7 Analyzing H1-B demand across firms

7.1 Motivation and background

Demand for foreign workers from US employers has fluctuated with economic cycles over the last twenty years and reflects a wide range of employers needs for high-skilled temporary workers. The H- 1B visa program is the primary mechanism through which US firms satisfy this demand, as it allows them to temporarily employ foreign workers on a nonimmigrant basis in specialty occupations. There are several reasons why employers use the H-1B visa program to hire workers. To be competitive in the global economy, firms must possess high-skilled labor forces with strong capabilities in science, technology, engineering and mathematics. However, the supply of these skills is limited and the demand for highly skilled workers is unevenly distributed and poorly matched geographically to the supply. As a result, US employers have difficulty sourcing and recruiting qualified domestic workers. Obtaining an employment-based green card for foreign workers can also take several years, whereas the approval process for an H-1B generally takes only a few months.

Before an employer can file an H-1B petition with US Citizen and Immigration Services, they must submit a Labor Condition Application (LCA) to the US Department of Labor. The LCA requires the employer to determine the prevailing wage for the position in the geographic area and the actual wage paid by the employer to other individuals with similar experience and qualifications for that type of work. The higher of the prevailing wage and the actual wage must be paid to the H-1B worker. The Department of Labor reviews the LCA and certifies the application within seven days of the filing of the application as long as it is complete and contains no obvious inaccuracies. The Department of Labor makes this LCA data available to the public on a quarterly basis, including employers names and locations, wage rates offered, number of H-1Bs sought, the occupations in which the H-1Bs will be employed, and whether the LCAs were certified or denied. It is important to note that the LCA data contains records for every request submitted, but does not contain the final outcome of each LCA. There are many more LCAs filed than H-1Bs granted. In addition, an LCA is submitted for every H-1B request, whether new or a renewal. Despite these points, the LCA data is the best available measure of the total demand for H-1B workers. The act of filing an LCA for a worker accurately measures the employers demand for H-1B labor and this data describes the flow of H-1B demand. Therefore, it can be used to make inferences and predictions with regards to the economic health and outlook of companies, industries and geographic areas of the United States.

7.2 Dataset

The dataset consists of LCAs filed with the Department of Labor from January 2002 through June 2017. In addition to H-1B LCAs, the data also includes LCAs for E-3 and H-1B1 visas and these should be included in the analysis.

LCA data prior to 2008 and the relevant schema are located here:
http://www.flcdatacenter.com/CaseH1B.aspx
LCA data and associated schemas from 2008 onwards are located here:
https://www.foreignlaborcert.doleta.gov/performancedata.cfm

7.3 Project Overview

The project will consist of the following phases:

1. Data acquisition, parsing and cleaning

   The students should write a script to download all of the LCAs from the Department of Labor websites. The LCA files will then need to be parsed according to their specification and aggregated into a tabular data structure. Fields should be validated and transformed into appropriate datatypes, and missing and erroneous values should be identified. Duplicate records should also be filtered. Statistical techniques should be employed to flag outliers and potential errors. Special attention should be paid to employer names which may vary over time or even filing to filing. Employer names should be matched to a unique, permanent identifier.

2. Exploratory data analysis and model development

   The students should conduct exploratory data analysis on the parsed and cleaned LCAs to better understand the characteristics of the dataset. This should include examining a number of summary statistics in the cross-section and time-series across various dimensions such as work city, work state, employer industry (NAICS), and occupation. In particular, the students should look for structural changes in LCA filings due to changes in the regulatory and/or political environment which could otherwise impact the analysis. For example, in October 2000 the H-1B visa cap was raised to 195,000 for fiscal years 2001, 2002 and 2003. In 2004, the cap returned to 65,000 with 20,000 additional visas for holders of advanced degrees.

   Once they are comfortable with the data, the students will propose and develop two models and/or methodologies:

   (a) Identifying acceleration/deceleration in H-1B demand

   This model should identify acceleration or deceleration in the growth or contraction of aggregate H1-B demand across industries, occupations and geographic locations on a fiscal year basis. Structural trends, seasonality and autocorrelation should be accounted for in the model. Statistics such as ACFs and PACFs should be examined to determine the time-series properties of H1-B demand across a particular dimension. The model should be able to produce output as of a particular point-in-time without look-ahead bias.

   (b) Identifying firms with similar occupational demand

   Using historical hiring patterns for each employer and the Standard Occupational Classification (SOC) system, the model should identify employers which have similar occupational demand according to LCA disclosures. Specifically, for N employers, the model should produce a symmetric N x N matrix where each element is a similarity score between 0 and 1, with 0 signifying no similarity and 1 signifying perfect similarity. The model should be able to produce output as of a particular point-in-time without look-ahead bias.

3. Implementation and evaluation

   The students should implement the project in a programming language adept at statistical analysis (Python is preferred). The students will need to write a script to retrieve the raw LCA data, preprocess it and transform it into a format that can be readily processed by the modelling code. They should develop evaluation criteria that will allow them to benchmark their models and methodologies. They should iterate over phase 2 and 3 appropriately.

7.4 Research Goals

The primary goal of the project is to identify occupations, industries and geographical locations that are experiencing above or below trend growth or contraction in H1-B demand. The secondary goal of the project is to identify firms that have similar H-1B demand across occupational lines. We encourage the students to be creative and go beyond these basic requirements.

7.5 Suggested Output

For each fiscal year end (September 30th), four files should be produced. The first file should list each industry (identified by NAICS) along with a cross-sectional z-score that represents the acceleration in growth for that industry for the fiscal year. The second file and third files should contain similar output for occupations and geographic locations respectively. The fourth file should output the employer similarity matrix as of the fiscal year end.

7.6 Follow-Ons

A potential follow-on is a visualization tool that will allow the user to interactively explore summary statistics across various dimensions in the time-series and cross-section as well as the trends and linkages identified by the models developed over the course of the project.

7.7 Mentor information

Dennis Walsh
Managing Director
Goldman Sachs