BGT technical document

*CAVEAT I’M ONLY WORKING ON MAIN TABLE I ONLY USED A FEW COLUMNS FROM MAIN*

## Table of Contents

1. Introduction
2. Burning Glass Technology (BGT)
   * Background
   * Review of Academic Literature
   * Data Review
     + BGT Job-Ad Data
     + Data Profiling
     + Exploratory Data Analysis
3. Virginia Open Data/Open Jobs Data
   * Background
   * Data Analysis
     + Data Profiling
     + Exploratory Data Analysis
4. BGT Job-ad Data Validation
   * Validation Results in the Literature
   * Comparison with Virginia Open Data/Open Jobs Data
   * Conclusions
5. Recommendations

## Introduction

This technical document evaluates the feasibility of using the real-time labor market information (LMI) collected by Burning Glass Technology to supplement survey and administrative data collected by federal and state governments. In contrast to designed and administrative data which have varying lag times between collection and dissemination and are often aggregated over broad occupation categories, the real-time job ads collected by Burning Glass Technology are made available within a day of the job-ad being posted and provide information at a granular level that links skill set requirements to specific occupations within broad occupation categories. In this document, the data are evaluated for use in identifying the skills sets necessary for a job in the skilled technical workforce and how these skill sets and employer demands vary within a state and across the country.

This document reviews the use of the BGT job-ads in academic research highlighting issues with the data and any validation results. It provides the results of profiling and exploratory data analyses, for both the BGT job-ads data and the Virginia Open Data/Open Jobs Data which is used to validate the BGT job-ads data. The document concludes with recommendations regarding fitness-for-use.

## Burning Glass Technology

#### Company Background

Burning Glass Technologies is a Boston-headquartered labor market analytics firm founded in 1999, that uses artificial intelligence to collect and host a massive repository of workforce and employment data. Burning Glass collects these data primarily for commercial purposes and secondarily for research purposes. The company markets both the data as a product as well as their consulting expertise on labor market questions to customers across a variety of industries, education institutions, local and regional governments, recruiting and staffing agencies, and corporate firms.

The company makes use of text-parsing and predictive matching methods to extract and aggregate LMI data, currently from over 40,000 sources (the number of websites has changed over the years). Data has been extracted since 2007 with the exception of 2008 and 2009. Job boards, corporate websites, and other places where job ads are posted are scanned daily and algorithms are used to identify and remove duplicate postings (close to 80% of all postings collected are marked as duplicates.) From the job-ad, the BGT software extracts the title, occupation, employer, and location, and uses natural-language technology to identify specific occupations, skills, and qualifications from the job description. Both the algorithms used to identify duplicates and scrape websites are proprietary as well as the addresses of the website that are scraped.

#### Review of Academic Literature

Favorite articles:

* Deming, David, and Lisa B. Kahn. “Skill Requirements across Firms and Labor Markets: Evidence from Job Postings for Professionals.” Journal of Labor Economics 36, no. S1 (2018): S337-S369.
* Liu, Yukun. “Vacancy Postings, Skill Requirements, and the Cross-Sectional Return Predictability.”
* Jeffrey Clemens, Lisa B. Kahn, and Jonathan Meer. “Dropouts Need Not Apply: The Minimum Wage and Skill Upgrading.” (Sept 2018)
* Paul Mohnen, Enrico Berkes, and Bledi Taska “Can Skill Mismatch Explain Geographic and Time Variation in the Returns to College Majors? Evidence from Online Job Postings” (2017) Northwestern University
* Papageorgiou, Theodore. “Worker sorting and agglomeration economies.” McGill University (2017)
* Wardrip, Keith, Stuart T. Andreason, and Mels De Zeeuw. “Uneven Opportunity: Exploring Employers’ Educational. Preferences for Middle-Skills Jobs.” Special Report of the Federal Reserve Banks of Philadelphia, and Atlanta (2017).
* Fuller, Joseph B., Jennifer Burrowes, Manjari Raman, Dan Restuccia, and Alexis Young. “Bridge the Gap: Rebuilding America’s Middle Skills.” Report, U.S. Competitiveness Project, Harvard Business School, November 2014
* Hershbein, Brad and Lisa B. Kahn (2018) Do Recessions Accelerate Routine-Biased Technological Change? Evidence from Vacancy Postings. American Economic Review 108(7): 1737-1772. \*\*cited all the time
* Templin, Thomas, and Lesley Hirsch. 2013. “Do Online Job Ads Predict Hiring?” New York: New York City Labor Market Information Services. \*\* validation example

#### Data Review

##### BGT Job-Ad Data

At the data inventory stage, potential data sources (identified in the data discovery stage) are further screened to determine if they would support the research questions. Specifically, the screening process involves assessing data sources on six factors including purpose, method, description, timeliness, selectivity, and accessibility.

* Purpose: Burning Glass collects this data primarily for commercial purposes and secondarily for research purposes. The company markets both the data as a product as well as their consulting expertise on labor market questions to customers across a variety of industries, such as higher education, local and regional government, recruiting and staffing agencies, and other corporate firms.
* Method: Burning Glass scrapes over 40,000 websites to collect data from online job postings. The data is cleaned and deduplicated to present a national view of the labor market landscape across time.
* Description: The data is centered around individual positions, with one-to-many related tables for requirements for skills, certification, and education. While the data includes various types of data, the bulk of this data is text-based.
* Timeliness: The data is collected in real-time and covers 2007 and 2010-2017. [How soon after collection is data available?]
* Selectivity: The data is intended to represent the universe of all US jobs posted online over the years mentioned above.
* Accessibility: The data is accessible via file-transfer protocol, but a data-sharing agreement may fetter its accessibility outside of the lab staff.

The structure of this data includes 8 discrete columns and 5 continuous ones.

|  |  |
| --- | --- |
| Metric | Value |
| rows | 63610 |
| columns | 13 |
| discrete\_columns | 8 |
| continuous\_columns | 5 |
| all\_missing\_columns | 0 |
| total\_missing\_values | 904 |
| complete\_rows | 63158 |
| total\_observations | 826930 |
| memory\_usage | 9902352 |

##### Data Profiling

This section profiles the Burning Glass data, reviews its quality, and seeks to determine its useability on the following metrics.

* Completeness.
* Value validity.
* Consistency.
* Uniqueness.
* Duplication.

Here, we focus on the postings themselves which represent the central table of the data.

This first table summarizes the postings dataset. Of note, this shows that 99.2894199 % of the data are complete cases.

This second table breaks down the postings dataset by column using a superficial check for blanks. Of note, this shows that only two columns have blanks: the FIPS county and FIPS value columns are both missing 452 observations. It is important to note, however, this table qualifies missing values as blanks or NA only.

|  |  |  |
| --- | --- | --- |
| Column | # Missing | % Missing |
| bgtjobid | 0 | 0.0000000 |
| jobdate | 0 | 0.0000000 |
| occfam | 0 | 0.0000000 |
| occfamname | 0 | 0.0000000 |
| employer | 0 | 0.0000000 |
| city | 0 | 0.0000000 |
| state | 0 | 0.0000000 |
| county | 0 | 0.0000000 |
| fipsstate | 0 | 0.0000000 |
| fipscounty | 452 | 0.0071058 |
| fips | 452 | 0.0071058 |
| lat | 0 | 0.0000000 |
| lon | 0 | 0.0000000 |

We would like also to capture invalid values in addition to the missing ones. This last table captures blanks, NA, and ‘na’ text values (Burning Glass entered). Here we see that more missing values are captured for the County column and new invalid values are captured for the Occupation Family code, Occupation Family Name, Employer, and City columns.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Column | Example | Class | Blanks | % Blank | NA | % NA | na | % na | Unique |
| bgtjobid | bgtjobid | 38247599675 | character | 0 | 0.0 | 0 | 0.0 | 0 | 0.0 | 63610 |
| jobdate | jobdate | 2017-07-07 | character | 0 | 0.0 | 0 | 0.0 | 0 | 0.0 | 31 |
| occfam | occfam | 29 | character | 0 | 0.0 | 0 | 0.0 | 2148 | 3.4 | 24 |
| occfamname | occfamname | Healthcare Practitioners and Technical Occupations | character | 0 | 0.0 | 0 | 0.0 | 2148 | 3.4 | 24 |
| employer | employer | na | character | 0 | 0.0 | 0 | 0.0 | 13546 | 21.3 | 6309 |
| city | city | Leesburg | character | 0 | 0.0 | 0 | 0.0 | 430 | 0.7 | 801 |
| state | state | Virginia | character | 0 | 0.0 | 0 | 0.0 | 0 | 0.0 | 1 |
| county | county | Loudoun | character | 452 | 0.7 | 0 | 0.0 | 0 | 0.0 | 132 |
| fipsstate | fipsstate | 51 | integer | 0 | 0.0 | 0 | 0.0 | 0 | 0.0 | 1 |
| fipscounty | fipscounty | 107 | integer | 452 | 0.7 | 452 | 0.7 | 0 | 0.0 | 132 |
| fips | fips | 51107 | integer | 452 | 0.7 | 452 | 0.7 | 0 | 0.0 | 132 |
| lat | lat | 39.0845 | numeric | 0 | 0.0 | 0 | 0.0 | 0 | 0.0 | 952 |
| lon | lon | -77.6066 | numeric | 0 | 0.0 | 0 | 0.0 | 0 | 0.0 | 989 |

We also perform a check for consistency, uniqueness, and duplication, and find that 0 have multiple dates associated with them. We can be reasonably confident the data is consistent and unique.

##### Exploratory Data Analysis

These histograms show the top value distribution for each column in the dataset (except the identifier column). For any variables with more than 50 possible values, only the top 20 values are shown below for clarity.

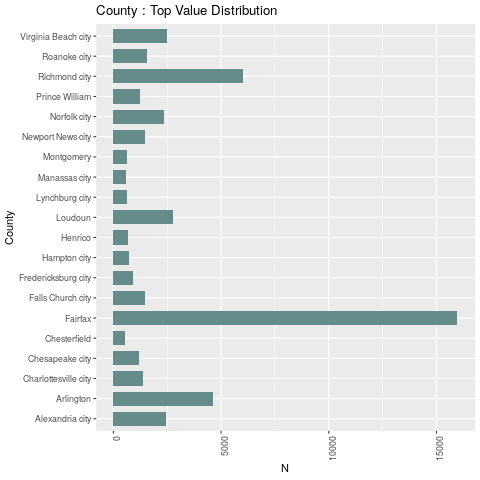


Figure 1. BGT Jobs in Virginia during July 2017 by County.

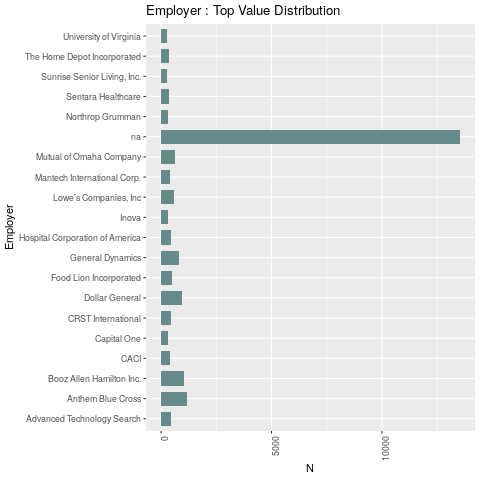


Figure 2. BGT Jobs in Virginia during July 2017 by Employer

The employer is not well populated, at 20% of the values appearing at ‘na.’ As seen below, the ‘na’ values are the most common value by far, followed by Anthem Blue Cross and Booz Allen Hamilton Inc. Fairfax County is the county with the greatest number of jobs. Richmond and Arlington follow with less than half.

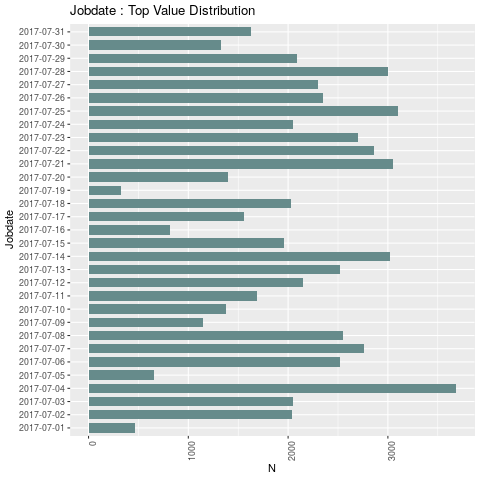


Figure 3. BGT Jobs in Virginia during July 2017 by Date

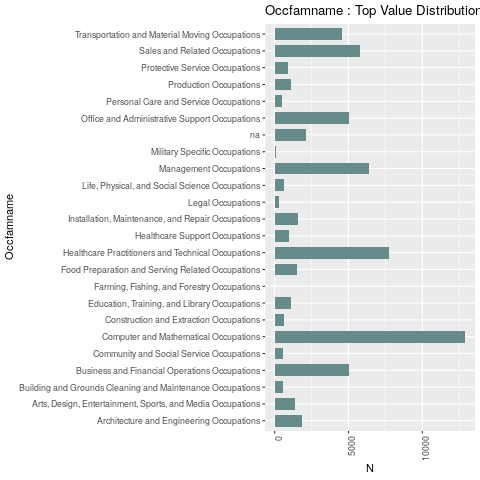


Figure 4. BGT Jobs in Virginia during July 2017 by Occupation Family

Throughout the month, there does appear to be a few spikes in jobs, the highest of which taking place on July 4, 2017. The next 3 days have comparable job numbers: July 25, July 21, and July 14. Throughout the month, there does appear to be a few spikes in jobs, the highest of which taking place on July 4, 2017. The next 3 days have comparable job numbers: July 25, July 21, and July 14.

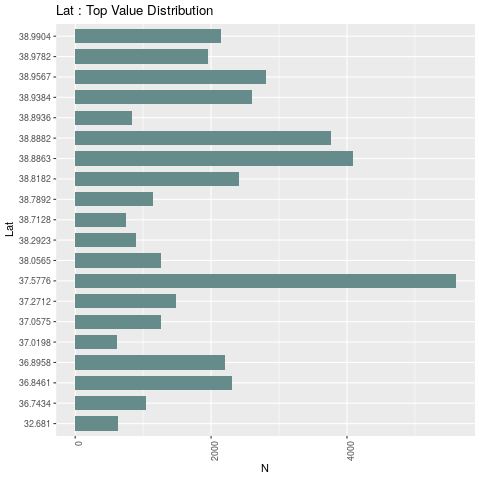


Figure 5. BGT Jobs in Virginia during July 2017 by Lat

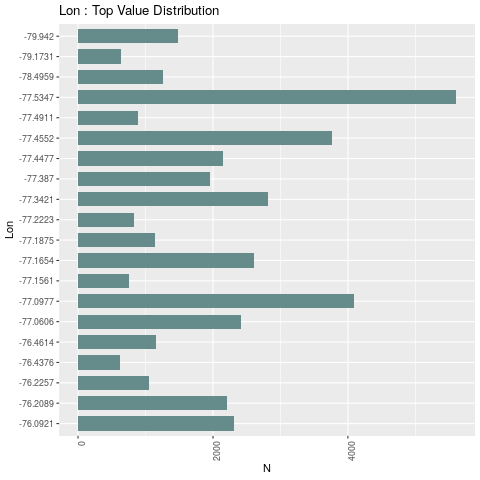


Figure 6. BGT Jobs in Virginia during July 2017 by :pm

It may be interesting to note that the top three coordinates given in the last two plots for the Latitude and Longitude columns, (37.5776, -77.5347), (38.8863, -77.0977), and (38.8882, -77.4552), respectively represent Richmond, Arlington (Clarendon), and Chantilly respectively.

#### Validation

1. BGT Job-ad Data Validation
   * Validation Results in the Literature
   * Comparison with Virginia Open Data/Open Jobs Data
   * Conclusions

FILTERING FOR VIRGINIA JULY 2017

The distribution of jobs throughout Virginia appears as expected, with coverage throughout all regions of Virginia and density surrounding the major urban areas. Notably, the increased density aligns with the areas around the DC metro area, the Norfolk/Virginia Beach area, and the city of Richmond.

If we look at this same map in aggregate across counties, we see the same density patterns (darker reds indicating greater density) except Blacksburg and Charlottesville also become notable areas of density.

Compare by Occupation Family

# A tibble: 23 x 4

# Groups: occfam [23]

occfam occfamname BGTcount OJcount 1 11 Management Occupations 6408 9796 2 13 Business and Financial Operations Occupations 5012 5701 3 15 Computer and Mathematical Occupations 12893 5313 4 17 Architecture and Engineering Occupations 1833 1997 5 19 Life, Physical, and Social Science Occupations 613 220 6 21 Community and Social Service Occupations 597 191 7 23 Legal Occupations 290 88 8 25 Education, Training, and Library Occupations 1110 571 9 27 Arts, Design, Entertainment, Sports, and Media … 1369 880 10 29 Healthcare Practitioners and Technical Occupati… 7743 3277 # … with 13 more rows

Compare by Economic Region

# A tibble: 10 x 4

# Groups: Region [10]

Region GOorg BGTcount OJcount 1 1 UVA Wise 1204 1132 2 2 VT Economic Development 3688 4141 3 3 Southside Planning District Commission 1547 1509 4 4 GROW Capital Jobs Foundation 9071 6914 5 5 Reinvent Hampton Roads 10126 8350 6 6 George Washington Regional Commission 2168 1193 7 7 Northern Virginia Regional Council 28690 20745 8 8 Northern Shenandoah Valley Regional Commission 2565 1964 9 9 Central Virginia Partnership for Economic Devel… 2807 1859 10 NA 1314 NA

Compare by Economic Region & County

# A tibble: 131 x 5

# Groups: Region, GOorg [10]

Region GOorg County BGTcount OJcount 1 1 UVA Wise Bland County 15 11 2 1 UVA Wise Bristol city 119 434 3 1 UVA Wise Buchanan County 48 15 4 1 UVA Wise Carroll County 49 15 5 1 UVA Wise Dickenson County 30 13 6 1 UVA Wise Galax city 76 67 7 1 UVA Wise Grayson County 27 15 8 1 UVA Wise Lee County 60 22 9 1 UVA Wise Norton city 34 34 10 1 UVA Wise Russell County 51 45 # … with 121 more rows

## Post Script? Questions Remaining for Burning Glass

* If occfam is missing, is occfamname always missing as well?
* what naming standardizations have been done to the employer column?
* How does burning glass handle subsidiaries?