.1 and Step 1.2 in Figure 1. We used this as our test set rather than a random sample from all queries because migration queries are only a small percent of all queries. Two researchers coded the 500 queries as either *migratory* or *non-migratory* according to the definition of migration queries. For samples with disagreement, the researchers discussed and reached consensus on a label. Coding results show 92 migration queries among these 500 housing and jobs queries.

Comparing these human labels to the results of our migration query filter indicates generally strong classification results. Our filter achieves high performance on all metrics, including precision (0.870), recall (0.891), F1 score (0.880) and accuracy (0.880). We note that our filter's performance is comparable to prior work that filtered search queries into job categories [13]. We applied this filter on a large sample of queries from 2014 to 2016, yielding just over 121 million migration queries in total (see Table 1). Among these, 83% are housing-related and 17% are job-related. Among housing migration queries, we observe a roughly equal amount of buy-specific and rent-specific queries. Among job migration queries, there are three times as many Manufacturing and Transportation related migration queries as there are Science and Technology related queries, which is consistent with prior work [13].

4 FINDINGS AND INSIGHTS

The data collection and filtering pipeline described above produced a corpus of more than 120 million migration queries from 2014, 2015 and 2016, each categorized with information about its origin state, destination state and specific types of migration intent. In this section, we detail our analyses and findings, organized into three parts. First, we validate our migration queries against external government statistics, showing that our data are highly correlated with core migration measures. Second, we prove that adding variables derived from these queries enhances the prediction of migration measurements over two robust baseline models from the migration literature. Third, we demonstrate that by extracting categorical descriptors of the type of query, such as whether it is about housing or jobs, we are able to shed light on the differential demands of migrants from varied demographic backgrounds and geographic interests.

We conduct all comparisons to government measures using data from the *American Community Survey* (ACS), focusing on migration between the 50 states, excluding the District of Columbia. ACS data is chosen for the following reasons. First, among all popular government migration data sources, only ACS provides migrants' detailed demographic information, which allows us to understand how migration-related search demands vary across different demographics. Second, despite being only a 1% sample of the total population, ACS has shown to be equally accurate to IRS migration data, which cover 87% of American households, on many important migration measures [48]. We focus on the state level because it is a popular aggregation unit used in the traditional migration literature (e.g. [12, 27, 53, 61, 64]) and because it bears smaller sampling errors compared to finer-grained units of aggregation.

Since state-level migration data are geographic data, we tested for the existence of spatial autocorrelation. Spatial autocorrelation describes a well-known phenomenon in spatial data that measurements taken at nearby locations tend to be correlated [43], which

	Inflow Correlations		
	Migration 14	Migration 15	Migration 16
Queries 14	0.72	0.78	0.75
Queries 15	0.67	0.73	0.73
Queries 16	0.65	0.71	0.73
	Outflow Correlations		
	Migration 14	Migration 15	Migration 16
Queries 14	0.32	0.59	0.60
Queries 15	0.26	0.53	0.52
Queries 16	0.21	0.46	0.44
	Netflow Correlations		
	Migration 14	Migration 15	Migration 16
Queries 14	0.11	0.55	0.58
Queries 15	0.06	0.45	0.59
Queries 16	0.13	0.42	0.61

Table 2: Correlations between migration rates and migration queries by year for inflow, outflow, and netflow respectively.

violates the independence assumption that underlies many statistical tools. We followed the state-of-the-art procedure for testing spatial autocorrelation in [21, 65], finding weak to no spatial autocorrelations on all variables used in our analyses (defined below), and thus we maintain use of conventional statistical approaches.

4.1 Correlations with migration statistics

In this section, we validate our query-based migration measures. To do so, we examine the relationship between the volume of migration queries and three core migration measures from government: (1) inflow migration rate, which is the number of people who move into a specific state in a year, normalized by the state population; (2) outflow migration rate, which is the number of people who move out of a specific state in a year, normalized by the state population; and (3) net migration rate, which is the difference between inflow and outflow migration rates. These measures are used widely in many social science studies [10, 12, 61] and are the primary indicators of the nation's overall migration status. Although they are estimated annually by different federal agencies (e.g. ACS and IRS), there is typically a two year lag time to publication. For the time periods considered in this paper, these measures typically were in the low single digit range per state. For example, in 2016, the net migration rate varied from -1.5% in Alaska to +1.2% in Arizona. Search volumes that correspond to these three measures are also normalized by state population.

Pearson's correlation coefficients between these three core migration measures and their corresponding search volumes (normalized by state population) are shown in Table 2. These results suggest two themes. First, we observe generally high correlations between the official migration measures and the volume of migration queries. Among the three migration variables, migration search queries are most accurate at measuring inflow migration, with a mean correlation coefficient of 0.72 and standard deviation of 0.039 (μ = 0.72, σ^2 = 0.039). Figure 2 shows a scatterplot of inflow migration rate in 2015 and inflow migration queries in 2014, controlling for state population, with a strong linear relationship between migration queries and actual migrants. The correlation coefficients are lower for outflow (μ = 0.44, σ^2 = 0.14, per Table 2) and net migrations (μ = 0.39, σ^2 = 0.22, per Table 2). There are at least two possible explanations for these lower correlations. First, the origin states that are used to aggregate outflow migration are inferred by reverse-IP location, which might be less accurate than the destination states that are directly extracted from the query. Second, it is possible that the 2014 ACS outflow migration measure is an outlier, thus lowering the overall correlation. The outflow migration measure from ACS in 2014 have considerably lower autocorrelation with the same variables from 2015 ($r_{out14,out15}$ = 0.90) and from 2016 ($r_{out14,out16}$ = 0.79) compared to the equivalent autocorrelation for inflow migration ($r_{in14,in15}$ = 0.98, $r_{in14,in16}$ = 0.96). If only considering 2015 and 2016, our search queries present medium to strong correlations with the outflow migration (μ = 0.52, σ^2 = 0.06) and with net migration (μ = 0.53, σ^2 = 0.08).

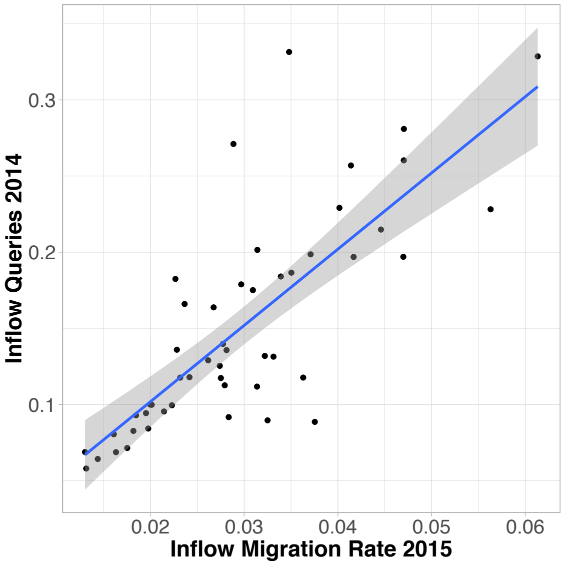

Figure 2: Incoming migration queries in 2014 and actual incoming migrants in 2015, both normalized by state population. Each point represents a U.S. state. Shaded area indicates 95% confidence interval of the linear regression.

The second theme from Table 2 is that search queries appear to have higher correlations with contemporaneous or future migration variables than with the past. Using inflow migration as an example, we group the correlation coefficients for inflow migrations in Table 2 by the number of years between the migration queries and the ACS statistics and compute the means for each group. Results in Figure 3 suggests that inflow search queries that precede the year of interest ("lead 2 years" and "lead 1 year") have higher correlations with the ACS inflow migration rates than the search queries that are contemporaneous ("same year") or follow the year of interest ("lag 1 year" and "lag 2 years"). We note that this leading effect of migration queries is consistent with previous findings on housing search queries, which demonstrates that web queries are usually good indicators of demand that will be eventually reflected in future consumer behavior [16, 68]. In the case of migration, it stands to reason that a typical prospective migrant would need to search

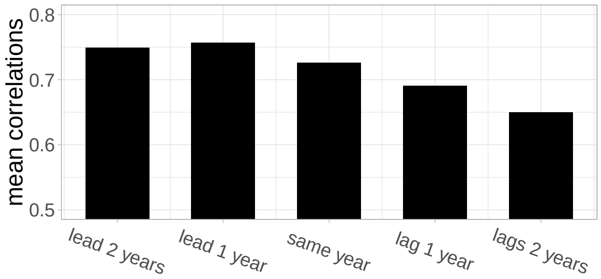

Figure 3: The leading effect of migration queries.

for a home or a job prior to actually moving. We provide further evidence for leading effect when demonstrating improvements to migration forecasting (see below).

Beyond the three primary migration measures, we also briefly highlight that the search query-based measures correlate strongly with secondary migration measures from government data. For example, ACS also collects data about the tenure of the migrants, which captures for each state the number of incoming migrants who currently live in owner occupied housing units and renter occupied housing units respectively. These tenure variables are very similar to the "Buy" and "Rent" variables we extracted (see Table 1). We thus compute the Pearson's correlation coefficients between the owner-renter ratio using ACS tenure data in 2016 and the buy-rent ratio using migration query data in 2015. Our result shows that these two variables indeed are strongly correlated ($\rho = 0.67$), further validating a query-based migration measurement.

4.2 Forecasting migrations statistics

The previous analyses demonstrated that migration queries alone can reasonably measure and even lead future migration variables. However, this doesn't necessarily indicate their incremental value in predicting future migration. In this section, we compare baseline migration prediction models with enhanced models that include search query measures, which is a typical methodology adopted by previous studies to demonstrate search queries' predictive power on economic indicators [3, 13, 16, 68]. Specifically, we follow these works to compare the goodness of fit (*adjusted* R^2) between the models with and without search query measures. We note that to be consistent with baseline models from the migration literature (e.g. [12, 61]), we construct models to predict *state level net migration.*

Our comparisons include two types of well-studied baseline models from migration literature. The first type of model is the autoregressive model. Because macroeconomics variables such as migration rates tend to be highly autocorrelated in time, simply forecasting future values from past data can provide good estimates. Traditionally known as the *Petersen-Greenwood hypothesis* in the migration literature [27, 53], this model treats current migration movement as a function of historical movement and has been shown to be effective in modeling state-to-state migration patterns [64]. We note that although more sophisticated nonlinear models are possible, we follow the practice common to these migration studies [27, 64] and use simple linear regression.

We first estimate the baseline model in Eq. 1 to predict 2016's ACS net migration rate using historical data from one year prior. This model is then compared with an enhanced model in Eq. 2 that adds search query variables from one year prior. Because domestic migration is a long decision making process, we hypothesize that adding query variables from multiple years might additionally enhance the prediction. As such, we also estimate an enhanced model in Eq. 3 that incorporates search query data from two years prior.

$$NetACS_{16} = \alpha + \beta_1 NetACS_{15} + \epsilon \tag{1}$$

$$NetACS_{16} = \alpha + \beta_1 NetACS_{15} + \beta_2 NetQueries_{15} + \epsilon \tag{2}$$

$$\begin{aligned} NetACS_{16} = \alpha{+}&\beta_1 NetACS_{15} + \beta_2 NetQueries_{15} \\ &+\beta_3 NetQueries_{14} + \epsilon \end{aligned} \tag{3}$$

We then estimate a stronger baseline model in Eq. 4 that includes the prior two years of ACS data. We note that we also explored models with longer histories of ACS migration data (e.g., an aggregate of the past five years) and the results were approximately the same or even worse in some cases. Similar to the previous comparison, we enhance this baseline by adding search variables of the prior one year (Eq. 5) and prior two years (Eq. 5) respectively.

$$NetACS_{16} = \alpha + \beta_1 NetACS_{15} + \beta_2 NetACS_{14} + \epsilon \tag{4}$$

$$\begin{aligned} NetACS_{16} = \alpha{+}&\beta_1 NetACS_{15} + \beta_2 NetACS_{14} \\ &+\beta_3 NetQueries_{15} + \epsilon \end{aligned} \tag{5}$$

$$\begin{aligned} NetACS_{16} = \alpha{+}&\beta_1 NetACS_{15} + \beta_2 NetACS_{14} \\ &+\beta_3 NetQueries_{15} + \beta_4 NetQueries_{14} + \epsilon \end{aligned} \tag{6}$$

The second type of baseline models we compare to is explanatory models that use socioeconomic factors to predict net migration. Among numerous such models (see comprehensive summaries in [11, 28]), we select the explanatory model proposed by Cebula and Alexander [12] as our baseline for the following reasons. First, it is designed and tested specifically to model state-level net migration rates. More importantly, it encompasses a comprehensive set of variables, including both traditional factors and novel "quality-of-life" factors [10], and has been shown to outperform models that include only economic factors [14, 56]. Lastly, as it was developed only a decade ago, the socioeconomic factors it considers should be the most pertinent to our context.

For model specification, we pick the best performing model from Cebular and Alexander [12] and include all the significant explanatory variables. We traced the source of these variables in the original model and found their most time-relevant counterparts [1]. Specifically, in Eq. 7, *MFI* is the median family income of each state; *COL* is the cost of living index for average four-person family; *EMPLGR* is the percent employment growth rate; *JANTEMP* is the daily maximum temperature in January; *HAZARD* is the percent distribution of hazardous waste sites of each state on the National Priority List; *TOXIC* is the toxic chemical releases per person; *EDPUP* is the government expenditures on elementary and secondary school, *STINCTAX* is the per capita state income tax burden. For brevity, we denote these socioeconomic variables with single notation *SOCIOECON* in ensuing model specifications and results interpretations. Readers should refer to the original paper [12] for detailed regression results for these variables. We enhance this baseline model by adding one year prior search variable (Eq. 8).

[1] We were unable find the exact counterpart data for a variable about the maximum January temperature, so we used the average winter temperature from NOAA.

	Dependent variable: $NetACS_{16}$									
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**
$NetACS_{15}$	0.641^{***}	0.493^{***}	0.571^{***}	0.518^{***}	0.312^{**}	0.386^{***}			0.389^{***}	0.310^{***}
	(0.098)	(0.099)	(0.108)	(0.103)	(0.099)	(0.106)			(0.081)	(0.085)
$NetACS_{14}$				0.148^{*}	0.185^{***}	0.180^{***}			0.136	0.184^{*}
				(0.056)	(0.048)	(0.047)			(0.079)	(0.078)
$NetQueries_{15}$		0.036^{**}	0.084^{**}		0.042^{***}	0.085^{**}		0.027^{**}		0.018^{*}
		(0.011)	(0.030)		(0.010)	(0.026)		(0.010)		(0.008)
$NetQueries_{14}$			-0.053			-0.05				
			(0.030)			(0.027)				
SOCIOECON							(Omitted)	(Omitted)	(Omitted)	(Omitted)
Obs. (States)	50	50	50	50	50	50	50	50	50	50
Adj. R^2	0.462	0.556	0.573	0.521	0.656	0.670	0.679	0.723	0.828	0.843

Note: Standard errors are shown in parentheses; SOCIOECON model coefficients are omitted for space considerations
Significance notation: $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 3: Linear regression to predict 2016 net migration rate

$$
\begin{aligned}
NetACS_{16} = \alpha &+ \beta_1 MFI + \beta_2 COL + \beta_3 EMPLGR \\
&+ \beta_4 WINTEMP + \beta_5 HAZARD + \beta_6 TOXIC \\
&+ \beta_7 EDPUP + \beta_8 STINCTAX + \epsilon
\end{aligned}
\tag{7}
$$

$$
NetACS_{16} = \alpha + \beta SOCIOECON + \beta_9 NetQueries_{15} + \epsilon \tag{8}
$$

Finally, we estimate a third class of baseline model that combines the explanatory model and the autoregressive model, shown in Eq. 9. We compare this to an enhanced model Eq. 10 that adds one year prior search query variables.

$$
\begin{aligned}
NetACS_{16} = \alpha &+ \beta SOCIOECON + \beta_9 NetACS_{15} \\
&+ \beta_{10} NetACS_{14} + \epsilon
\end{aligned}
\tag{9}
$$

$$
\begin{aligned}
NetACS_{16} = \alpha &+ \beta SOCIOECON + \beta_9 NetACS_{15} \\
&+ \beta_{10} NetACS_{14} + \beta_{11} NetQueries_{15} + \epsilon
\end{aligned}
\tag{10}
$$

Results in Table 3 show consistent improvement when adding search query variables to various baseline models. For the simplest autoregressive baseline (Eq. 1), adding one year prior search queries (Eq. 2) increased the adjusted R^2 from 0.462 to 0.556, a 20% of improvement. Likewise, for the stronger autoregressive baseline that includes the prior two years of historical data (Eq. 4), adding one year prior search queries (Eq. 5) increased the goodness of fit from 0.521 to 0.656, or a 26% of increase. We highlight that these improvements are considerably larger than those found in similar work that also demonstrate the power of search query data to predict macroeconomic phenomena (e.g. [68]). Including an additional prior year of search queries (Eq. 3 and Eq. 6) does increase the adjusted R^2. However, the estimated coefficients of $NetQueries_{14}$ are non-significant and in fact slightly negative, indicating a multicollinearity issue. Indeed, the VIF for $NetQueries_{14}$ and $NetQueries_{15}$ in both models are greater than 5, confirming the existence of multicollinearity [58]. As such, with building parsimonious models in mind, we drop $NetQueries_{14}$ in later models. We also note that dropping $NetQueries_{15}$ from Eq. 3 and Eq. 6 still provides models better than the baseline (Eq. 1 and 4), indicating that two year ahead query variables can still contribute to migration prediction.

Similarly, we observe improvement in adding search query variables to the SOCIOECON explanatory model. Comparing the results for Eq. 7 and Eq. 8, we observe that the adjusted R^2 increases from 0.679 to 0.723. Finally, for the model that encompasses both socioeconomic variables and historical data (Eq. 9 and Eq. 10), adding search queries still provides 0.02 increase to the model's Adjusted R^2, again an improvement comparable to similar work demonstrating the predictive power of search query data in the macroeconomic domain (e.g. [68]). For all the above models, we emphasize that all estimated coefficients of $NetQueries_{15}$ are positive and significant, showing non-zero effects of the search query variable.

In addition to predicting *NetACS16*, we also estimated models predicting inflow migration rate and outflow migration rate. Although inflow or outflow alone are easier to predict compared to net migration, we observe consistent, though smaller, improvements even when the baseline models are already very accurate. For instance, for a baseline model that predicts 2016 inflow migration rate using previous two years' history data, adding search query measures enhances the adjusted R^2 from 0.871 to 0.893, again, an increase that is comparable to related work.

Finally, we note that we experimented with different specifications for these models. For example, we tried adding contemporaneous *NetQueries* variables to all models, e.g. using $NetACS_{15}$, $NetQueries_{15}$ and $NetQueries_{16}$ to predict $NetACS_{16}$. We observed little to even negative changes to the adjusted R^2, which offers additional support for our model specification and for the previous finding about the leading effect of search queries.

Overall, the analyses in this section demonstrate that that migration queries consistently improve the prediction of core migration outcomes across several strong, theory-based baseline models.

4.3 Understanding Migration Demand

One additional advantage of our approach is that out filter extracts detailed metadata for each migration query, indicating whether the search is housing-related related or job-related, even the sub-types for housing and jobs. We thus investigate how the intent to migrate varies on two important dimensions. First, we study how migration demand (indicating the sum of all expressed intent to migrate, even if this expression goes unrealized) varies *demographically* by age, income and educational attainment. These attributes have been shown in migration studies to have differential impacts on domestic migration [37, 39, 57]. Second, we demonstrate that the type of migration demand (e.g., driven by housing versus jobs)

	Housing	Jobs	Housing-Rent	Housing-Buy
Age 18-24	-0.40	0.40	-0.06	-0.19
Age 25-44	-0.19	0.19	0.41	-0.18
Age 45-64	0.48	-0.48	-0.40	0.30
Age > 65	0.53	-0.53	-0.50	0.31

(a)

	Housing	Jobs	Jobs-Sci/Tech	Jobs-Mfg/Transp
< 25k	-0.17	0.17	-0.19	0.09
25k-50k	0.27	-0.27	0.17	0.06
50k-75k	0.23	-0.23	0.16	-0.42
> 75k	0.15	-0.15	0.20	-0.26

(b)

	Jobs-Sci/Tech	Jobs-Mfg/Transp
< High School	-0.23	0.37
High School	-0.32	0.35
Some College	-0.12	0.36
Bachelors	0.25	-0.44
Graduate	0.27	-0.43

(c)

Table 4: Correlation between migrants demographics and specific types of migration queries

varies *spatially*, even among states with similar net migration rates.

4.3.1 Migration Demand and Demographics. We hypothesize that the type of migration demand will differ based on varying demographics. Specifically, we focus on differential migration demand in relation to age, income, and education. These demographic attributes have previously been shown to affect related concepts, such as demand for employment, as measured by internet search queries (e.g. [13]). We obtained data for these demographic attributes from 2016 ACS provided through NHGIS [44]. Following our previous analyses, we compare these government data with search data from 2015, removing the District of Columbia. However, different from previous analyses which focused on net migration, here we study inflow migration for the following reasons. First, instead of prediction, this section focuses on explaining the relationship between where people are moving to and relevant demographics, and hence focusing on only one side of migration makes this relationship clearer. Second and more importantly, when segmenting total net migration into different groups, the margins of error of each group, as computed according to ACS's recommendation [25], can sometimes be larger than the net statistics, endangering the accuracy of our analysis. Furthermore, we chose to focus on inflow migration using traditional migration literature that studies the economic impact of international migration (e.g. [40, 42]) as guidance. We briefly discuss the outflow side at the end of the results, which present a generally similar trend.

First, we investigate how the types of migration demand, proxied by the types of migration queries, varies across age groups. To do so, we use ACS Table B07001 which captures for each state, the number of in-migrants for different age groups. Census publications uses a variety of age group breakdowns. To reduce the margins of error, we select the one with the fewest age groups, which is also used widely in decennial census publication (e.g. [31]). Motivated by the migration literature that argues that age plays an important role in housing-related migration (e.g. [33]), we study the relationship between age and housing-related migration searches, first in comparison to job related migration searches, and then within the buy versus rent housing query sub types. We normalize ACS age group data with the total number of inflow migrants, normalize housing and job queries by the total number of migration queries, and normalize rent and buy queries with the total number of housing queries, effectively turning every raw count into a proper percentage. We do so because we are most interested in understanding their relative relationships, rather than predicting their raw values.

Table 4 (a) presents the correlations between age groups and different types of migration demand. We note that because housing queries and jobs queries sum to 1 after our normalization, their correlation coefficients always have opposite signs and equal absolute values. Our results suggest that the percentage of older age migrants (e.g., age > 45) has positive correlations with the percentage of housing queries, whereas the percentage of younger migrants has positive correlations with the percentage of job queries to a state. In other words, states that attract a relatively large portion of housing queries in 2015, received large portion of older migrants in 2016 and states that drew job queries (or low housing queries) in 2015, pulled in more young migrants. These results are in line with prior migration theory which argues that young migrants are more motivated by employment and older migrants by housing [33]. To further unpack the relationships between age group and housing demand, the results in Table 4 (a) show that states that attract older age migrants in 2016 are positively associated with higher housing buy demand in 2015, whereas states that receive younger age group are positively associated with higher housing rent demand.

Next, we examine the relationship between income and migration demand. For income data, we again uses ACS Table B07010, that records the number of migrants for each income bracket. Similar to the previous analysis, we merge some income groups for ease of interpretation. Motivated by migration theories that demonstrate the interaction between migration, income and employment [17, 23], we investigated the relationship between income level and demand for employment related migration, with a particular focus on two job categories currently driving important economic disparities in the U.S.: science and technology versus manufacturing and transportation. We pre-process both ACS data and job related migration queries data with the same normalization procedure as in the above analysis on age.

Results in Table 4 (b) highlight two trends. First, as expected, our results indeed show that states with more higher income migrants are associated with more high tech job related inbound migration searches. Although the effect size is relatively small, these correlations shows distinctions when compared to the negative correlations between low income migrants and high tech job migration searches. For states that attract more low income migrants, our results did not suggest that they also attract large manufacturing and transportation jobs searches. Looking at this further, we did notice a trend that these same states saw increases in inflow job related migration searches in general, suggesting a possibility that

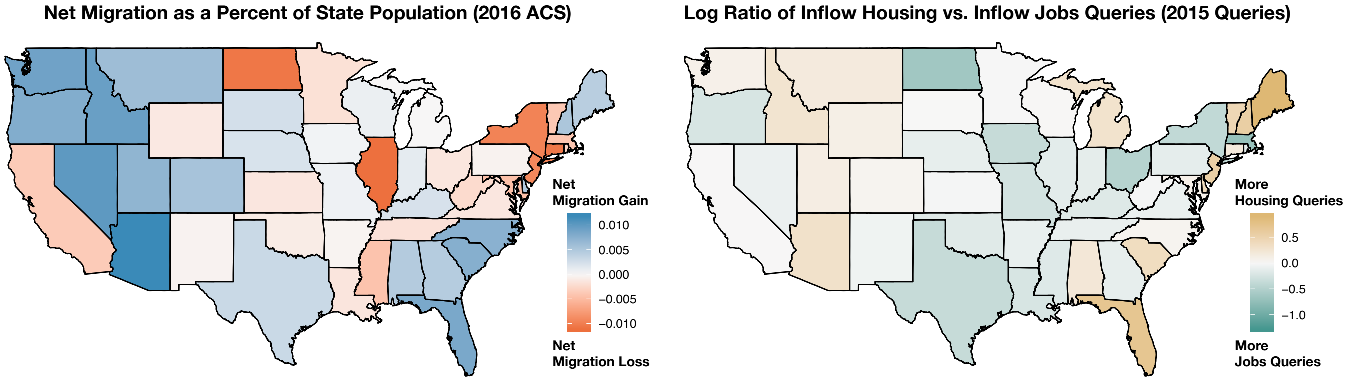
Net Migration as a Percent of State Population (2016 ACS)
Net Migration Gain
0.010
0.005
0.000
−0.005
−0.010
Net Migration Loss
Log Ratio of Inflow Housing vs. Inflow Jobs Queries (2015 Queries)
More Housing Queries
0.5
0.0
−0.5
−1.0
More Jobs Queries

Figure 4: Left: Net domestic migration per state from the 2016 American Community Survey, shown as a percentage of state population. Right: the ratio of a state's 2015 inbound housing migration queries relative to its incoming jobs queries, normalized by that same ratio over all queries, shown here in log scale.

these lower income migrants might be looking for jobs that do not requires special skills. Non-specific job related migration may be a direction for future exploration.

Finally, we explore the relationship between education level and migration demand. ACS data record the number of migrants in five different education attainment groups ranging from "Below High School" to "Graduate". Like the previous analysis, we correlate these categories with science and technology versus manufacturing and transportation employment migration searches based on prior research that looks at employment searches and demographics [13]. We applied the same data pre-processing before conducting the correlation analyses. Results in Table 4 (c) indicate that states with greater inflow migration queries in science and technology see more migrants with bachelor and graduate degrees. By contrast, states with more inflow migration queries for manufacturing and transportation jobs attract more migrants with low education attainment. These results demonstrate that the relationships between employment demands and demographics that are shown in previous research [13] also ring true in the migration context.

As previously mentioned, we replicated the above analyses using outflow job queries and outflow migrants demographics. We generally observed similar trends, although some effect sizes are smaller. We hypothesize that this might be due to the fact that on the inflow side, certain states serve as the "magnets" that attract particular demographics, while such an effect is weaker on the outflow side.

4.3.2 Migration Demand and Geography. Domestic net migration is a core migration statistic that always draws extensive commentary and interpretation from columnists, policy makers and scholars upon its publication each year [1, 41, 63]. While much of this discussion centers around which states are growing and shrinking the fastest, raw migration statistics are not, on their own, able to contextualize the economic impact to local housing and job markets.

The left map in Figure 4 visualizes net domestic migration in the 48 contiguous United States using data from ACS 2016. In this figure, we see that states such as Florida (+0.83%), Texas (+0.31%), Arizona (+1.17%) and Nevada (+1.00%) are experiencing net population growth from domestic migration, while states such as New York (-0.96%), California (-0.36%), North Dakota (-1.06%), and Illinois (-1.11%) are losing more people than they are attracting.

Our data allow us to characterize the flow of migration queries into (or out of) a state as to whether the queries are disproportionately about housing or about jobs. Following the previous analysis, we focus only on an analysis of inflow; the analysis of outflow, which may help explain those states in deficit, follows analogously.

First, for each state s, we compute the proportion of incoming housing queries s receives relative to the housing queries in all other states:

$$\rho_{\text{in}}^{\text{housing}}(s) = \frac{\text{total inbound housing queries to } s}{\sum_{s'} (\text{total inbound housing queries to } s')}.$$

Similarly, we compute the proportion of incoming jobs queries:

$$\rho_{\text{in}}^{\text{jobs}}(s) = \frac{\text{total inbound jobs queries to } s}{\sum_{s'} (\text{total inbound jobs queries to } s')}.$$

All things equal, the ratio of job and housing related migration searches would be similar across states, however in practice, $\rho_{\text{in}}^{\text{housing}}(s)$ and $\rho_{\text{in}}^{\text{jobs}}(s)$ can be very different. The search behaviors that underlie each index are influenced by numerous factors, including the competitiveness of the local housing and job markets, and demographic differences among the migrants themselves, such as whether they are early in their career, or whether they are retired.

In the right side of Figure 4, we plot $\log\left(\rho_{\text{in}}^{\text{housing}}(s) \middle/ \rho_{\text{in}}^{\text{jobs}}(s)\right)$ for each state s using queries from 2015, highlighting which states attract a disproportionate fraction of housing migration searches relative to the proportion of job migration searches they attract, and vice versa. For example, while Florida (+0.83%) and Texas (+0.31%) are both net-positive in domestic migration, their inbound migration queries have very different profiles. Florida attracts disproportionately more housing searches (+0.70 in log ratio), indicating a stronger demand for its local housing market perhaps because of its status as a retirement destination, while Texas attracts more incoming job queries (-0.36 in log ratio), a possible reflection of a stronger demand for its labor market. Similarly, while neighboring states Nevada (+1.00%) and Arizona (+1.17%) are both domestic migration winners, Nevada attracts more job queries (-0.09 in log ratio) and Arizona attracts more housing queries (+0.30 in log ratio).

5 DISCUSSION

Forecasting and understanding large-scale migration patterns has long been a key problem of interest both for academics seeking to better understand the forces that shape society, and for government officials seeking to attract and retain more people to their regions. The impact of migration trends can be massive, with "winning" regions attracting a significant influx of skilled workers, and “losing" regions seeing their talent pool drift away to other shores.

In this work, we explore the use of internet search engine queries as a signal for predicting and understanding domestic migration trends that can augment and improve on government data-based models. The correlations with government data (Table 2, Figures 2, 3) indicate that the search query-based migration intent signals correlate with, and in fact even slightly lead current gold standard migration metrics. This leading relationship is most evident in the net migration flow correlations, in which the correlations between the search query metric and net migration jumps from about 0.10 when contemporaneous to 0.50 to 0.60 when leading official migration statistics. Thus, internet searching appears to be a precursor to migration, which is sensible given that people are likely to research possible migration destinations prior to moving.

The fact that queries tend to lead official metrics motivates our exploration of using search query data to *forecast* migration. Indeed we demonstrate that a migration query measure adds variance explained to regression models that predict future migration. Our model adds more than 10% variance explained over an autoregressive baseline to predictions of *net* migration, which is much harder to forecast than inflow or outflow migration alone. While the boost in variance explained in column 8 of Table 3 is more modest over the more robust baseline shown in column 7, bear in mind that this baseline incorporates many socioeconomic and environmental variables, such as median income and school expenditures, most of which are only available at annual timescales at best. The search query measure still adds variance explained over all of these individual measures, and is available in near real time. Indeed, since government migration data tend to be released two years after the period in question, the fact that search queries can provide a real-time, leading indicator of migration trends is testament to the potential of this method, which provides insights years in advance of any official government tally. Even once government data are released, combining such data with our data provides an even stronger model for predicting subsequent years' migration trends.

Beyond prediction, as shown in Table 4, we illustrate how migration queries shed light on some of the more important questions surrounding aggregated demand for migration. Perhaps the most fundamental question is whether people move for housing or for jobs. Using our migration query taxonomy, we aligned housing and job related migration queries with known demographics of migrants from the ACS, suggesting that younger people and people making the lowest income are more likely to migrate for employment reasons, while conversely older and wealthier people migrate for housing. Furthermore, among people who are searching for housing, older migrants are more likely to be looking to buy property, compared to their younger counterparts who are more likely to be renting. Similar relationships can be seen in looking at the data through he lens of geography. Figure 4 reveals that in states like Florida and Arizona, which attract a considerable number of retirees, incoming migrants are much more likely to be searching for housing than for jobs.

Regarding migration for employment, we highlight how query data can shed light on arguably the most pressing issue of employment today: high-tech versus manual labor jobs. We see a clear split between the less wealthy and less educated versus the wealthier and more highly educated. States attracting migrants in the lowest income bracket are the least likely to see job migration queries in the science and technology sector, while those with the higher income brackets are less likely to see migration queries in the manufacturing and transportation sectors. Education levels reveal a similar story, with science and technology migration queries aligning with high education states, and manufacturing and transportation migration queries aligning with lower education states.

These relationships demonstrate that search query data could be used to forecast the types of housing and employment incoming migrants to a state will seek, which could be extremely valuable for state and even city level planning. For instance, governments through permitting processes and developers could better plan for not only the overall increase or decrease in demand for housing, but whether that housing should be for rent or purchase. These forecasts simply are not feasible with government data, outside of autoregressive approaches that forecast based simply on past trends. Similarly, local governments and business leaders can benefit from forecasting incoming employees across industry sectors, again something not available as leading indicators in government data.

This then, indicates usage scenarios for government planning and policy agencies. At 10% of the population per year, churn in population for any given state is significant, and can drastically impact the social and economic development of a region. States will want to plan and draft housing and economic policies proactively in order to strategically encourage growth in desirable industries, to make housing stock as responsive to changes in demand, and so forth. Predictive policy and economic growth is far from trivial, but at the very least the relevant organizations could be using data that, like we have shown with these search query-based metrics, is at sufficient scale, is leading, and is nuanced in explanation.

6 CONCLUSION

In this paper, we proposed a novel approach that utilizes search query data to forecast and understand domestic migration trends in the United States. We developed a fast and robust migration query filter that achieves 0.88 F1 score. Our migration search data were validated against government data and showed high correlations. More importantly, we demonstrated that adding our migration search data could enhance two types of migration prediction models. Details of our search queries also offer nuanced explanations to how migrants' demands differ demographically and spatially. Overall, our approach will significantly benefit computational social scientists, migration researchers and local policy makers.

REFERENCES

[1] [n. d.]. California Losing Residents Via Domestic Migration [EconTax Blog]. ([n. d.]). https://lao.ca.gov/laoecontax/article/detail/265

[2] [n. d.]. LinkedIn Workforce Report | United States | August 2018. ([n. d.]). https://economicgraph.linkedin.com/resources/linkedin-workforce-report-august-2018

[3] Nikolaos Askitas and Klaus F Zimmermann. 2009. Google econometrics and unemployment forecasting. *Applied Economics Quarterly* 55, 2 (2009), 107–120.

[4] René Böheim and Mark P Taylor. 2002. Tied down or room to move? Investigating the relationships between housing tenure, employment status and residential mobility in Britain. *Scottish Journal of Political Economy* 49, 4 (2002), 369–392.

[5] Ilaria Bordino, Stefano Battiston, Guido Caldarelli, Matthieu Cristelli, Antti Ukkonen, and Ingmar Weber. 2012. Web search queries can predict stock market volumes. *PloS one* 7, 7 (2012), e40014.

[6] Andrei Broder. 2002. A taxonomy of web search. In *ACM Sigir forum*, Vol. 36. ACM, 3–10.

[7] Davide Buscaldi. 2011. Approaches to disambiguating toponyms. *Sigspatial Special* 3, 2 (2011), 16–19.

[8] Davide Buscaldi and Paolo Rosso. 2008. Map-based vs. knowledge-based toponym disambiguation. In *Proceedings of the 5th Workshop on Geographic Information Retrieval*. ACM, 19–22.

[9] Gavin Cameron and John Muellbauer. 1998. The housing market and regional commuting and migration choices. *Scottish Journal of Political Economy* 45, 4 (1998), 420–446.

[10] RJ Cebula and JE Payne. 2005. Net migration, economic opportunity and the quality of life, 1999-2000. *International Review of Economics and Business* 52, 2 (2005), 245–254.

[11] Richard J Cebula. 1979. *The determinants of human migration*. Lexington Books.

[12] Richard J Cebula, Gigi M Alexander, et al. 2006. Determinants of net interstate migration, 2000-2004. *Journal of Regional Analysis and Policy* 36, 2 (2006), 116–123.

[13] Stevie Chancellor and Scott Counts. 2018. Measuring Employment Demand Using Internet Search Data. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 122.

[14] Alberta H Charney. 1993. Migration and the public sector: a survey. *Regional Studies* 27, 4 (1993), 313–326.

[15] Yong Chen and Stuart S Rosenthal. 2008. Local amenities and life-cycle migration: Do people move for jobs or fun? *Journal of Urban Economics* 64, 3 (2008), 519–537.

[16] Hyunyoung Choi and Hal Varian. 2012. Predicting the present with Google Trends. *Economic Record* 88, s1 (2012), 2–9.

[17] Thomas J Courchene. 1974. *Migration, income, and employment: Canada, 1965-68*. CD Howe Research Institute.

[18] Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

[19] Gordon B Dahl. 2002. Mobility and the return to education: Testing a Roy model with multiple markets. *Econometrica* 70, 6 (2002), 2367–2420.

[20] Francesco DâĂŹAmuri and Juri Marcucci. 2010. 'Google it!'Forecasting the US unemployment rate with a Google job search index. (2010).

[21] J Paul Elhorst. 2010. Applied spatial econometrics: raising the bar. *Spatial Economic Analysis* 5, 1 (2010), 9–28.

[22] Michael Ettredge, John Gerdes, and Gilbert Karuga. 2005. Using web-based search data to predict macroeconomic statistics. *Commun. ACM* 48, 11 (2005), 87–92.

[23] Riccardo Faini, Alessandra Venturini, et al. 1994. *Migration and growth: the experience of Southern Europe*. Technical Report. CEPR Discussion Papers.

[24] Lee Fiorio, Guy Abel, Jixuan Cai, Emilio Zagheni, Ingmar Weber, and Guillermo Vinué. 2017. Using twitter data to estimate the relationship between short-term mobility and long-term migration. In *Proceedings of the 2017 ACM on Web Science Conference*. ACM, 103–110.

[25] Sirius Fuller. 2018. Using American Community Survey Estimates and Margins of Error. (April 2018). https://www.census.gov/content/dam/Census/programs-surveys/acs/guidance/training-presentations/20180418_MOE.pdf

[26] US Census Bureau Geography. [n. d.]. 2010 Geographic Terms and Concepts - Place. ([n. d.]). https://www.census.gov/geo/reference/gtc/gtc_place.html

[27] Michael J Greenwood. 1969. An analysis of the determinants of geographic labor mobility in the United States. *The review of Economics and Statistics* (1969), 189–194.

[28] Michael J Greenwood. 1975. Research on internal migration in the United States: a survey. *Journal of Economic Literature* (1975), 397–433.

[29] Emily Gross. 2005. Internal revenue service area-to-area migration data: Strengths, limitations, and current trends. In *Proceedings of the Section on Government Statistics*. 2005.

[30] Andrew Henley. 1998. Residential mobility, housing equity and the labour market. *The Economic Journal* 108, 447 (1998), 414–427.

[31] Lindsay M Howden and Julie A Meyer. 2010. Age and sex composition: 2010. *2010 Census Briefs, US Department of Commerce, Economics and Statistics Administration. US CENSUS BUREAU* (2010).

[32] David Ihrke. [n. d.]. Why Did You Move?: An Overview and Analysis of the Annual Social and Economic SupplementâĂŹs Reason for Move Write-In Expansion. SEHSD-WP2016-22 ([n. d.]). https://www.census.gov/library/working-papers/2016/demo/SEHSD-WP2016-22.html

[33] David Ihrke. 2014. Reason for Moving: 2012 to 2013. *Current population reports. Washington, DC: US Census Bureau* 20 (2014), 574.

[34] David Ihrke. 2017. United States mover rate at a new record low. *Census Blogs. Available at: https://www. census. gov/newsroom/blogs/randomsamplings/2017/01/mover-rate. html* (2017).

[35] Richard Jackman and Savvas Savouri. 1992. Regional migration versus regional commuting: the identification of housing and employment flows. *Scottish Journal of Political Economy* 39, 3 (1992), 272–287.

[36] Geraint Johnes and Thomas Hyclak. 1994. House prices, migration, and regional labor markets. *Journal of Housing Economics* 3, 4 (1994), 312–329.

[37] Kenneth M Johnson, Paul R Voss, Roger B Hammer, Glenn V Fuguitt, and Scott McNiven. 2005. Temporal and spatial variation in age-specific net migration in the United States. *Demography* 42, 4 (2005), 791–812.

[38] John D Kasarda. 1988. Jobs, migration, and emerging urban mismatches. *Urban change and poverty* (1988), 148–198.

[39] John Kennan and James R Walker. 2011. The effect of expected income on individual migration decisions. *Econometrica* 79, 1 (2011), 211–251.

[40] Sari Pekkala Kerr and William R Kerr. 2011. *Economic impacts of immigration: A survey*. Technical Report. National Bureau of Economic Research.

[41] Andy Kiersz. [n. d.]. Here's how each US state's population changed between 2016 and 2017 because of people moving in and out. ([n. d.]). https://www.businessinsider.com/state-domestic-migration-map-2016-to-2017-2018-1

[42] Robert J LaLonde and Robert H Topel. 1997. Economic impact of international migration and the economic performance of migrants. *Handbook of population and family economics* 1 (1997), 799–850.

[43] Pierre Legendre. 1993. Spatial autocorrelation: trouble or new paradigm? *Ecology* 74, 6 (1993), 1659–1673.

[44] Steven Manson, Jonathan Schroeder, David Van Riper, and Steven Ruggles. 2017. IPUMS National Historical Geographic Information System: Version 12.0 [Database]. *Minneapolis: University of Minnesota* (2017).

[45] Warren F Mazek and John Chang. 1972. The chicken or egg fowl-up in migration: comment. *Southern Economic Journal* (1972), 133–139.

[46] Nick McLaren and Rachana Shanbhogue. 2011. Using internet search data as economic indicators. (2011).

[47] Johnnatan Messias, Fabricio Benevenuto, Ingmar Weber, and Emilio Zagheni. 2016. From migration corridors to clusters: The value of Google+ data for migration studies. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE Press, 421–428.

[48] Raven Molloy, Christopher L Smith, and Abigail Wozniak. 2011. Internal migration in the United States. *Journal of Economic perspectives* 25, 3 (2011), 173–96.

[49] Richard F Muth. 1971. Migration: chicken or egg? *Southern Economic Journal* (1971), 295–306.

[50] 2017-2018 Session of the Vermont State Senate. 2017. S.94, Act 257. An act relating to promoting remote work. (2017). http://www.leg.state.vt.us/jfo/fiscal_notes/2018_S_94_fiscal_note_house_commerce.pdf

[51] Adam Pasick. [n. d.]. Facebook's big data glimpse at human migration and the growth of mega-cities. ([n. d.]). https://qz.com/159279/facebooks-big-data-glimpse-at-human-migration-and-the-growth-of-mega-cities/

[52] Jaroslav Pavlicek and Ladislav Kristoufek. 2015. Nowcasting unemployment rates with google searches: Evidence from the visegrad group countries. *PloS one* 10, 5 (2015), e0127084.

[53] William Petersen. 1969. *Population*. London: MacMillan Co.

[54] Tobias Preis, Helen Susannah Moat, and H Eugene Stanley. 2013. Quantifying trading behavior in financial markets using Google Trends. *Scientific reports* 3 (2013), 1684.

[55] Tobias Preis, Daniel Reith, and H Eugene Stanley. 2010. Complex dynamics of our economic life on different scales: insights from search engine query data. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 368, 1933 (2010), 5707–5719.

[56] Stephen M Renas and Rishi Kumar. 1978. The cost of living, labor market opportunities, and the migration decision: A case of misspecification? *The Annals of regional science* 12, 2 (1978), 95–104.

[57] Aba Schwartz. 1976. Migration, age, and education. *Journal of Political Economy* 84, 4, Part 1 (1976), 701–719.

[58] Simon Sheather. 2009. *A modern approach to regression with R*. Springer Science & Business Media.

[59] Robert Shimer. 2007. Mismatch. *American Economic Review* 97, 4 (2007), 1074–1101.

[60] David A Smith and Gregory Crane. 2001. Disambiguating geographic names in a historical digital library. In *International Conference on Theory and Practice of Digital Libraries*. Springer, 127–136.

[61] Paul M Sommers and Daniel B Suits. 1973. Analysis of net interstate migration. *Southern Economic Journal* (1973), 193–201.

[62] Donald N Steinnes. 1978. Causality and Migration: A Statistical Resolution of the" Chicken or Egg Fowl-up". *Southern Economic Journal* (1978), 218–226.

[63] Lyman Stone. 2017. Reviewing the 2017 Census Population Estimates. (Dec. 2017). https://medium.com/migration-issues/reviewing-the-2017-census-population-estimates-62a8a9ad5d17

[64] James D Tarver and R Douglas McLeod. 1973. A test and modification of ZipfâĂŹs hypothesis for predicting interstate migration. *Demography* 10, 2 (1973), 259–275.

[65] Jacob Thebault-Spieker, Loren Terveen, and Brent Hecht. 2017. Toward a geographic understanding of the sharing economy: Systemic biases in UberX and TaskRabbit. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 3 (2017), 21.

[66] Antonio Toral and Rafael Munoz. 2006. A proposal to automatically build and maintain gazetteers for Named Entity Recognition by using Wikipedia. In *Proceedings of the Workshop on NEW TEXT Wikis and blogs and other dynamic text sources.*

[67] Ingmar Weber, Emilio Zagheni, et al. 2013. Studying inter-national mobility through IP geolocation. In *Proceedings of the sixth ACM international conference on Web search and data mining.* ACM, 265–274.

[68] Lynn Wu and Erik Brynjolfsson. 2015. The future of prediction: How Google searches foreshadow housing prices and sales. In *Economic analysis of the digital economy.* University of Chicago Press, 89–118.

[69] Jeffrey E Zabel. 2012. Migration, housing market, and labor market responses to employment shocks. *Journal of Urban Economics* 72, 2-3 (2012), 267–284.

[70] Emilio Zagheni, Venkata Rama Kiran Garimella, Ingmar Weber, et al. 2014. Inferring international and internal migration patterns from twitter data. In *Proceedings of the 23rd International Conference on World Wide Web.* ACM, 439–444.

[71] Emilio Zagheni and Ingmar Weber. 2012. You are where you e-mail: using e-mail data to estimate international migration rates. In *Proceedings of the 4th annual ACM web science conference.* ACM, 348–351.