(Machine) Learning About Immigration's Impact in Local Labor Markets With Classified Ad Text

L. Jason Anastasopoulos George Borjas Gavin Cook
Michael Lachanski

March 20, 2019

Abstract

The decennial Census provides detailed information about local labor markets but at very low-frequencies, making it difficult for researchers to identify short-run effects in local labor markets. Machine learning techniques open the door to studies analyzing local labor markets at high frequency by allowing researchers to analyze large text corpora, such as those found in classified ads. Motivated by Census tabulations showing a Cuban immigration surge in the Miami metro area around 1980, we document a decline in the Help Wanted Index, a monthly count of help-wanted classifieds, for Miami relative to geographic baselines. We conduct a supervised learning exercise demonstrating a statistically significant decline in the fraction of job openings for roles typically associated with less-educated workers using a 1/29 sample of the *Miami Herald*'s classifieds. We explore these results using topic modeling, which naturally generates occupational and industry categories with minimal researcher degrees of freedom. This exercise uncovers a statistically significant increase in the proportion of job openings advertised for positions in accounting, sales, and engineering and a corresponding decrease in the proportion advertised for positions in the food and automotive industry. Analysis of the Current Population Survey's Annual Social and Economic Supplement confirms the relative importance of job openings for less-educated workers' wage outcomes, but the linkage with employment outcomes is statistically indistinguishable from zero for all workers. We conclude with directions for future research and suggestions for agencies aiming to modernize their statistical releases.[1]

[1]Title XIII and XXVI statistics approved for release by the Disclosure Review Board (DRB) under Approval Number CBDRB-FY19-190.

1 Introduction

Classified advertisements provide a rich and mostly untapped source of information about labor markets. While traditional data sources such as the Current Population Survey (CPS) and decennial Census provide us with overviews of labor markets and employment trends at the national and state level, high resolution data about local labor markets, particularly in smaller metropolitan areas, has not been available to researchers at high frequency. In recent years however, machine learning techniques such as topic models, which allow researchers to extract relevant economic concepts from text data, open the door to studies comparing local labor markets.[2] In principle, the techniques we utilize can be used to study a wide variety of policy shocks and economic events like minimum wages increases, natural disasters, and trade shocks. We expect future policy changes to result in increased legal immigration to the United States by 2050 (reflecting the current bipartisan consensus).[3] Thus, to estimate the expected impact of these future policy shocks, we apply our methods to a well-known refugee shock. In particular, we study the 1980 Mariel Boatlift, with the intuition that the changing labor market conditions we identify can be attributed to the refugee shock, and thus prove useful for policymakers and academics trying to anticipate likely impacts of future increases in immigration.

The outline of the essay is as follows. In the next section, we discuss the *Help*

[2]This working paper was made possible by the US 2050 project, supported by the Peter G. Peterson Foundation and the Ford Foundation. The statements made and views expressed are solely the responsibility of the author. The authors would like to thank Jonathan Rothbaum, and Rebecca Chenevert for helpful comments.

[3]For instance, consider the most recent State of the Union address. https://www.politico.com/story/2019/02/05/trump-state-of-the-union-legal-immigration-1148629

Wanted Index (HWI), the longest-running job vacancy proxy for the United States. Next, we use Census tabulations from the 2000 Decennial Long Form to show both a spike in Cuban refugees to the United States in 1980 and tax filings to show that their primary economic impact was in the Miami MSA. We begin our analysis by estimating a supervised learning algorithm using the O*NET classifications to divide job advertisements into those targeting the less-educated and those targeting the more-educated. We then describe how unsupervised machine learning techniques can be used to extract relevant information from the text of classifieds and formally analyze a sample of classifieds from the *Miami Herald* before and after the Mariel Boatlift using topic models. The last analysis we conduct uses traditional data sources like the Bureau of Labor Statistics Job Openings and Labor Turnover Survey (JOLTS) and the Current Population Survey's Annual Social and Economic Supplement (CPS-ASEC) to understand what wage and employment consequences are associated with changes in job openings. We conclude and make suggestions for future research.

2 American Help-Wanted Advertisements in the Twentieth Century

The HWI is a commonly-used proxy of unfilled job vacancies and, by extension, slack in the labor market.[4] The Conference Board contacted 51 newspapers, each corresponding to a different metropolitan statistical area (MSA), and recorded the number of classified ads placed each month in each paper. This number was adjusted

[4]Job vacancies are usually thought of as a stock analogous with unemployment.

for seasonality and day-of-the-week bias to create a monthly index. The raw counts have not to our knowledge ever been made available, and so we work with the HWI throughout.

2.1 History

The US government did not maintain any continuous official statistics on job openings until the year 2000, when the Bureau of Labor Statistics (BLS) introduced the Job Openings and Labor Turnover Survey (JOLTS).[5] The first edition of the HWI was launched in 1964 and combined data from previous surveys to create a monthly time series stretching to 1951. Zagorsky (1998) combined previous surveys of help wanted classifieds by the Metropolitan Life Insurance Company with the HWI to create a help wanted index ranging from 1923 to 1994. Apart from the removal of the *Newark Evening News* in 1971 and a swap of the *Dallas Times Herald News* for the *Dallas Morning News* in the early 1990s, the list of papers and total number of cities surveyed has not changed since 1970.[6] This enables far-reaching historical uses of the data.

The Conference Board introduced a measure of job advertising online called the Help-Wanted Online (HWOL) to track online job ads in 2005. Despite Barnichon (2010) showing that a measure combining the HWI, and HWOL into a single job vacancy measure outperforms either measure individually from the mid-1990s to the late 2000s, the Conference Board ceased releasing the HWI to the public in July

[5]Prior to 2000, researchers interested in American job openings or vacancies relied on the Conference Board's HWI or, secondarily, on state-level job vacancy data.

[6]See Table 1 in the Appendix.

2008.[7]

2.2 Early Debates

The HWI has seen use in a number of widely-cited articles in leading journals of economics. Five articles from *The Review of Economics and Statistics* serve as early examples of academic writing on the HWI. Cohen and Solow (1967) note that "economists have used the National Industrial Conference Board index of help-wanted advertising as an indirect measure of demand pressure in labor markets" "in the absence of direct statistics on unfilled vacancies". They note that there is no natural way to utilize the HWI in a macroeconometric framework. Consequently, they normalize it by accounting for the growth of the national economy so they can estimate a relationship between the HWI and unemployment. Burch and Fabricant (1968) notes that "the sharp increase in the (HWI) in the second half of the 1965 has evoked considerable interest" because it showed a "divergence between the HWI and the unemployment rate". Both conclude the relationship is stable and inverse. Gujarati (1969) challenges this assertion by estimating the HWI-unemployment relationship in each phase of business cycle from 1951 to 1968. He finds that the relationship between HWI and unemployment is always inverse, statistically significant, and economically significant, but unstable. In particular, coefficients appear statistically difference not only by phases of the business cycle but even between different expansions. In a rejoinder, Cohen and Solow (1970) notes that they worked

[7]The Conference Board stopped internal data collection for the index in October 2010. The HWOL is now the only time series tracking contemporary help-wanted advertising volume from the Conference Board.

with first differences in the HWI in their original article while Burch and Fabricant (1968) and Gujarati (1969) worked with HWI levels. When first-differences are used, the HWI-unemployment relationship appears reasonably stable over the two decades of data available at that time.[8] In addition to showing that the HWI was of academic interest to economists in their own right, this debate shows that researchers must think carefully about how they utilize the HWI which has no natural units.[9]

Abraham and Wachter (1987) represents one of the most comprehensive summaries of the benefits of, issues with, and changes in, the HWI data. As a check against the HWI, she compares the index's data for Minnesota against state-collected job opening numbers numbers.[10] Her comparisons lead her to conclude: "the Minnesota data...suggest that the normalized help-wanted index is a reasonably good vacancy proxy".[11]

Abraham also records a few shifts in the content of advertisements. Abraham notes that the HWI has drifted upward since 1970 and posits that "it is likely that EEO and affirmative action concerns have caused at least some increase in the volume of help-wanted advertising". The number of employers advertising white-collar jobs

[8] In the debate's last entry, Burch and Fabricant (1971) notes that Gujarati (1969) used his own scheme for assigning dates to expansions and contractions rather than official NBER dating. When official NBER recession dating is used, even the HWI level and unemployment relationship appears stable conditional on adding one structural shift variable in 1957 and whether or not one is in an expansion (recession).

[9] Taking up where Cohen and Solow (1970) left off, throughout we work with log-differences and percent differences of a normalization of the HWI, and carefully consider business cycle factors wherever appropriate.

[10] Minnesota was one of the only states to collect job openings data and release it for public use.

[11] She also observes that "declining competition in the newspaper industry" may also help drive this shift because "employers may have become more likely to advertise any particular job opening in the surviving papers". The total number of papers in the MSAs covered by the HWI dropped "from 148 in 1960 to 87 in 1985", and the weighted mean share of circulation commanded by the papers rose "from 60 percent in 1960 to 80 percent in 1985".

for college graduates also increased during the same period in which the upward drift occurred. This indicates a shift to white-collar jobs in classified advertising, which mirrors employment trends at large, but Abraham's data suggests that there is actually an over-representation of white-collar jobs in classifieds (Table 3, Abraham and Wachter (1987)). James Tobin's comment at the end of the paper hints at the impact of this shift. He "questioned whether positions advertised in newspapers all represent genuine vacancies" and cited an article in Fortune that found that "many of the positions advertised required specific skills that few individuals would possess" and "did not represent jobs relevant to the unemployed". This follows from Abraham's intuition that "EEO and affirmative action concerns" may have prompted employers to preemptively advertise openings to defuse potential claims of discrimination.[12]

2.3 Recent Debates

For the period of time our work covers, the HWI was the gold standard for labor market data. Substantiating this, Berman (1997) on labor matching in Israel uses a dataset which explicitly compares to the HWI.[13] Abraham and Katz (1986) use the HWI to show that structural changes do not appear to drive cyclical decreases in employment. Andolfatto (1996) shows that modeling the job search process ex-

[12]At the time of writing, Abraham did not have "direct evidence on the influence of EEO and affirmative action pressures" on advertising or hiring practicing. We anticipate being able to answer this question in future work.

[13]Some recent scholarship has focused on the waning relevancy of the HWI (Kroft and Pope, 2014), but many scholars continue to use the HWI for both historical research (Lee, 2016) and as a proxy for job vacancies in macroeconometric models that require long vacancy time series to estimate (Pater, 2017).

plicitly can help real business cycle theories of macroeconomic activity better match U.S. time series. On the other hand, Shimer (2005) used the HWI to construct a time-series of historical data and finds that both JOLTS and the HWI do not vary enough during recessions according to standard job market matching specifications, launching a debate on the relevance of job market matching to quantitative business cycle activity. These recent debates are not particularly germane[14] to our research, which attempts to use the content of classified ads to conduct careful case studies on specific economic events and policy shocks.

3 Even More Data on the Mariel Boatlift

The 1980 Mariel boatlift is the largest American refugee shock which exists contemporaneously with modern economic statistics. In Table 2, we tabulate the number of Cuban foreign-born entering the U.S. in each year according to the 2000 US Census Decennial Long Form.[15] In other words, the 2000 U.S. Decennial Long Form largely replicates the findings derived from public use microdata as in Card (1990) and Borjas (2017) on the size of the Mariel supply-shock. Between 120,000 and 130,000 Cubans, the majority of whomst were high-school dropouts, entered the

[14] In fact, there may be a way to connect our comparative case study method using classifieds to the macroeconomic consequences of job search. Newspaper classifieds were a job search technology especially relevant to less-educated workers by all accounts. If newspaper strikes and shutdowns, which are often plausibly exogenous to local economic conditions, have no impact on local area outcomes for these workers or high-turnover occupation-industry pairs, then it stands to reason that job search may not be very important for understanding aggregate outcomes like unemployment at business-cycle frequencies as suggested in Shimer (2007).

[15] Noise has been added to the estimates according to the interim 2018 rules for privacy protection, and the figure has *not* been mortality-adjusted, but we hope that this new data enables researchers new and old to grasp the relative sizes of the 1980 and post-1995 Cuban refugee shocks.

United States and primarily settled in the Miami metropolitan area. Because of immigrant labor market downgrading, many high school and college graduates worked in occupations usually taken by high school dropouts, making the effective labor supply shock for less-educated workers even larger. Unsurprisingly, this refugee shock's impacts on housing and rental prices (Saiz (2003)), fertility (Seah (2018)), workers' wages (Yasenov and Peri (2018)), and local product demand (Bodvarsson, Van den Berg, and Lewerd (2008)) have been carefully studied.

3.1 Administrative Data Meets the Mariel Boatlift

Rather than retreading old ground, we bring new administrative data to this question. In particular, we use the U.S. 2000 Decennial linked for selected IRS 1040 Filing years 1970, 1975, 1980, 1985, 1990, and 1995 and track likely Marielitos filing locations through time in Table 3.[16] Two important implications emerge. First, a small but statistically significant (relative to zero) number of Marielitos filed taxes before their stated year of entry to the United States. This demonstrates three potential sources of error. First, they may have incorrectly stated their year of entry to the United States on the Decennial 2000 Long Form, which may have been interpreted as their full year in the country, first citizenship, or first full year of citizenship.[17] A second source of error lies in the tax filings. For instance, the name or address on an

[16] A likely Marielito is someone who defines themselves as having been born in Cuba and entered the United States in 1980 in the U.S. 2000 Decennial Long Form. There were Cubans who arrived to the United States in 1980 that were not part of the Mariel boatlift, but as the 1979 and 1981 number of entrants in Table 2 suggests, they were a very small fraction of the total.

[17] There may also have been tabulation errors or incorrect transcription of the year from the paper form.

IRS 1040 may have been incorrectly transcribed.[18] A third and final source of error is in the linkage process between US Census records, a protected identification key (PIK) and tax records' PIK.[19]

The second key takeaway is that, of the Marielitos participating in the labor market, the vast majority settled in the Miami MSA. This fills in the gaps between 1980 and the 1990 and 2000 U.S. decennial Censuses.[20] This table confirms the existence of a large labor supply shock in Miami by 1984, the first full year for which the Marielitos report earnings in 1985.

3.2 Miami's HWI Versus Simple Geographic Baselines

As one might intuit, this supply-shock induced a persistent decline in Miami's HWI relative to geographic controls, which we present in the figure below.[21] Both the national and South Atlantic HWI are constructed by the Conference Board by weighting the constituent MSAs in the according to the procedures described in Preston (1977). The trends in the raw HWI data are visually striking.[22] The index for Miami declined very rapidly after 1980, reaching a nadir towards the end of 1982, and then

[18]There is often little incentive to correct wrong administrative records after enough time has passed.

[19]The process of linking records at the US Census is described in Alexander et al. (2014).

[20]This also helps confirm the results of smaller special purpose surveys conducted at the time which found that most Marielitos stayed in Miami. See, https://www.upi.com/Archives/1990/05/05/A-decade-ago-on-May-5-President-Jimmy-Carter/3418641880000/, accessed 2019/02/21

[21]We normalize the HWI for Miami, the US-as-a-whole, and the South Atlantic region so that they equal 1 at some point in 1977 to 1979. We do this to adjust for different initial levels of the indices. This is akin to taking fixed effects.

[22]Geographic comparison groups have been attacked on a number of fronts over the last decade. Using a number of synthetic comparison groups and those of historical interest, Anastasopoulos et al. (2019) finds the same results.

began a slow recovery through the 1980's. By 1989, the value of the HWI index for Miami was again similar to the national index (although it was still lower than the index for the South Atlantic region).[23] The obvious implication is that as the Marielitos joined the labor market, they took positions that would otherwise have been advertised in the *Miami Herald*. This did not happen in the nation-as-a-whole or the South Atlantic region, and so Miami lagged behind these entities for a decade.

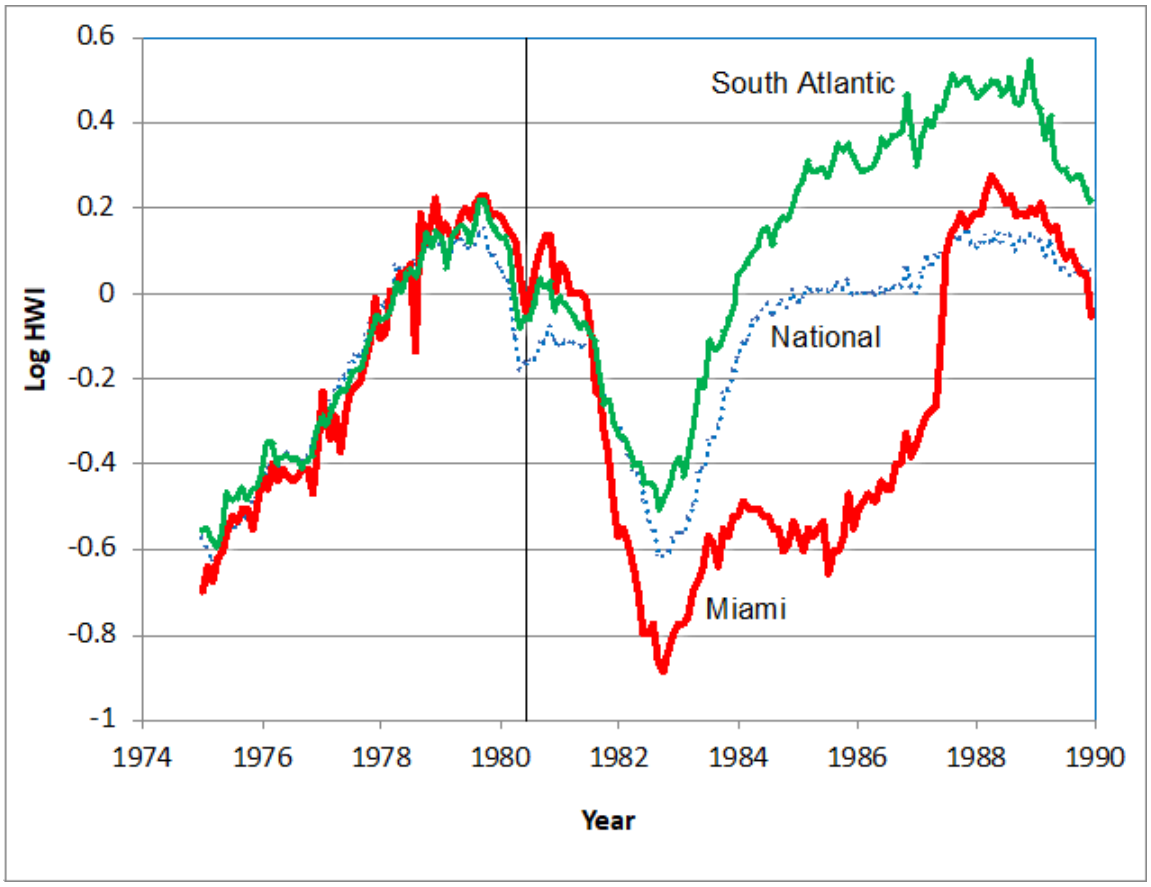

Figure 1: In this figure, we plot the log of a normalized Miami HWI versus that of the National and South Atlantic regions.

[23]Note that for this exercise, we have not dropped Miami from the South Atlantic region.

Learning what happened to the HWI is important in its own right, but there are a number of reasons to study the content of the ads.[24] First, prior studies on Mariel find that the effect of the Mariel surge was limited to the wages of high school dropouts, but Anastasopoulos et al. (2019) finds that the HWI correlates with high school graduate wages and the wages of those with some college. Thus, its a puzzle why a declining HWI did not also correlate with declining wages for high school graduates and those with some college education in Miami. To see how content analysis can solve this puzzle, suppose that all job vacancies were advertised in a single broadsheet with jobs for two classes of worker: less-educated and more-educated. For simplicity, suppose that the broadsheet is composed equally of both. If the supply of less-educated workers suddenly increases and they take all the open vacancies available, then the broadsheet will have 50% less vacancies but for unemployed less-educated workers 100% fewer job openings will be available. Furthermore, in a model with job vacancies, we might expect the complementarity between more and less-educated to increase the number of job vacancies for more-educated workers.[25] In other words, even small changes to the HWI may obscure important compositional changes. Statistical text analysis allows us to systematically detect these compositional effects.[26] Second, beyond the education- and wage-level changes in job openings advertised,

[24]One benefit we do not concentrate on in this essay is that directly counting advertisements allows us to sanity check the construction of HWI.

[25]In general, accounting for heterogeneity in job search by education level, increases the expected impact of composition effects on labor market outcomes. Less-educated people might be less qualified for job vacancies on average, and have fewer contacts, reducing the efficacy of both direct contact of employers and social networks - the two dominant forms of American job search in the 20th century.

[26]Specifically, if we see that the occupations advertised in the classifieds are associated with higher wages or education requirements, then this lends itself to the classical interpretation of the Mariel boatlift as a positive shock to the supply of less-educated.

we might be interested in how the labor demanded by particular occupations and industries evolve.[27] Recent work by Bana (2018) has shown that local occupation-level shocks are especially important for determining individual labor market outcomes.[28] Classifieds currently offer the only data source that could enable the measurement of day-to-day fluctuations in the quantity of labor demanded for a given occupation.[29]

4 Statistical Text Analysis for Economics

Statistical text analysis is a relatively new method in the social sciences, but it has already garnered a great deal of attention for its proven potential and promise.[30] Econometric data are much more readily available in structured, numerical format, but many sources of data of interest to political scientists are unstructured and hidden behind the complicated format of human language.[31]

[27]Parallel work by Atalay et al. (2018) uses a similar strategy to ours to study how technology shocks drive inequality by increasing the returns-to-education.

[28]One could also imagine using this approach to test more heterodox claims in economics. For instance, Jayadev and Bowles (2006) claim that economic inequality causes increases in the number of workers devoted to monitoring, guarding, and surveillance. They present between-country evidence supporting this hypothesis. On the other hand, Geloso and Kufenko (2017) present econometric evidence that the intensity of drug prohibition is a more robust, statistically significant, and economically significant determinant of the proportion of the working population devoted to guard labor-type tasks. Can we distinguish between these hypotheses using classifieds at the local level?

[29]They also offer insight into other hard-to-observe labor markets. For instance, the federal government released regular few economic statistics on minority populations prior to the 1970s. Minority-targeted classifieds offer one of the few ways to gain systematic and high-frequency data on the labor-demanded for specifically black and Latinx populations prior to the existence of official tabulations. We plan to pursue this in future work.

[30]It is most often applied in political science and economics.

[31]The methods we use, particularly for applied statistical text analysis in general were covered exhaustively in a previous draft of this essay. Readers wanting a more comprehensive introduction to these methods are directed to that draft, Anastasopoulos, Moldogaziev, and Scott (2017) which describes these methods in a public administration context, or Anastasopoulos et al. (2019). We present only an abbreviated description of our methods here in the interests of complying with conference space requirements.

Gentzkow, Kelly, and Taddy (2017) provide a thorough introduction to the text analysis in the social sciences that focuses on economic applications. They note many advantages of textual data for economic applications, particularly its ability to generate up-to-date estimates: "Important variables such as unemployment, retail sales, and GDP are measured at low frequency, and estimates are released with a significant lag. Others, such as racial prejudice or local government corruption are not captured by standard measures at all. Text produced online such as search queries, social media posts, listings on job websites, and so on can be used to construct alternative real-time estimates of the current values of these variables. By contrast with the standard exercise of forecasting future variables, this process of using diverse data sources to estimate current variables has been termed 'nowcasting'".

Financial economists have made ample use of advances in NLP in recent work.[32] Jegadeesh and Wu (2013) use a content analysis technique powered by text regression to analyze reports from the Securities and Exchange Commission. Manela and Moreira (2017) employ a similar strategy with a support vector machine (SVM) on articles from the Wall Street Journal.

Grimmer and Stewart (2013) write that text analysis is most effective when the text "is focused on the expression of one idea for classification". In this respect, classified ads are an ideal medium because they are designed for a single purpose and convey similar types of information. This gives our corpus of classifieds a degree of structure that many natural texts lack. Below we describe our sampling strategy, learn a supervised model, and calculate changes to the occupational structure of

[32]But see Lachanski and Pav (2017) for some common pitfalls in this type of work.

Miami's labor market.[33]

4.1 Supervised Learning Meets the Mariel Boatlift

We use supervised machine learning as a simple generalization of regression.[34] There are multiple directions in which one may relax the assumptions of the OLS regression model. Deep-learned neural nets, support vector machines, boosting, and tree models all address the problem of y not being a linear function of X in some sense. The famed LASSO and Ridge regressions and their assorted shrinkage estimators allow one to generate point-estimates for cases where the number of variables is larger than the number of observations. Strictly speaking, these do not exhaust the possibilities to extend regression, but the first case is in fact what we use supervised machine learning for. In addition to the problem of text variables' discreteness, the education level associated with a particular help-wanted ad is unlikely to be a linear function of its text in any case. We will feed a supervised machine learning algorithm many examples of job ads and their associated occupation's education level. The supervised machine learning model will use this labeled data to "cut out the middleman" and build a non-linear statistical model mapping a job ad's text directly to the education level associated with that text. Then, we will use that statistical model to estimate the fraction of job ads targeted towards the less-educated for the pre- and post-Mariel periods.

[33]This exercise complements Anastasopoulos et al. (2019) which conducts a similar analysis using the raw output from an O*NET autocoder designed by the U.S. Department of Labor and uses of a number of alternative baselines in a difference-in-difference causal inference approach.

[34]The machine learning is supervised in the sense that it is given access to an O*NET autocoder's output occupation labels for each ad. An unsupervised machine learning algorithm does not utilize any such labels for the text.

The field of natural language processing, or NLP for short, is fundamentally concerned with representing natural language units in a form that can be understood by and manipulated through a machine for the purpose of systematically analyzing them.[35] Once the features of documents can be represented on a machine as numerical values, all of the computational and statistical methods which can be applied to numerical data can then also be applied to texts.[36]

4.1.1 Sampling Strategy

The most basic unit of analysis in NLP is the *term*, which is analogous to a word or group of words which frequently co-occur.[37] *Documents* are comprised of groups of terms and are typically the main unit of analysis when analyzing texts. Documents can be anything from sentences and paragraphs to entire works of literature. Here, the documents that we refer to are help wanted ads.[38] Finally, a *corpus* is a collection of documents that are analyzed.[39]

[35]Natural language units include words, phrases, sentences, paragraphs, essays, etc. Historically, NLP has analyzed linguistic representations derived from both sound and text, but we focus entirely on the latter category.

[36]In the social sciences, natural language processing methods have been used to estimate media bias (Young and Soroka, 2012), identify the politically relevant features of texts (Barberá, 2014; Bond and Messing, 2015; Lowe et al., 2011; Monroe, Colaresi, and Quinn, 2008) and to measure agendas in political texts (Grimmer, 2009).

[37]Terms can be of two types: *words* or *n–grams*. Words in NLP are the same as they are in common usage while n–grams are typically two or more words that are treated as a single unit. For example, names of places such as "New York", "Los Angeles", "Ohio State University" etc. are n–grams that might be treated as a single term for purposes of analysis.

[38]Statistical text analysis can be distinguished from natural language processing in general in that the methods of statistical text analysis exploit relatively few features of a particular language. In theory, one could apply the techniques we utilize here to Spanish or Chinese classifieds with few modifications. By contrast, many natural language processing approaches achieve their objectives by leveraging particular linguistic features of the language in which the text is written such as sentence structure (e.g. subject-object-verb as in Korean versus subject-verb-object as in English).

[39]A corpus can be thought of as the equivalent to a "dataset".

In our case, the single corpus that we are analyzing is a 1/29 sample of the *Miami Herald* for the selected years: 1978, 1979, 1983, and 1984.[40] Equation 1 provides an overview of the NLP processing hierarchy. Here we see that the corpus (universe of help-wanted ads for selected dates) is comprised of multiple documents (job openings advertised) which are each in term comprised of terms (words in each ad).

$$terms \subseteq document \text{ (classified ads) } \subseteq corpus \text{ (1/29 sample of } Miami\ Herald) \quad (1)$$

Once dates were selected, we scanned the *Miami Herald* for that date and selected every page or area of a page containing a classified ad for conversion to PDF.[41] Once these pages were digitized, they were transcribed and broken into individual job openings using Amazon Mechanical Turk. This procedure yields 15,412 ads for 1978, 15,666 ads for 1979, 6,342 ads for 1983, and 11,435 for 1984. From this total of 48,885 ads, we drop all low-quality transcriptions for a total of 46,072 ads in both our supervised and unsupervised analyses. Ads were inspected by hand. The breakdown of clean ads by year is: 15,062 ads for 1978, 14,303 ads for 1979, 6,342 for 1983, and 10,362 for 1984. Figure 2 is a plot of the most frequent words found in our sample. Unsurprisingly, the most frequent stemmed terms are "call", "salari", "experi", and "benefit".

[40] See Table 4 for the exact dates selected into the sample and an explanation of the sampling procedure.

[41] In this way, for each date selected into the sample we obtain the approximate universe of help-wanted ads bar human error. Wherever available, we used the "First Edition" of the *Herald*, as this is most likely to reflect MSA-wide trends and job opportunities facing actual labor market participants. If the "First Edition" was not available or in such a deteriorated condition that it could not be transcribed, we sampled from the "Final Edition" aimed at the entire Miami area, rather than a county or region-specific edition. Luckily, for the sample used in this article, this procedure was sufficient to obtain the approximate universe of job ads for each date in Table 4.

Figure 2: Word cloud representing the most frequent terms found in the *Miami Herald* ads for our sample.

4.1.2 Gradient Boosted Trees Meet the Mariel Boatlift

We conduct our analysis of the classified ads at the time of Mariel using a supervised machine learning approach with gradient boosted trees. We examine the text of these classifieds to document: (a) the types of ads that typically appeared in the help-

wanted section of that newspaper; and (b) whether there was a relative decline in classifieds for less-educated workers in Miami.[42] Our evidence suggests that positions for the less-educated comprised about forty percent of the advertised jobs at the time, and that there was a statistically significant decline in the fraction of ads advertising vacancies for the less-educated after Mariel.[43]

Our supervised machine learning analysis involves three steps: (1) coding of a subsample (also known as the training data) of help-wanted ads into vacancies associated with occupations requiring more and less-education; (2) using the coded data to train, fine-tune, and test the algorithm that will be used to classify ads into less- and more-educated vacancies; and (3) assigning all ads outside of the training data to one of the two education groups.

We first need to acquire high-quality, coded classified ads to teach the algorithm how to classify a particular ad into a specific education group. We selected a random sample of 25 percent of the 46,072 classifieds in our sample. The text of these classifieds, with the dates stripped, were then classified by the O*NET-SOC AutoCoder v12.4, which classified them as belonging to an occupation. The random sample of classifieds was then used to train a gradient-boosted trees classifier to distinguish

[42]We define a less-educated position in the following manner. Rank all occupations in descending order by the total number of high school dropouts they employ using the 1990 US Census Decennial. Any occupation in the top 50 in an occupation for the less-educated. The IPUMS 1990 US Census about 500 occupations. Thus, this designation corresponds with the 10% of occupations with the most high school dropouts. Results were robust to a variety of reasonable alterations to this definition.

[43]It would be fascinating to compare the initial occupation of the Mariel refugees with the advertising trends uncovered by the topic model. Unfortunately, the first time we observe the occupation of the refugees in a reasonably sized sample is in the 1990 US Census's Decennial, by which time they have already acquired 10 years of experience in the U.S. labor market. Business tax records required that businesses report their industry and so one could imagine future work using these records being used to pin down the likely occupational choices of the Marielitos.

between ads using only the words in the ads. Gradient-boosted trees are a popular type of decision tree algorithm which, like their random forests predecessor, grows multiple trees from random subsets of the training data and use a majority vote of the trees to generate the final class label. Gradient-boosted trees have become popular in the economics literature and have been used for text classification because of their transparency and their ability to generate highly accurate predictions in various contexts (Athey, Tibshirani, and Wager (2019), Chalfin et al. (2016)). Gradient-boosted trees tend to exhibit significantly improved classification performance over vanilla random forests because they have several hyperparameters that can be fine-tuned using cross-validation methods.

Training the algorithm to identify ads for the less-educated and more-educated in the training data involved the following steps: (1) text pre-processing; (2) conversion of text into a document-term matrix; (3) algorithm training and fine-tuning via cross-validation; and (4) performance assessment on the test data. The text pre-processing stage involves standardizing the text of the ads so that only the words (or parts of words) with the highest amount of useful information are retained (Denny and Spirling (2018); Gentzkow, Kelly, and Taddy (2017); and Grimmer and Stewart (2013)). The processed text is then converted into a document-term matrix.[44] The entry in each cell of the matrix is the number of times the word appeared in the help-wanted ad.

The training process then involves randomly selecting a training and test set. We opt for a 75/25 train/test split which is the default setting on most machine learning

[44]The rows in this matrix contain each classified ad in the training data, and the 1,355 columns contain the number of "cleaned" words left after text pre-processing.

software packages (Pedregosa et al., 2011). Model training involves prediction of the less- or more-educated classifications for each ad in the training data using only the words contained in the document term matrix. This is accomplished through growing multiple trees via an iterative loss minimization process using an objective function, $O(\theta)$ which is comprised of a logistic regression loss function $L(\theta)$ of the tree parameters θ and a regularization term, $\Gamma(f_k)$, which is a function of the number of k trees grown where each tree is represented by a function $f_k \in F$ in the function space F of all possible trees:

$$O(\theta) = L(\theta) + \sum_{k=1}^{K} \Gamma(f_k) = l((c_i, c_i^p)) + \sum_{k=1}^{K} \Gamma(f_k) \tag{2}$$

The goal of training is to minimize $O(\theta)$ by simultaneously accounting for the difference between the true and predicted skill classification for the classified ads (c_i and c_i^p) and the regularization term $\sum_{k=1}^{K} \Gamma(f_k)$ which prevents overfitting of the model. An important part of the training process involved hyper-parameter tuning using 10-fold cross-validation on the training data to select the model with the minimum average cross-validated test error as defined by the objective function in equation 2. This exercise reveals decline in the relative number of low-skill ads published in the Miami Herald between the pre-Mariel period, 1978-1979, and the post-Mariel period, 1983-1984 which is statistically significant at any standard level. The full results of the supervised machine learning model we estimated are in the final row of Table 5.

One of the benefits of tree-based classifiers is that they provide some understanding of how the classifier makes its decisions. In our context, the tree provides

information about which words the classifier used to best distinguish between low- and high-skill vacancies. Figure 3 plots the terms (ranked from most to least important) that help distinguish classifieds using the information gain across trees. Interestingly, terms such as "manag" and "driver" which are terms overrepresented in ads targeting workers for the most and least educated occupations help to best distinguish between the classifieds.

4.2 Unsupervised Learning Meets the Mariel Boatlift

Unsupervised algorithms are models which do not use class labels to classify data, but rather automatically generate groups or clusters from independent variables (features). For this study, we provide our unsupervised machine learning algorithm with all of the job ad text. It will look for distinct clusters of words in the job advertisements which it will say belong to different categories with some probability. Most importantly, the algorithm will assign each ad to some combination of topics probabilistically. In words, our unsupervised algorithm might generate clusters for the sales and management occupations and the engineering occupations. Then, an ad for a "managing engineer", for example, that went into the topic model could be assigned a probability of being drawn 50% from the sales and management cluster and 50% from the engineering cluster.

Unsupervised algorithms include k–means clustering, hierarchical clustering, principal components analysis, multidimensional scaling and the Latent Dirichlet Allocation (LDA) of which there are a number of variants.[45] In the context of text analysis,

[45]These include correlated topic models, structural topic models, etc

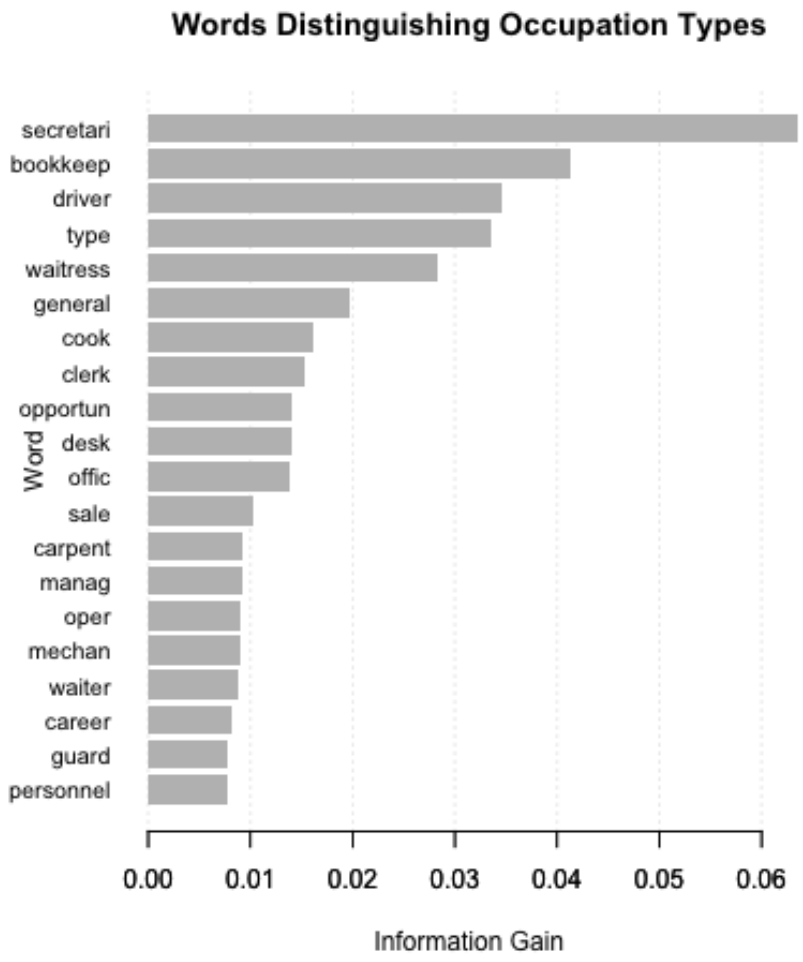

Figure 3: The importance of words, as measured by information gain, for distinguishing between ads targeting less- and more-educated workers are displayed from top (most important) to bottom (less important).

unsupervised learning methods compute the similarities between documents using the document–term matrix based on some distance metric.

Here, we propose the use of LDA, an unsupervised algorithm, as a general method for break job ads into industry and/or occupational categories. LDA is a nonpara-

metric Bayesian method and a feature of the LDA approach, which makes it an excellent method for analyzing texts, is that it has the ability to cluster texts and also provides information about how to *interpret* the clusters (Blei and Lafferty, 2007). For this reason, LDA is often colloquially referred to as a "topic model" because it scales text into a multidimensional set of topics that reflect underlying document themes (Blei and Lafferty, 2007; Grimmer, 2009). These unique features also allow us to identify theoretically relevant aspects of institutional documents as we do with job advertisements below.[46]

4.2.1 A Very Brief Introduction to Topic Modeling

Before describing the topic modeling process in more detail, we must first understand what a topic model does and how it is structured. The topic model allows us to do two things with texts. First, it allows us to get a sense of the common latent thematic elements across a corpus or collection of documents. For example, in our corpus of classified ads the topic model would be able to tell us information about job clusters common to ads in both pre- and post-Mariel Miami. Second, topic models allow us to measure how much of each topic is contained within each document. For instance, if we discover that the common themes across our corpus are related to more- or less-educated labor, then we would conceptualize the corpus as having two underlying dimensions, or topics, that correspond to this particular set of classified ads.

For each classified ad in that corpus, then, the topic model would be able to tell us whether each ad is likely to be related to an employment opportunity requiring more

[46] In what follows, we focus primarily on the LDA method, but do emphasize that it is a specific example of a more general category of unsupervised classification methods.

or less- education. For example, an exploration of the distribution of topics in a set of classifieds might suggest that an ad for janitorial work has a predicted probability of belonging to the less-educated "topic" equal to 0.90 and a predicted probability of belonging to the more-educated topic of 0.10. In this situation we can use either the predicted probability as a measure of the *extent* to which the ad contains tasks relevant to less-educated labor or we can simply classify the ad as belonging to the less-educated category using the heuristic if $P(LessEducated) > 0.50$ then classify the ad as belonging entirely to the less-educated category, otherwise classify the ad as belonging to the more-educated category.

Empirical evaluation of models with a different number of topics might reveal a better fit; for instance, perhaps less-educated work diverges clearly into food service and janitorial work and is thus better represented by two topics instead of one. In this sense, topic modeling is very similar to exploratory factor analysis as a means of reducing the dimensionality of texts. We discuss methods for testing and selecting the number of topics further in the section to follow. In summary, the topic model gives us pieces of information for any collection of documents: (1) a number of topics which are contained within a corpus and; (2) for each document contained within the corpus, what proportion of each of the topics is contained within the each document.

As with all text analysis problems which we discussed above, the fundamental unit of data used in topic models are *terms* as represented in the document–term matrix. Terms are treated as items from a *vocabulary*, indexed by a set of numbers $\{1, ..., V\}$. The vocabulary are all of the terms in a given *corpus* or collection of documents as discussed above.

A *document* is a bag of N terms. We describe a document as a "bag of terms" rather than a series or sequence of terms in a particular order because the topic model does not take the order of terms or words into account. These N terms can be represented by a vector $\mathbf{w} = (w_1, w_2, ..., w_N)$. A *corpus*, as above, is a collection of M documents which can be represented by $D = \{\mathbf{w_1}, \mathbf{w_2}, ..., \mathbf{w_M}\}$. The topic model treats each document within a corpus as a mixture of a fixed number of k latent topics which is represented by a distribution over words.

4.2.2 A Topic Model of Classified Ads

Topic models, with careful theoretical guidance and the appropriate data, can shed light on many theoretically relevant questions in labor economics which have traditionally been limited by the inability to systematically analyze large quantities of documents. One of these questions relates to understanding which industries and positions are most affected by less-educated migration. Specifically, we are interested in exploring whether the large exogenous immigration shock brought about by the Mariel Boatlift affected employment opportunities for less or more-educated natives at the time. By using topic models applied to classified ads, we can directly address this question across time.

Unlike survey data, classified ads can give us a window into the demand side impacts of immigration. Traditionally, one would have research assistants or other coders read each of these classifieds and use their own personal judgment to figure out which which employment categories they belong to. There are two major issues with such an approach: time and expertise. First, time constraints make gathering and

reading potentially hundreds or thousands of classifieds unfeasible, thus requiring that inferences be made on a small subset of classifieds. Second, systematic analysis of classified ads requires the researcher to rely exclusively on coders and requires that these coders to consistently apply the same criteria for labeling each ad.

The topic model solves both of these problems simultaneously. Topic modeling can extract latent themes from hundreds of thousands of classified ads in a systematic manner and it puts the power back in the hands of the domain expert. Below we describe how classified ads are modeled with topic models and then move on to descriptive analysis of *Miami Herald* classifieds before and after Mariel.

4.2.3 Modeling Classified Advertisements

The essential first step toward modeling any set of texts using topic models is division of these texts into *corpora* and documents. For our purposes, we define:

- **Document** - A classified a within an issues of a newspaper d, in year t is represented as a sequence of N terms $\mathbf{w} = (w_1, w_2, ..., w_N)$.
- **Corpus** - The collection of classified ads collected. This corpus includes over 46,000 classified ads collected from four years: two years selected from before and two years selected from after the Mariel Boatlift.

The LDA is a generative probabilistic model of the corpus of classified ads treated as a random mixture over k latent topics. Each topic as a distribution over the terms. But how do we know how many topics a corpus contains? Because the LDA does not automatically select the number of topics that a corpus is comprised of, the

researcher must ultimately decide how many topics that the corpus is comprised of on the basis of a number of factors.[47]

In addition to this, another popular method for topic selection involves estimating a topic model using $k = \{2, \cdots, n\}$ topics, measuring the perplexity of each model and choosing the model for which the marginal perplexity stops decreasing (Blei and Lafferty, 2007; He et al., 2013; Hinton and Salakhutdinov, 2009). Perplexity is an information theoretic metric which measures how well probability models predict a sample which we describe in further detail below. Lower values of perplexity imply models that better fit the data.[48] Thus, we use perplexity as a guide along with theoretical guidance to determine the number of topics present in our corpus. Using these criteria, we discovered that the best model contained $k = 10$ topics. In words, this implies that there are a total of 10 latent topical categories within the corpus of classified ads, and that each of these ads is a mixture over these 10 latent topical categories which can be interpreted as categories of jobs. The proportion of the ad devoted to each topic is represented by θ_{ad}, the distribution of topic proportions for each classified.

Figure 4 is a graphical model of the LDA as applied to the *Miami Herald* ads using plate notation to denote replicates of the ads D and the terms within each ad N. Each of the nodes represents a random variable. The only observed variable is the collection of terms which comprise the corpus. All other variables are unobserved latent variables which are estimated by the LDA. The graphical model above

[47] In most cases, theoretical guidance provided by domain expertise combined with human interpretability should serve as the first guide.

[48] While perplexity often provides a good means of guiding researchers, many argue that it should only be used as a guide rather than the sole means of choosing the appropriate number of topics.

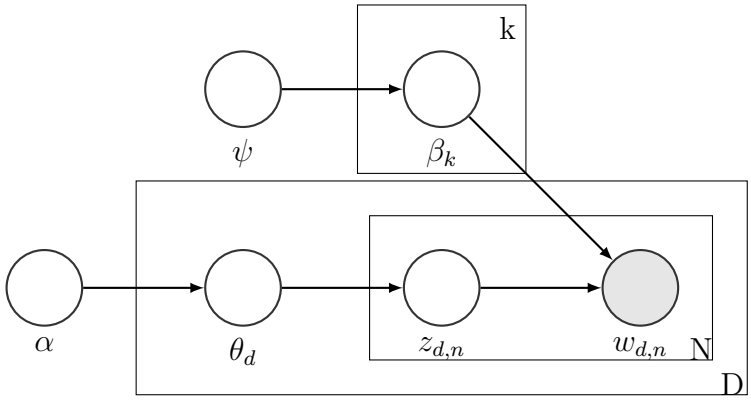

Figure 4: Graphical representation of the topic model as applied to classified ads using plate notation

assumes that $w_{d,n}$, each term in each classified ad in the corpus is generated from both a distribution over latent topics, which corresponds to each individual ad and a distribution over words which comprises the employment category.

We define:

1. $\beta_k \sim Dir(\psi)$, where $k \in \{1, ..., 10\}$ - the distribution over words that defines each of the $k = 10$ latent topics estimated from the 1978-1979 and 1983-1984 ads.

2. $\theta_d \sim Dir(\alpha)$, where $d \in \{1, ..., 46072\}$ – the distribution over topics for each classified ad in the sample .

3. $z_{d,n}$ - topic assignment of the n^{th} word in the d^{th} ad.

4. $w_{d,n}$ - the n^{th} word of the d^{th} ad.

The probability distributions of topic proportions for each ad $p(\theta_d|\alpha)$ and of each topic in all ads $p(\beta_k|\psi)$ are distributed Dirichlet with hyperparameters α and ψ

respectively. Thus topic proportions for each ad has the distribution:

$$p(\theta_d|\alpha) = \frac{\prod_{i=1}^{k}\Gamma(\alpha_i)}{\Gamma\big(\sum_{i=1}^{K}\alpha_i\big)}\prod_{i=1}^{7}\theta_{di}^{\alpha_i-1}$$

And each topic k across all ads has the distribution over terms:

$$p(\beta_k|\psi) = \frac{\prod_{i=1}^{N}\Gamma(\psi_i)}{\Gamma\big(\sum_{i=1}^{N}\psi_i\big)}\prod_{i=1}^{N}\beta_{ki}^{\psi_i-1}$$

The remaining distributions that we need in order to specify the model including topic assignment conditional on topic distribution $p(z_{d,n}|\theta_d)$ and term conditional on topic assignment $p(w_{d,n}|z_{d,n},\beta_k)$ are multinomial with:

$$z_{d,n} \sim Multinom(\theta_d) \tag{3}$$

$$w_{d,n} \sim Multinom(\beta_k) \tag{4}$$

Putting all this together, we arrive at the fully specified model over all classified ads in the corpus:

$$p(\theta, \mathbf{z}, \mathbf{w}, \beta|\psi, \alpha) = \prod_{k=1}^{7} p(\beta_k|\psi)\prod_{d=1}^{D}\left(p(\theta_d|\alpha)\prod_{n=1}^{N}p(z_{d,n}|\theta_d)p(w_{d,n}|z_{d,n},\beta_k)\right) \tag{5}$$

Estimating $p(\theta_d|\alpha)$, which we use for understanding changes in demand for each of the seven employment categories identified and all other relevant hidden parameters requires posterior inference using the variational expectation-maximization algorithm

(VEM) algorithm (Blei and Lafferty, 2007) which is implemented in **R** packages such as *topicmodels* and *lda*.

4.2.4 Results

Tables 6 and 7 contains the final estimated topics from the topic model along with topic labels.[49] The major topical employment groups include accountants, engineering, management and sales related jobs, secretarial jobs, the medical industry, food service, and the automotive industry.[50] Three of the categories we had trouble interpreting and so we collapsed these topics into a miscellaneous category comprising a little over a fifth of all total ads. Table 6 contains the final topics we used with labels.[51]

Using these topics, we broke down ads further into less- and more-educated labor categories and explored trends in each category. According to these estimates, jobs advertised for the more-educated in the *Herald* remained roughly the same for secretarial and medical service positions, but *increased* for sales, engineering, and accounting positions as predicted. A glance at ads targeted towards the less-educated, however, tells a different story. We see trends suggesting declines in the availability of positions in the highly elastic food service industry and automotive sectors that would be most sensitive to a large increase in the supply of the less-educated

[49] Unfortunately, the topics that our perplexity score chose do not respect the distinction between industries and occupations. Nonetheless, it is straightforward to estimate the educational content typically associated with job advertisements for each category, and by hand-inspecting ads associated with each category we verified these estimates.

[50] The medical industry advertisements primarily target nurses, billing, and hospital roles.

[51] Previous drafts of this article and Anastasopoulos et al. (2019) contain numerous examples of help-wanted classifieds.

workforce. Figure 5 is a plot of pre- and post-Mariel trends for the less- and the more-educated categories. Here we notice that ads for most less-educated job categories like food service and the automotive industry have declined by several percentage points.[52] Table 5 summarizes these trends.

A reader who has been following the Mariel literature will notice how congruent these findings are with it in at least three senses. First, the decline in the the fraction of ads targeted towards less-educated in Miami aligns with the results reported in Borjas (2017) and Yasenov and Peri (2018).[53] Second, work on Mariel has not unearthed a significant impact on female wages.[54] Perhaps then we should not be surprised that we are unable to detect a change in advertising for the then-feminized secretarial professions and those in the medical field. Medical occupations have a large amount of occupational licensing, preventing refugee entry.[55] Finally, as noted by Bodvarsson, Van den Berg, and Lewerd (2008)), immigrants are consumers as well as producers, and so an expansion in the fraction of help-wanted advertisement targeting sales and management roles should be expected.[56]

[52]Declines and increases are statistically significant at any standard level.

[53]A divergence in causal estimates on the impact of less-educated immigration between these two articles arises from disagreements over the correct counterfactual, which we do not address here; see Yasenov and Peri (2018), Figure 9.

[54]This is likely because female Marielitos did not have the English-language experience required to compete in less-educated occupations held by less-educated native-born females in Miami at that time.

[55]Additionally, the response of job applications to job ads is probably smaller for the credentialized medical application than for the less credentialized sales and management, food service, or automotive sectors.

[56]As noted above however, some of this relative expansion in white-collar roles may be driven by increasing EEO compliance unconnected to changing labor market fundamentals.

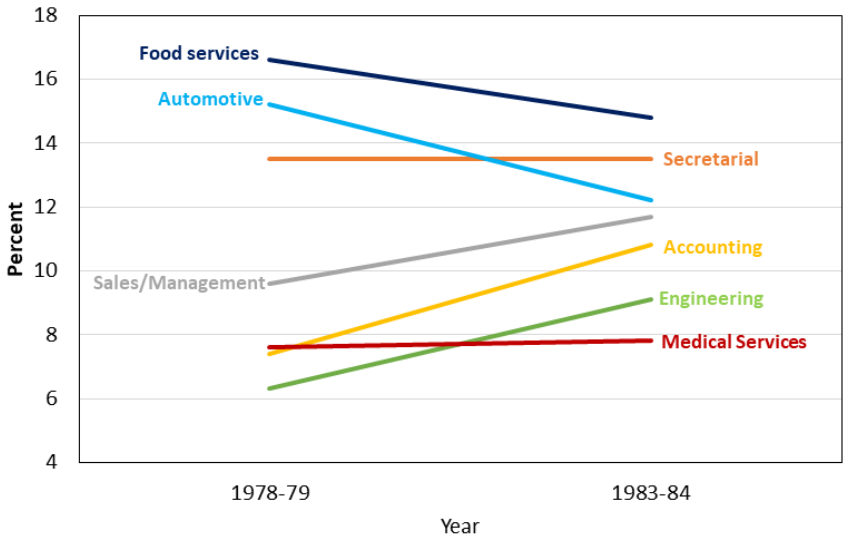

Figure 5: Pre- and post-Mariel trends for categorical topics.

5 American Job Openings in the 21st Century

Anastasopoulos et al. (2019) presents evidence that increases in job openings as proxied by the help-wanted index benefit some workers more than others. In particular, they conjecture that the distributional consequences of an increase in unfilled positions are progressive at least in the short-run in the sense that increases in the HWI yield the largest expected wage and employment levels for high-school dropouts, then high school graduates, then workers with some college, and finally college graduates.[57] Starting in December of the year 2000, the Bureau of Labor Statistics began making public statistics on monthly regional job openings through their Job Openings and Labor Turnover Survey (JOLTS) available. The JOLTS time series, which

[57]Unfortunately, it is unclear if this reflects biases in the construction of the HWI, which likely excludes many job openings for the highly-educated because those job openings were not always posted in the help-wanted section of the newspaper.

starts right when the statistical exercise with the HWI in Anastasopoulos et al. (2019) ends, provides an opportunity to conceptually replicate the findings reported in Anastasopoulos et al. (2019) ex-sample.[58] Consider the following regression model:

$$y_{rst} = \theta_r + \theta_s + \theta_t + \beta_0 \log V_{rt} + \beta_1 \theta_s \log V_{rt} + \epsilon \tag{6}$$

where y_{rst} is a labor market outcome for region r, education group s, and calendar year t; θ_r is a vector of region fixed effects; θ_t is a vector of calendar year fixed effects; and V_{rt} is a the average number of job openings in region r over year t. By including the city fixed effects, the coefficient vector (β_0, β_1) is essentially estimating the correlation between a within-region change in job openings and the corresponding change in labor market outcome y, and how that correlation varies across education groups.

We estimate equation 6 using four education groups: high school dropouts, high school graduates, some college, and college graduates.[59] We are interested in two labor market outcomes: wages and employment. The average wage for cell (r, s, t) is calculated from residuals to individual-level regressions estimated in the CPS data. We use a simple regression model to calculate the age-adjusted mean wage of an education group in cell (r, s, t). Specifically, we estimate the following individual-level earnings regression separately in each CPS cross-section:

[58] If the findings replicate using JOLTS, then it is likely that the association between low-wage outcomes and HWI reported in Anastasopoulos et al. (2019) reflects a robust statistical relationship between increased aggregate job openings and better outcomes for the less-educated in particular.

[59] High school dropouts have less than 12 years of education. High school graduates have exactly 12 years of education. People with between 13 and 15 years of education are in the some college group. All those with more than 15 years of education are in the college education group.

$$\log y_{irst} = \alpha_t + \gamma_t \boldsymbol{A}_i + \epsilon \tag{7}$$

where y_{irst} is the wage/employment level of worker i in region r with education s at time t; and $\boldsymbol{A}_i$ is a vector of age fixed effects. We used seven age groups to create the fixed effects for wages which exhaust the range of ages we examine for the wage regressions.[60] For our employment regressions, we use nine age groups to create the fixed effects which again exhausts the range of ages we examine for this exercise.[61] The average residual from this regression for cell (r, s, t) gives the age-and-sex adjusted mean wage of that cell, which is the outcome we regress on in (6). We use seasonally adjusted job openings.[62] We estimate regression (7) using weekly wage data from the March CPS covering the period 2000-2017 (Flood et al., 2018). The calculation of the average log wage in the cell weighs each individual observation by the product of the persons earnings weight times the usual number of hours worked weekly. The employment variable is defined as the number of workers in cell (r, s, t) who either worked at some point in the calendar year prior to the survey.

The various panels of Table 8 report the coefficients in the vector β for various specifications of the regression model. Consider initially the regression coefficients in the first row of the top panel of the table. These coefficients measure the correlation between wage trends in the March CPS and job openings. The interaction of the job openings with the education fixed effects indicate that the wage of less-educated

[60] Those age groups are: 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, and 55-59.
[61] Those age groups are: 19-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, and 55-59, 60-64.
[62] Results do not significantly differ if we use raw job openings.

workers is more strongly correlated with job openings than the wage of workers with more education. In fact, as in Anastasopoulos et al. (2019) the point estimates for the correlation decline monotonically with education. A 10 percent increase in the number of job openings is associated with a 0.39 percent increase in the wage of high-school dropouts (relative to college graduates), a 0.03 percent increase in the wage of high school graduates, and a 0.02 percent increase in the wage of workers with some college.[63] Note that this is the same pattern of decline in the job openings coefficients by education as in Anastasopoulos et al. (2019), but with an order of magnitude fewer observations.

The second row of the panel shows the correlation between employment trends and job openings. The March CPS shows the same pattern in the correlation between the BLS's job openings measure and employment as in Anastasopoulos et al. (2019) in that it is largest for the least educated. A 10 percent increase in the index is associated with a 0.12 percent relative increase in the number of workers employed who are high school dropouts, a -0.11 percent relative increase in the number of workers who are high school graduates, and a -0.0055 percent relative increase in the number of workers with some college. Only the coefficient on high school graduates is significant at any standard level and the results are sensible in the sense that higher levels of job openings predict a (statistically insignificant) increase in employment for all education groups.

While these results are interesting, they ultimately provide mixed evidence for the findings reported in Anastasopoulos et al. (2019). High school dropouts appear

[63] The latter two coefficients are not distinguishable from zero.

to be the biggest beneficiaries of a "hot" economy, but we find no statistically significant evidence for the monotonic decline in the relevance of job openings for labor market outcomes as Anastasopoulos et al. (2019) did. There are at least three reasons why would expect a weaker relationship than that reported in Anastasopoulos et al. (2019). First, we have fewer years and regions and consequently, much lower power than they did. Second, consider the size of the regions themselves. The HWI provided estimates for 51 metro areas and aggregated these into a number of regions corresponding with US Census categorizations. For the job openings dataset we only have 4 large regions, each of which encompasses several states. If the regions sampled in the CPS do not correlate with the areas in which the JOLTS firms are located, then we are much more likely to be measuring second-order impacts of job openings than in the HWI exercise - given that anyone in the CPS sample of the metro area could have opened up a newspaper and applied for the jobs therein. Finally, the demographic and personal characteristics of who falls into these categories has changed over time.[64]

6 Conclusions and Suggestions

Motivated by tabulations suggesting a surge of less-educated immigrants to the Miami area in the early 1980s, we conducted a comparison of the number of classified

[64]In the first 40 years of the post-war era, it is easy to find MSAs in which between one-in-four and one-in-five workers do not have high school diplomas. Today, only seven percent of the workforce does not have a high school degree. Similarly, educational attainment in the United States is higher than ever and expected to continue growing. With the proliferation of for-profit universities and online education, we expect the reference category used in this analysis to become even less informative over time.

ads in the Miami area versus those found in some simple geographic comparison groups. We find clear declines in the Miami index. Investigating the content of the classifieds, we construct an estimate of the fraction of ads targeting the less-educated and the more-educated in the *Miami Herald* using a supervised machine learning algorithm. This approach uncovers a statistically significant decline in the fraction of jobs advertised for the less-educated. Digging deeper, we found that the declines were partially offset by changes in composition favoring those pursuing roles in management and sales, accounting, and engineering. On the other hand, compositional changes magnified apparent decreases in job opportunities for those seeking positions in the automotive industry and food services.

Our last exercises suggest that job openings today are strongly correlated with wages for all workers, but especially the least-educated. Much has been made of the most recent years' increases in the number of job openings, but our analysis cannot detect any regional increase in employment (adjusting for worker age and sex) resulting from increases in the JOLTS. Given that one of the stylized facts of the job search literature is that higher levels of job openings should lead to more employment, this suggests that today's JOLTS dataset is, if anything, a little less informative than yesterday's HWI, short-lived state-level vacancy statistics, short-lived national manufacturing vacancy statistics, or the measures we derived directly from classifieds in this paper.[65] For assessing and predicting worker outcomes in the

[65]If educational attainment is not an important moderator of the JOLTS-unemployment relationship, then adding a moderator like education should further reduce the power of the regression unnecessarily but not impact the point-estimates of the coefficient. Collapsing the high school graduate and high school dropout into a single category, or collapsing all education categories but college graduates did not significantly alter the results we present in Table 8 for unemployment. A regression of the same form that includes only regional fixed effects and a national time trend

21st century, we strongly recommend that the occupational, industrial, and regional categories in the JOLTS survey be increased.

Paradoxically, in attempting to forecast the impact of expected future policy changes for this project, we have reached deep into America's past. For many, the Mariel boatlift is becoming more of an economic history topic rather than one which can inform current policy debates. Changes to America's social safety net, demographic structure, and educational system since 1980 lend some credence to this perspective. On the other hand, environmental policy shocks from the 1970s and inequality over the last century are still considered relevant for conducting short- and long-run policy analyses. Prudence requires generalizing from key features from previous episodes in economic history onto future situations that academics and policymakers in the public and private sector may encounter. In some cases, the most relevant empirical analysis must make use of events before the relatively data rich present and recent past.

In the case of immigration's impact on native-born worker wages and employment, contemporaneously collected public-use microdata have not lent themselves to a decisive conclusion. While we were able to find a well-known dataset (the HWI) and construct a new dataset from newspapers that lend themselves to an uncomplicated assessment of the Marielitos' impact, deteriorating historical data may threaten this

but no education variables found that regional job openings' impact on contemporaneous regional unemployment is not statistically different from zero. While it is possible to discover specifications that restore the significance of the link between employment levels and the JOLTS at the regional level, it should be alarming for data users that adding a small set of indicator variables is sufficient to break the contemporaneous link between job vacancies and employment in a standard regression set-up using all public data available.

capability for analyzing other local economic shocks in future.[66] In contradistinction with academic and private sector best-practices, most statistical agencies globally have no legislated or internal mandate to maintain and preserve their historical data stores, but these data stores will only grow in value with time for both historical purposes and as inputs to data-hungry private sector and academic algorithms.[67] Agencies of all kinds have created chief data officer roles to engage with public and internal data consumers and to create more accessible data products. Given the high potential value added by exploiting and maintaining heritage data stores for long-run forecasting, such as that called for by the US 2050 Initiative, perhaps agencies and legislators looking to modernize should consider the creation of chief heritage data officers, who would create value by searching for, preserving, organizing, and documenting older data stores for internal and public use.[68]

[66]For instance, we discovered no official written record of the newspapers actually used in the HWI existed, as the Conference Board kept it a trade secret. Zagorsky (1993) lists nine large city newspapers used in the HWI, but the years of their inclusion is unknown. Since the HWI proper has been discontinued, we contacted the now semi-retired Kenneth Goldstein of the Conference Board to uncover the final list in Table 1. Had we not been able to make contact, 42 of the 51 papers used would have been lost forever. Even so, we were unable to confirm the newspaper used to characterize vacancies in the Oklahoma City metro area.

[67]One option could be to adopt something akin to the 72-year rule for all surveys. Allowing this data to enter the public domain would greatly facilitate the preservation of these datasets, ensure comparability of the data across time, and allow for the reassessment of prior research in social science in addition to all the benefits listed in the main text. Given that that surveys like the CPS were voluntary, a shorter rule could be appropriate. In the case of the basic monthly CPS files from 1962, which are the earliest data available from IPUMS for the CPS and whose raw microdata are not available in any form for public use (as far as we could tell), it would be almost two decades before these files could be released if the 72 year rule were in effect.

[68]Jarmin (2019) notes the challenge of preserving respondent privacy in publicly available microdata and tabulations going forward, but for many heritage datasets the disclosure risks are relatively limited. Many survey respondents are deceased. Still living survey participate less in the social media and digital ecosystem that lends itself to large disclosure risks. As a general rule, the US Census has greater difficulty successfully linking survey participants from older surveys than from more recent ones (see Bond et al. (2014) for an example).

References

Abraham, Katharine G, and Lawrence F Katz. 1986. "Cyclical unemployment: sectoral shifts or aggregate disturbances?" *Journal of political Economy* 94 (3, Part 1): 507–522.

Abraham, Katharine G, and Michael Wachter. 1987. "Help-wanted advertising, job vacancies, and unemployment." *Brookings Papers on Economic Activity* 1987 (1): 207–248.

Alexander, J. Trent Alexander, Todd Gardner, Catherine G. Massey, and Amy O'Hara. 2014. "Creating a Longitudinal Data Infrastructure at the Census Bureau." : 1–14.

Anastasopoulos, L. Jason, George Borjas, Gavin Cook, and Michael Lachanski. 2019. "Job Vacancies and Immigration: Evidence from Pre- and Post-Mariel Miami.".

Anastasopoulos, L. Jason, Tima T. Moldogaziev, and Tyler A. Scott. 2017. "Computational Text Analysis for Public Management Research: An Annotated Application to County Budgets." : 1–47.

Andolfatto, David. 1996. "Business cycles and labor-market search." *The American Economic Review*: 112–132.

Atalay, Enghin, Phai Phongthiengtham, Sebastian Sotelo, and Daniel Tannenbaum. 2018. "New technologies and the labor market." *Journal of Monetary Economics* 97: 48–67.

Athey, Susan, Juliet B. Tibshirani, and Stefan Wager. 2019. "GENERALIZED RANDOM FORESTS." *The Annals of Statistics* 47 (2): 1148–1178.

Bana, Sarah. 2018. "Identifying Vulnerable Displaced Workers: The Role of Local Occupation Conditions.".

Barberá, Pablo. 2014. "Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data." *Political Analysis* 23 (1): 76–91.

Barnichon, Regis. 2010. "Building a composite help-wanted index." *Economics Letters* 109 (3): 175–178.

Berman, Eli. 1997. "Help wanted, job needed: Estimates of a matching function from employment service data." *Journal of labor economics* 15 (1, Part 2): S251–S292.

Blei, David M, and John D Lafferty. 2007. "A correlated topic model of science." *The Annals of Applied Statistics*: 17–35.

Bodvarsson, Örn, Hendrik Van den Berg, and Joshua J. Lewerd. 2008. "Measuring immigration's effects on labor demand: A reexamination of the Mariel Boatlift." *Labour Economics* 15 (4): 560–574.

Bond, Brittany, Brown J. David, Adela Luque, and Amy O'Hara. 2014. "The Nature of the Bias When Studying Only Linkable Person Records: Evidence from the American Community Survey." : 1–30.

Bond, Robert, and Solomon Messing. 2015. "Quantifying social medias political space: Estimating ideology from publicly revealed preferences on Facebook." *American Political Science Review* 109 (1): 62–78.

Borjas, George. 2017. "The Wage Impact of the Marielitos: A Reappraisal." *Industrial and Labor Relations Review* 70 (5): 1077–1110.

Burch, Susan W, and Ruth A Fabricant. 1968. "A further comment on the behavior of help-wanted advertising." *The Review of Economics of Statistics*: 278–281.

Burch, Susan W, and Ruth A Fabricant. 1971. "Cyclical Behavior of Help-Wanted Index and the Unemployment Rate: A Reply." *The Review of Economics of Statistics* 53 (1): 105–106.

Card, David. 1990. "The impact of the Mariel Boatlift on the Miami labor market." *Industrial and Labor Relations Review* 43: 245-57.

Chalfin, Aaron, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan. 2016. "Productivity and Selection of Human Capital with Machine Learning." *American Economic Review* 106 (5): 124–127.

Cohen, Malcolm, and Robert Solow. 1970. "The Behavior or Help-Wanted Advertising: A Reply." *The Review of Economics of Statistics* 52 (4): 442–443.

Cohen, Malcolm S., and Robert M Solow. 1967. "The behavior of help-wanted advertising." *The Review of Economic and Statistics*: 108–110.

Denny, Matthew J., and Arthur Spirling. 2018. "Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It." *Political Analysis* 26 (2): 168189.

Flood, Sarah, Miriam King, Renae Rodgers, Steven Ruggles, and J. Robert Warren. 2018. "Integrated Public Use Microdata Series, Current Population Survey.". *Version 6.0 [dataset].* Minneapolis, MN: IPUMS, https://doi.org/10.18128/D030.V6.0.

Geloso, Vincent, and Vadim Kufenko. 2017. "THE DEMAND FOR "GUARD LABOR": ANOTHER EXPLANATION." *Economic Affairs* 37 (3): 373–381.

Gentzkow, Matthew, Bryan T. Kelly, and Matt Taddy. 2017. Text as data. Technical report National Bureau of Economic Research.

Grimmer, Justin. 2009. "A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases." *Political Analysis* 18 (1): 1–35.

Grimmer, Justin, and Brandon M Stewart. 2013. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political analysis* 21 (3): 267–297.

Gujarati, Damodar. 1969. "Cyclical Behavior of Help-Wanted Index and the Unemployment Rate." *Review of Economics and Statistics* 51 (4): 482–484.

He, Yulan, Chenghua Lin, Wei Gao, and Kam-Fai Wong. 2013. "Dynamic joint sentiment-topic model." *ACM Transactions on Intelligent Systems and Technology (TIST)* 5 (1): 6.

Hinton, Geoffrey E, and Ruslan R Salakhutdinov. 2009. "Replicated softmax: an undirected topic model." In *Advances in neural information processing systems.* pp. 1607–1614.

Jarmin, Ronald. 2019. "Evolving Measurements for an Evolving Economy: Thoughts on 21st Century US Economic Statistics." *Journal of Economic Perspectives* 33 (1): 165–184.

Jayadev, Arjun, and Samuel Bowles. 2006. "Guard Labor." *Journal of Development Economics* 79 (2): 328–348.

Jegadeesh, Narasimhan, and Di Wu. 2013. "Word power: A new approach for content analysis." *Journal of Financial Economics* 110 (3): 712–729.

Kroft, Kory, and Deving G. Pope. 2014. "Does Online Search Crowd Out Traditional Search and Improve Matching Efficiency? Evidence from Craigslist." *Journal of Labor Economics* 32 (2): 259–303.

Lachanski, Michael, and Steven Pav. 2017. "Shy of the Character Limit." *Econ Journal Watch* 14 (3): 302–346.

Lee, W. 2016. "Slack and slacker: Job seekers, job vacancies, and matching functions in the U.S. labor market during the roaring twenties and the great contraction, 1924-1932." *The Journal of Economic History* 76 (3): 840–873.

Lowe, Will, Kenneth Benoit, Slava Mikhaylov, and Michael Laver. 2011. "Scaling policy preferences from coded political texts." *Legislative studies quarterly* 36 (1): 123–155.

Manela, Asaf, and Alan Moreira. 2017. "News implied volatility and disaster concerns." *Journal of Financial Economics* 123 (1): 137–162.

Monroe, Burt L, Michael P Colaresi, and Kevin M Quinn. 2008. "Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict." *Political Analysis* 16 (4): 372–403.

Pater, Robert. 2017. "Is there a Beveridge curve in the short and the long run?" *Journal of Applied Economics* 20 (2): 283–303.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg et al. 2011. "Scikit-learn: Machine learning in Python." *Journal of machine learning research* 12 (Oct): 2825–2830.

Preston, Noreen L. 1977. *The Help-Wanted Index : Technical Description and Behavioral Trends*. Conference Board.

Saiz, Albert. 2003. "Room in the Kitchen for the Melting Pot: Immigration and Rental Prices." *Review of Economics and Statistics* 85 (3): 502–521.

Seah, Kelvin. 2018. "The effect of immigration shocks on native fertility outcomes: evidence from a natural experiment." *IZA Journal of Development and Migration* 8 (18): 1–34.

Shimer, Robert. 2005. "The cyclical behavior of equilibrium unemployment and vacancies." *American Economic Review* 95 (1): 25–49.

Shimer, Robert. 2007. "Mismatch." *American Economic Review* 97 (4): 1074–1101.

Yasenov, Vasil, and Giovanni Peri. 2018. "The Labor Market Effects of a Refugee Wave: Synthetic Control Method Meets the Mariel Boatlift." *Journal of Human Resources*: 0217_8561R1.

Young, Lori, and Stuart Soroka. 2012. "Affective news: The automated coding of sentiment in political texts." *Political Communication* 29 (2): 205–231.

Zagorsky, Jay L. 1993. "Understanding the economic factors affecting help-wanted advertising." *Journal of Advertising* 22 (3): 75–84.

Zagorsky, Jay L. 1998. "Job vacancies in the United States: 1923 to 1994." *Review of Economics and Statistics* 80 (2): 338–345.

7 Appendix

Table 1: Metro Area Newspapers Sampled by the Conference Board

Metro Area	**Paper Used for HWI Since 1970**
Albany	*The Times Union*
Allentown	*Allentown Morning Call*
Atlanta	*Atlanta Constitution**
Baltimore	*Baltimore Sun*
Birmingham	*Birmingham News*
Boston	*Boston Globe*
Charlotte	*Charlotte Observer*
Chicago	*Chicago Tribune*
Cincinnati	*Cincinnati Enquirer*
Cleveland	*Cleveland Plain Dealer*
Columbus	*Columbus Dispatch*
Dallas	*Dallas Times Herald***
Dayton	*Dayton Daily News*
Denver	*Denver Rocky Mountain News*
Detroit	*The Detroit News*
Gary	*Gary Post-Tribune*
Hartford	*Hartford Courant*
Houston	*Houston Chronicle*
Indianapolis	*Indianapolis Star*
Jacksonville	*Florida Times-Union*
Kansas City	*Kansas City Star*
Knoxville	*Knoxville News-Sentinel*
Los Angeles	*Los Angeles Times*
Louisville	*Louisville Courier-Journal*
Memphis	*Memphis Commercial Appeal*
Miami	*Miami Herald*
Milwaukee	*Milwaukee Sentinel*
Minneapolis	*Minneapolis Star Tribune*
Nashville	*Nashville Tennessean*

Continued on next page

Table 1 – continued from previous page

Metro Area	**Paper Used for HWI Since 1970**
New Orleans	*The Times-Picayune*
New York	*New York Times*
Newark	*Newark Evening News*
Oklahoma City	*The Daily Oklahoman****
Omaha	*Omaha World-Herald*
Philadelphia	*Philadelphia Inquirer*
Phoenix	*Phoenix Arizona Republic*
Pittsburgh	*Pittsburgh Post-Gazette*
Providence	*Providence Journal*
Richmond	*Richmond Times-Dispatch*
Rochester	*Rochester Times-Union*
Sacramento	*Sacramento Bee*
Salt Lake City	*Salt Lake Tribune*
San Antonio	*San Antonio Express-News*
San Bernardino	*San Bernardino Sun*
San Diego	*San Diego Union*
San Francisco	*San Francisco Examiner*
Seattle	*Seattle Post-Intelligencer*
St. Louis	*St. Louis Post-Dispatch*
Syracuse	*Syracuse Herald Journal*
Toledo	*Toledo Blade*
Tulsa	*Tulsa World*
Washington, D.C.	*Washington Post*

*Replaced by the *Atlanta-Journal Constitution* in 1982.
**Replaced by the *Dallas Morning News* in 1991.
***We have been unable to confirm that the surveyed paper was the *Daily Oklahoman*.

Table 2: U.S. 2000 Decennial Long Form Tabulation: Year of Entry for Cuban-born Respondents (Weighted).

Year	Entrants	Year	Entrants
$\leq$ 1935	2924	1968	38920
1936	141	1969	35740
1937	203	1970	38050
1938	231	1971	28370
1939	275	1972	11070
1940	633	1973	9982
1941	374	1974	11750
1942	404	1975	3879
1943	467	1976	2760
1944	710	1977	2527
1945	1294	1978	4510
1946	1373	1979	12950
1947	1522	1980	100600
1948	1699	1981	7467
1949	1417	1982	6076
1950	2101	1983	6847
1951	1406	1984	9309
1952	1742	1985	9687
1953	2147	1986	7523
1954	3737	1987	6362
1955	5647	1988	12440
1956	7679	1989	12680
1957	6706	1990	14030
1958	5537	1991	13540
1959	9990	1992	13860
1960	33460	1993	17600
1961	45080	1994	26810
1962	45310	1995	51000
1963	11860	1996	20470
1964	9614	1997	16880
1965	17480	1998	23040
1966	30850	1999	30730
1967	36350	2000	7812

All numbers preserve only four significant figures to comply with DRB regulations.

Table 3: Number of Likely Marielitos (Weighted) Filing IRS Form 1040s by Geographic Areas for Selected Years: 1970, 1975, 1980, 1985, 1990, and 1995.

Region	1970	1975	1980	1985	1990	1995
New England (CT, ME, MA, NH, RI, and VT)	20	60	20	500	300	350
Mid-Atlantic (NJ, NY, and PA)	250	450	600	6600	5700	5200
East North Central (IL, IN, MI, OH, and WI)	80	150	100	850	800	950
West North Central (IA, KS, MN, MS, NE, ND, and SD)	20	N< 15	40	150	250	300
South Atlantic (DE, FL, GA, MD, NC, SC, VA, DC, and WV)	450	1400	1800	34500	43000	47000
East South Central (AL, KY, MS, and TN)	(D)	(D)	(D)	90	60	90
West South Central (AR, LA, OK, and TX)	60	100	150	950	800	1000
Mountain (AZ, CO, ID, MT, NV, NM, UT, and WY)	40	30	30	450	550	600
Pacific (AK, CA, HI, OR, and WA)	40	150	200	2500	2500	2300
United States	950	2300	3000	46500	54000	58000
Selected States						
New Jersey	100	200	300	4400	3900	3500
New York	150	150	250	2000	1600	1400
Florida	450	1300	1700	34000	42000	46500
Texas	(D)	(D)	(D)	600	450	600
California	(D)	150	200	2400	2400	2200
Miami MSA	400	1200	1500	31000	39000	42000

Table constructed by linking US 2000 Decennial Census to IRS 1040 filings. Because of DRB-mandated rounding, region cells may not add up to total national filers by year. DRB rules require suppression of cell output where individual disclosure risk is high. Suppressed cells have had their contents replaced by "(D)" in this table to reflect that they were not judged releasable by the DRB.

Table 4: 1/29 Sample Used for Supervised and Unsupervised Learning.

Dates Sampled Before the Mariel Boatlift	Dates Sampled After the Mariel Boatlift
1/4/1978	1/18/1983
2/2/1978	2/16/1983
3/3/1978	3/17/1983
4/1/1978	4/15/1983
4/30/1978	5/14/1983
5/29/1978	6/12/1983
6/27/1978	7/11/1983
7/26/1978	8/9/1983
8/24/1978	9/7/1983
9/22/1978	10/6/1983
10/21/1978	11/5/1983
11/19/1978	12/3/1983
12/18/1978	1/1/1984
1/16/1979	1/30/1984
2/14/1979	2/28/1984
3/15/1979	3/28/1984
4/13/1979	4/26/1984
5/12/1979	5/25/1984
6/10/1979	6/23/1984
7/9/1979	7/22/1984
8/7/1979	8/20/1984
9/5/1979	9/18/1984
10/4/1979	10/17/1984
11/2/1979	11/15/1984
12/1/1979	12/14/1984
12/30/1979	

We selected dates by picking two anchor days for the before and after-Mariel periods. Our post-Mariel day, selected by selecting a random number between two and three years after the start of the boatlift was December 20, 1982. Once that date was determined, we worked backwards to pick a similarly situated anchor date in the years before Mariel: December 18, 1978. From each anchor date, we sampled every date 29 days away in the selected years 1978, 1979, 1983, and 1984. This procedure achieved the following objectives: (1) it ensured all dates were selected from economic expansions to minimize contamination from business cycle effects, (2) it ensured at least one date was selected for each calendar month, (3) it allowed us to build a sample more than an order of magnitude larger than our previous draft, and (4) because the sample rotates the day-of-the-week chosen, it ensures neither the before nor after Mariel sample is biased by selecting too many of any one day-of-the-week, e.g. Mondays.

Table 5: Summary of results from supervised and unsupervised machine learning exercises.

Unsupervised Machine Learning	1978 - 1979	1983 - 1984	Difference
Topics			
Food services	16.6	14.8	-1.90 (0.35)
Secretarial	13.5	13.5	0.02 (0.33)
Sales/management	9.6	11.7	2.04 (0.30)
Accounting	7.4	10.8	3.36 (0.28)
Automotive	15.2	12.2	-3.04 (0.32)
Engineer	6.3	9.1	2.78 (0.26)
Medical	7.6	7.8	0.16 (0.25)
Not classifiable	23.6	20.3	-3.43 (0.40)
Supervised Machine Learning			
% of Ads for the Less-Educated	43.9	36.3	-7.62 (0.47)

Robust standard errors in parentheses. The sample consists of a random sample of 46,072 help-wanted classifieds that appeared in the *Miami Herald* in the 1978, 1979, 1983, and 1984 calendar years.

Table 6: Truncated $k = 10$ topic model with relevant topic labels.

Food Service	Secretarial	Sales and Management	Accounting	Automotive	Engineering	Medical
person	offic	sale	experi	experienc	resum	time
appli	type	train	exp	pay	year	full
experienc	secretari	will	account	mechan	experi	posit
hotel	good	manag	oper	need	requir	part
refer	clerk	commiss	must	work	send	avail
cook	bilingu	new	help	top	manag	servic
restaur	skill	look	bookkeep	man	engin	call
waitress	general	career	need	must	system	shift
need	call	store	necessari	driver	seek	day
live	phone	earn	assist	shop	respons	hospit

Table 7: Full k = 10 topic model.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
call	person	offic	sale	experi	experienc	benefit	resum	time	opportun
must	appli	type	train	exp	pay	salari	year	full	employ
area	experienc	secretari	will	account	mechan	excel	experi	posit	equal
week	hotel	good	manag	oper	need	good	requir	part	insur
need	refer	clerk	commiss	must	work	open	send	avail	paid
hour	cook	bilingu	new	help	top	compani	manag	servic	personnel
dade	restaur	skill	look	bookkeep	man	work	engin	call	includ
licens	waitress	general	career	need	must	call	system	shift	offer
car	need	call	store	necessari	driver	start	seek	day	excel
per	live	phone	earn	assist	shop	immedi	respons	hospit	pleas

Table 8: Who Benefits from an Additional Job Opening?: Evidence from the CPS-ASEC and JOLTS, 2000 - 2017

		Log (Vacancies) Interacted With		
	Log (Vacancies)	Less than High School	High School	Some College
1. Log weekly wage	0.0598	0.0390	0.0029	0.0022
	(0.0302)	(0.0145)	(0.0108)	(0.0116)
2. Employment propensity	0.0145	0.0128	-0.0111	-0.0055
	(0.0169)	(0.0086)	(0.0051)	(0.0052)

Robust standard errors in parentheses. The number of observations of in each regression is 288 in the CPS-ASEC. The wage variable gives the age-adjusted mean wage in the year-city-education cell for workers aged 25-59. The employment variable in the March CPS gives the weighted number of persons who worked at some point during the calendar year for individuals aged 19 to 64. The individual age-and-sex regressions are weighted by the standard ASEC weight for the ASEC while the regressions in this table are weighted by the number of observations. Both models are linear, but employment results do not significantly differ for a probit specification.