Submitted on: 04.08.2015

Promoting Public Library Sustainability through Data Mining: R and Excel

Sarah Bratt
Syracuse University School of Information Studies, Syracuse, NY, USA
sebratt@syr.edu

Kusturie Moodley
University of Technology, Durban, South Africa
kmoodley@dut.ac.za

Abstract:

Information professionals have a vested interest in leveraging data to advocate for, justify, and support libraries' political and financial activities. This research explores New York State public library data by analyzing the economic and employment disparities among New York State public libraries, affording steps toward a greater balance in terms of data accessibility and transparency.

Acquiring and warehousing data is neither meaningful nor useful unless a workflow around data mining and analysis is established to ground assessment, recruiting, budgeting, decision-making, benchmarking, and community empowerment. The potential impacts of a dearth in best practices for quickly summarizing and interpreting public and enterprise data culminate not only in lost opportunities but also neglected resources.

This report documents the workflow and insights from an analysis of the Institute of Museum and Library Services' (IMLS) voluntary annual survey of public libraries in the United States from 2008-2011 using the statistical analysis tools R and MS Excel. The authors explored trends in New York State public libraries and found statistical correlations between library location, resources, and employee education with analysis steps that could be reproducible for libraries globally and nation-wide.

Libraries may ostensibly seem behind the curve in understanding how to quickly assess the community. Yet librarians can leverage local data and international trends to better serve their respective communities, taking business insights and transforming them into public library ethos.

Keywords: data mining, library & information science, R, knowledge management

---

Introduction

The Institute of Museum and Library Services (IMLS) conducts a voluntary annual survey of public libraries in the United States. The 2010-2011 survey is mandated by the Museum and Library Services Act of 2010 (PL 111-340) (Institute of Museum and Library Services, 2013). The 2008-2009 survey is mandated by the Museum and Library Services Act of 2003 (SEC 210) (Institute of Museum and Library Services, 2011). Four separate datasets from the years 2008-2011 were combined into a master data frame. When downloading the CSV from the IMLS website, the zip file includes 3 compressed files, and the largest size file available was chosen to ensure inclusivity.

Some variables of particular interest were the number of ALA MLS librarians, count of print and electronic materials, count of bookmobiles, and number of reference transactions, among others. Though the data sets' size varies by year, they are of approximately the same size: 9284 rows and 150 columns (before cleaning) and each about 756 rows and 25 columns (after cleaning and transformation). The combined master set is 3024 rows and 28 variables.*

* Original data files (in CSV format) and a text file of code are available upon request. Note that the code file does not have robust comments for ease of reproducibility. Robustly commented data files are available upon request.

Preparation/Cleaning

Data cleaning is essential as it ensures the integrity and improves the quality of the data (Mathew et al., 2014). Data cleaning involves identifying and removing errors, deciding which variables to retain and which ones to delete, and ensuring that the data is consistent (Ekbia et al., 2014).

After downloading the individual data sets (2008-2011) from the IMLS website in CSV format, the cleaning process began with the deletion of all rows with states other than New York. Because only 25 out of 150 variables were of interest, unnecessary columns (such as children's circulation statistics, specific program data, etc.) were deleted.

The variables retained were as follows: state (NY), legal basis, library name, total branch libraries, bookmobiles count, total hours worked by ALA-MLS librarians per week, total staff hours, total income, print materials expense, electronic materials expense, other materials expense, total materials expense, physical units and downloadable titles of audio/video/ ebooks, total licensed databases, annual reference transactions, registered borrowers, programs, program attendance, public internet computers count, use of public internet computers count, county population, and locale description (rural, town, suburb, city, etc.), latitude, and longitude

(Institute of Museum and Library Services, 2013). A column was added identifying the year of the dataset because the data used spanned from 2008-2011. A column populated was also added with the calculation of the % of hourly employment of ALA MLS librarians (=MASTER/TOTSTAFF). These variables were chosen because they are key indicators that inform the core question: Do libraries that employ more ALA MLS librarians per hour provide greater overall value for that library community, as indicated by resources and programming data?

Transformations:

The documentation and data dictionary were reviewed to change the column names of the 2010 and 2011 data so that it was consistent with previous years. Namely, the variable AUDIO_PH was changed to AUDIO and VIDEO_PH changed to VIDEO (Institute of Museum and Library Services, 2013).

Missing Data Mitigation: Of the attributes of interest, there were no missing values.

Once the data was cleaned and transformed, the data sets were combined into a master set using the rbind() function. A column called "Threshold" was added indicating "yes" or "no" as to whether a library was above the hourly ALA MLS hiring threshold.

Visualization Summary

Visualizations were produced with R; the hiring rate (%) dispersion was visualized with a boxplot and the hiring rate trend from 2008-2011 was displayed in a scatter plot. The frequency of locale type in libraries with a hiring rate of > 55% is shown in a simple bar chart for each year (2008-2011). Relative resource expenditures are visualized with a series of heat maps, as compared with reference questions, computer usage, and programming attendance counts among the libraries with the highest rate of ALA MLS hiring (i.e., those libraries with a rate above the established threshold). Finally, the rate of hourly employment of ALA MLS librarians is represented on a Google map, with locations plotted by latitude and longitude. Circle circumference indicates the hourly employment rate ALA MLS librarians (above the threshold).

Questions

This report investigates trends in the hourly rate of employment of ALA MLS librarians in NY state public libraries. Examples of questions specifically addressed are as follows:

1. What is the yearly trend of NYS libraries above the ALA MLS hourly employment threshold (55% hourly employed MSLIS librarians) in the years from 2008-2011?

2. In which NY locale is there the greatest hourly employment of ALA MLS librarians? What locales rank 2nd and 3rd in the hourly employment of ALA MLS librarians?

3. What are the electronic resources development, programming, print resources development, program attendance, ebook collection development, etc. practices of NYS libraries' over the (externally) established threshold from 2008-2011?

4. Where in NYS are ALA MLS librarians hired the most (% of total hourly employment rate) in 2011?

5. What attributes and characteristics describe (and "predict") the library systems with the greatest or lowest hourly employment of ALA MLS librarians?

Question 1: What is the yearly trend of NYS libraries above the ALA MLS hourly employment threshold (55% hourly employed MSLIS librarians) in the years from 2008-2011?

First, analysis revealed an increase in the yearly trend of hourly employment rate of ALA MLS librarians in NYS public libraries. The scatterplot displays an increasing trend of NY public libraries' hourly employment rates of ALA MLS librarians above the threshold (55%) within the years 2008-2011 (See Figure 1). The plot shows that in 2008, no libraries in NYS had 100% hourly employment of ALA MLS librarians. However, in the following years (2009-2011) there are data points that reach the 100% level. There is also a significant leap in increasing the hourly employment of ALA MLS librarians, as seen in the scatterplot columns of the years 2009, 2010, and 2011.

The key element of this chart is the jump from NYS boasting no libraries in 2008 with 100% hourly employment of ALA MLS librarians, to libraries *having* 100% hourly employment of ALA MLS librarians on their staff. Accordingly, an interpretation of this jump is that there might be an important shift in hiring and employment practices in NYS public libraries in recent years. Granted, 2008 might be an anomaly year because there are no public libraries in NYS in 2008 that had 100% of their staff with ALA MLS degrees. Nevertheless, libraries are generally stable employers. Therefore, it can be validly conjectured that such a jump is significant because it suggests that more value is placed on ALA MLS degree-holding librarians.

The implications extend beyond the public library sector. This shift in valuing ALA MLS degree-holding librarians impacts Library & Information Science schools because it suggests a viable and growing career option. The increased value targets a number of audiences: library directors who would hire personnel for their library, new job seekers who are looking for a position in the public library, currently employed library staff (without an ALA MLS degree) who aim to advance their career with a promotion, and library school faculty, staff, and recruitment administrators to market their institution.

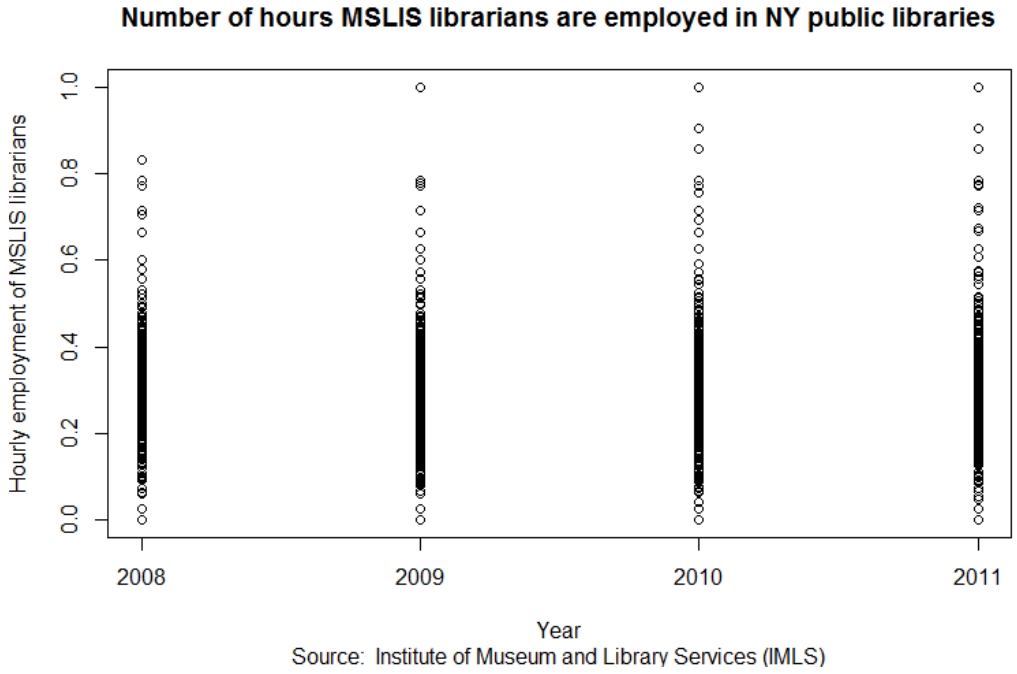

Figure 1 A multi-year visualization of the distribution of the hourly employment of degree-holding librarians. Note in 2008 there were no libraries with 100% employment of MLS degree-holding librarians.

Question 2: In which NYS locale is there the greatest hourly employment of ALA MLS librarians? What locales rank 2nd and 3rd in the hourly employment of ALA MLS librarians?

Overall, the highest rates of hourly employment of ALA MLS librarians for each year were in the NY locale 42, i.e., "Distant, Rural" region (Institute of Museum and Library Services, 2013). A visual breakdown by year of the locales where ALA MLS librarians are employed above the threshold to show the granular yearly changes of locale frequency are below (Figure 2). One possible reason why most libraries above the threshold are in less urbanized areas could be because recent ALA MLS librarian graduates are attracted to rural libraries: First, they have a higher likelihood of securing a job in a remote area where competition is scarce (say, compared to New York City); Secondly, because new graduates have more discretion over the direction of the library as one of a fewer number of staff and less bureaucracy; Third, new librarians in smaller libraries have the opportunity to experience and learn without the added pressure of a big city environment.

Figure 2
*Note the scaled axes of Figure 2 below vary between years; interpret the results with this in mind.

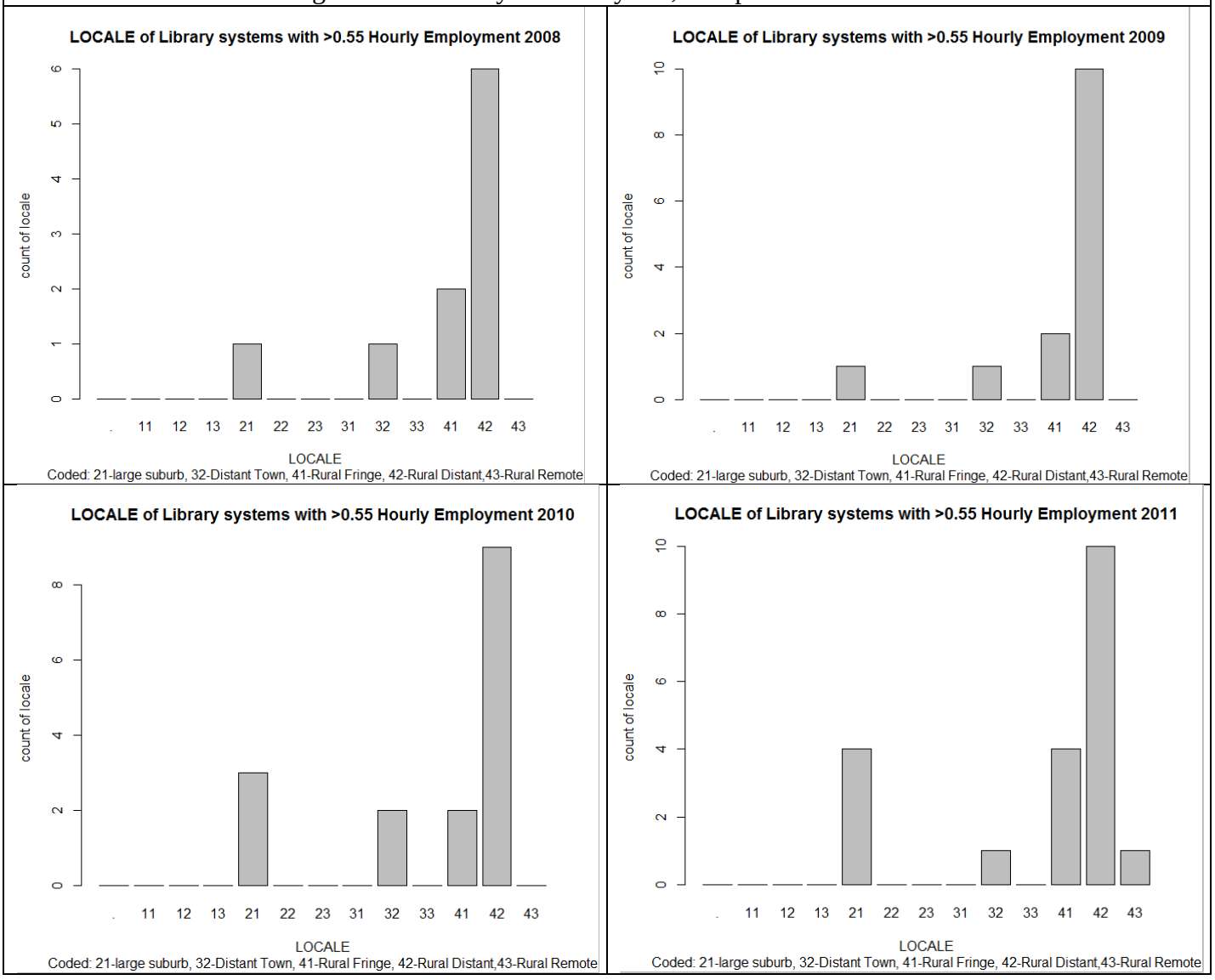

Figure 2: In 2008, there were 6 libraries in the Rural Distant locale (42). An increase of 4 libraries occurred in 2009, and went down by 1 in 2010. The final chart in figure 2 shows the number of NY public libraries with >.55 hourly employment of MSLIS librarians increased in 2011 to a total of 10 libraries. It is critical to note that the change in the count of libraries over the 4 years is not a simple case of new libraries being added above the threshold or previous libraries above the threshold dropping off. The landscape of hourly hiring is much more dynamic. For example, a year-by-year analysis shows that some libraries previously above the threshold dropped off, or vice versa.

To illustrate these changes, a subset of locale 42 by library name was extracted for the yearly details. The table below shows the libraries in Distant Rural areas above the threshold (>.55) in 2008. There are 6 libraries including Arvilla E. Diver Memorial Library (Figure 4.1).

```
> LibraryNameLoc6<-sqldf("SELECT LIBNAME, LOCALE, Year from combolibdata where
PercentHourMLS>.55 and Year=2008 and LOCALE=42")
> LibraryNameLoc6
                              LIBNAME LOCALE Year
1          ULYSSES PHILOMATHIC LIBRARY     42 2008
2 LOUISE ADELIA READ MEMORIAL LIBRARY     42 2008
3               LIVINGSTON FREE LIBRARY     42 2008
4             NORTHVILLE PUBLIC LIBRARY     42 2008
5           CORDELIA A. GREENE LIBRARY     42 2008
6    ARVILLA E. DIVER MEMORIAL LIBRARY     42 2008
```

FIGURE 4.1

Figure 4.2 shows 2009 libraries in locale 42.

```
> LibraryNameLoc7<-sqldf("SELECT LIBNAME, LOCALE, Year from combolibdata where
PercentHourMLS>.55 and Year=2009 and LOCALE=42")
> LibraryNameLoc7
                              LIBNAME LOCALE Year
1                 HILLVIEW FREE LIBRARY     42 2009
2            AUSABLE FORKS FREE LIBRARY     42 2009
3                WELLS MEMORIAL LIBRARY     42 2009
4            ULYSSES PHILOMATHIC LIBRARY     42 2009
5  LOUISE ADELIA READ MEMORIAL LIBRARY     42 2009
6                LIVINGSTON FREE LIBRARY     42 2009
7            CAMDEN LIBRARY ASSOCIATION     42 2009
8             NORTHVILLE PUBLIC LIBRARY     42 2009
9            CORDELIA A. GREENE LIBRARY     42 2009
10      RENSSELAERVILLE PUBLIC LIBRARY     42 2009
```

FIGURE 4.2

Now, there are 5 new libraries added to the list because Arvilla E. Diver Memorial Library dropped off the list and HillView Free Library, Ausable Forks Free Library, Wells Memorial Library, Camden Library Association, Rensselaerville Public Library joined the above-threshold group. Although the list increased room 6 to 10, a faulty assumption would be that, therefore there are 4 new libraries, when the story is turns out to be subtler.

Question 3: What are NYS libraries' over the threshold electronic resources development, programming, print resources development, program attendance, ebook collection development, etc. practices from 2008-2011?

The heatmaps show each NY libraries' resource allocation, programming practices and attendance, as well as reference questions and internet computer count and usage ordered by decreasing rate of hourly employment of ALA MLS librarians (Figures 5 & 6).

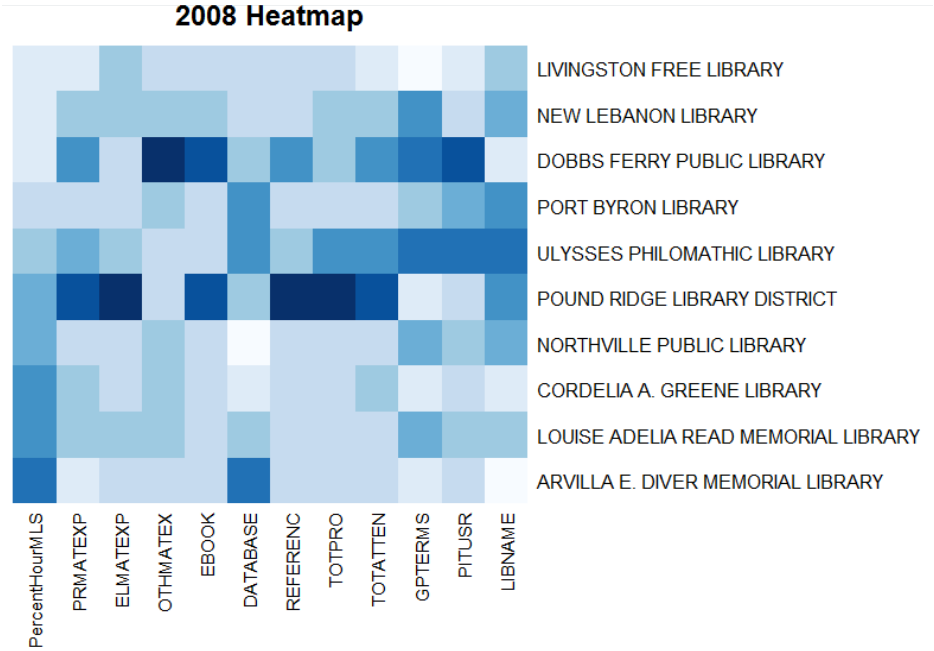

Figure 5 In the 2008 heatmap, patterns of library resource collection underscore great differences between each library. For example, Arvilla Library has the highest rate of hourly employment of ALA MLS librarians and the most database subscriptions.

Recall that the Arvilla E. Diver Memorial Library dropped off the list of above the threshold libraries from 2008 to 2009 (Figures 4.1 & 4.2). It is possible that the low amount of internet computers (indicated by the lighter colored grid squares) and low level of internet computer users accounts for this abrupt drop off. The interpretation of this pattern is that the hourly employment rate is affected by multiple factors, not a single cause (e.g. database subscription alone).

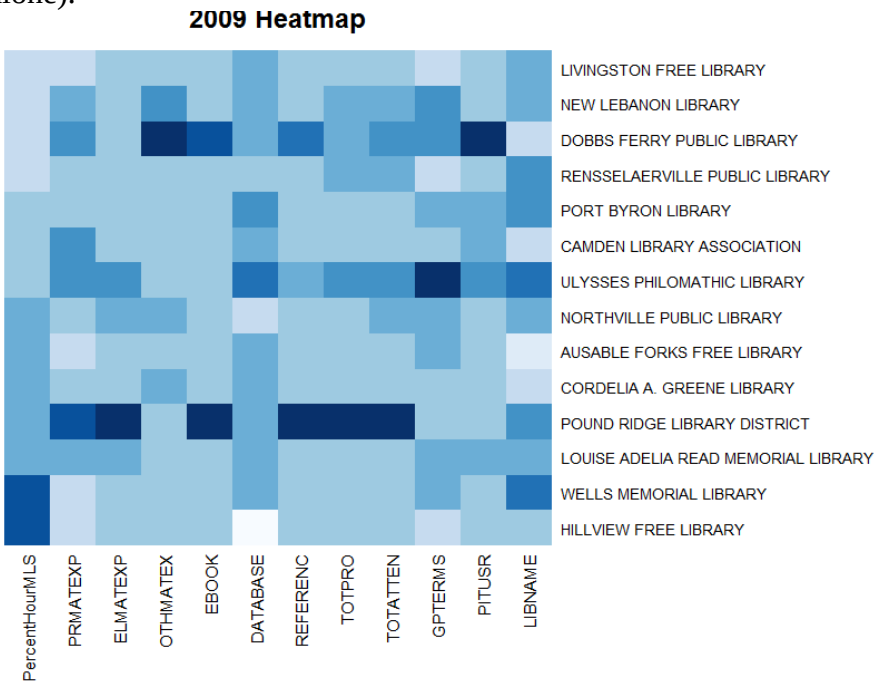

Figure 6 In 2009, Pound Ridge Library District demonstrates the changes in the relationship between resource and programming practices and the hourly employment rate of ALA MLS librarians in NY public libraries. When you compare the heatmaps of 2008 and 2009, Pound Ridge Library District ascended in the ranks from 5th in hourly employment to 4th. There was a significant increase in their program attendance and ebook purchases (darker colored grid squares). The increase of digital resources and social outreach (programs) suggests that the added value of increased rate of degree-holding librarians positively affects the library.

Question 4: Where in NYS are the libraries that employ the highest hourly rate of ALA MLS librarians in 2011?

The use of geolocation confirms findings and provides a visual representation of the geographic distribution of libraries with above threshold rates of hourly employment of ALA MLS librarians. As seen on the 2011 map below (Figure 7), rural areas have circles indicating the hourly employment of ALA MLS librarians. Findings showed that there are no libraries above the threshold in Central and Upstate NY in 2008. However, as seen in the map, by 2011 many more had bubbled up in Kingston, near the Vermont border, north of NYC, and close to Albany. Also note that there was a growth in currently existing libraries above the threshold in areas such as just north of New York City, which grew from 0.6 in 2008 to a considerably large 0.8 in 2011. Overall, most of the circles grew as time progressed, indicated by increased circle size in Figure 6.4, with the minimum value jumping from 0.6 in 2008 to a consistent value of 0.75 in 2009, 2010, and 2011.

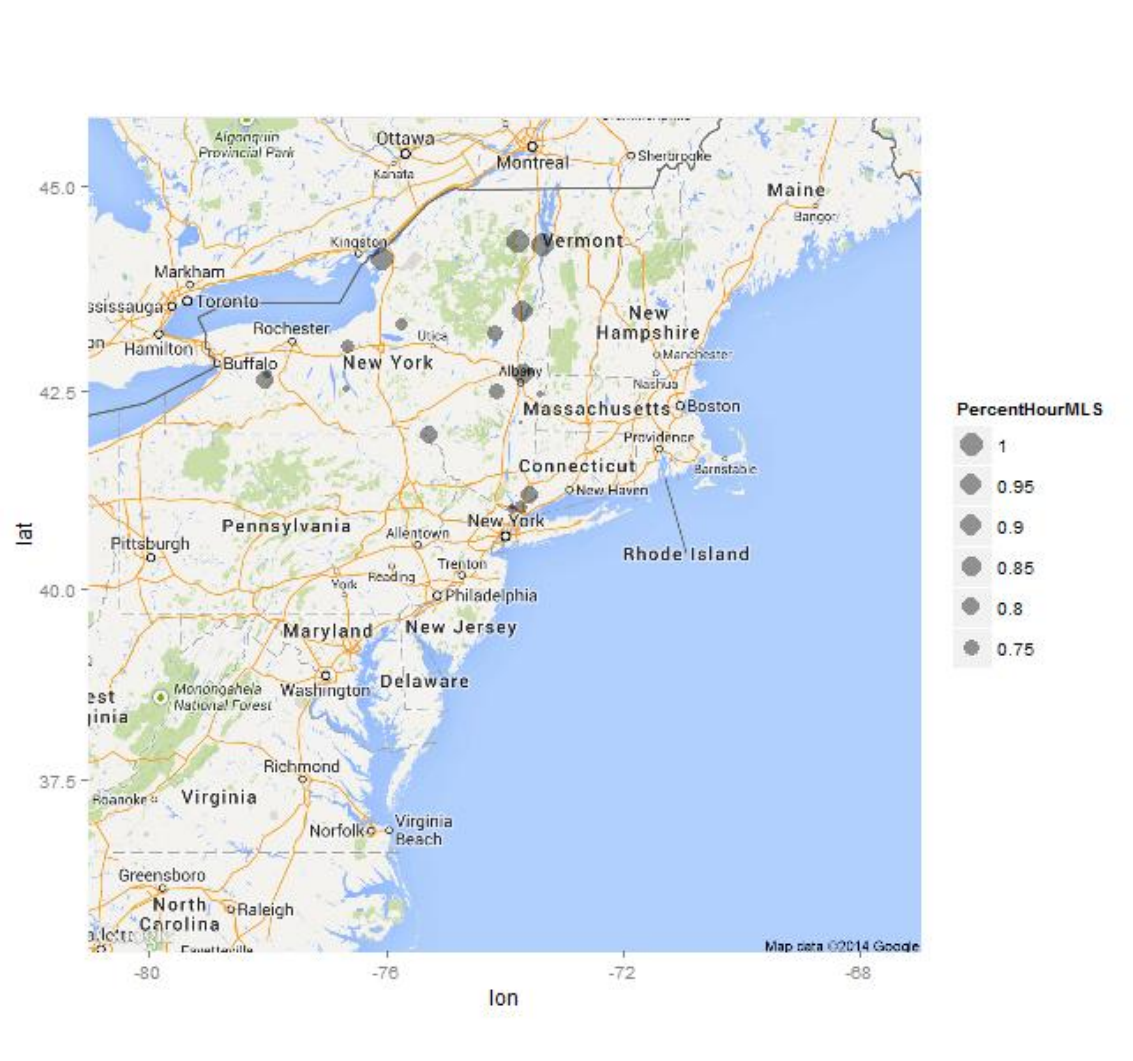

Figure 7 The 2011 map of NYS public library and the hiring rates of MSL-degree-holding librarians.

Question 5: What attributes and characteristics describe (and "predict") the NY public libraries with the greatest or lowest hourly employment of ALA MLS librarians?

Running association rules corroborated the previous 4 questions results. Beginning with apriori association analysis in the default setting mode, ~~we found~~ the strongest rules had a right hand side (RHS) of "Threshold=No." The exact same rules appeared when we ran association rules (apriori) with specified parameters. The association rules were set with specific parameters, with adjusted confidence and support as well as a specified RHS to try and produce rules which show correlated characteristics of libraries with hourly hiring rates above the threshold. However, there were no strong rules with RHS of "Threshold=Yes." The following rules are those we

handpicked as the most 'interesting' rules, in terms of the highest support confidence, and lift as well as with the greatest relevance to our question:

```
 lhs                rhs             support confidence      lift
6  {ELMATEXP=0,
   OTHMATEX=0,
   EBOOK=0}         => {Threshold=No} 0.05853175  1.0000000 1.0202429
7  {ELMATEXP=0,
   OTHMATEX=0}      => {Threshold=No} 0.08564815  0.9961538 1.0163189
15 {DATABASE=14}    => {Threshold=No} 0.05059524  0.9935065 1.0136180
16 {OTHMATEX=0,
   EBOOK=0}         => {Threshold=No} 0.08994709  0.9927007 1.0127959
17 {GPTERMS=7}      => {Threshold=No} 0.07771164  0.9915612 1.0116333
20 {ELMATEXP=0,
   EBOOK=0,
   Year=2008}       => {Threshold=No} 0.09193122  0.9893238 1.0093506
32 {ELMATEXP=0,
   EBOOK=0}         => {Threshold=No} 0.25000000  0.9805447 1.0003938
39 {EBOOK=0,
   GPTERMS=4}       => {Threshold=No} 0.06018519  0.9784946 0.9983022
```

As graphically represented in Figure 8 (below), there is a high likelihood that there are no ebooks and no expenditures on "other materials" or electronic materials" when the hourly employment of ALA MLS librarians in NY public libraries is beneath the threshold.

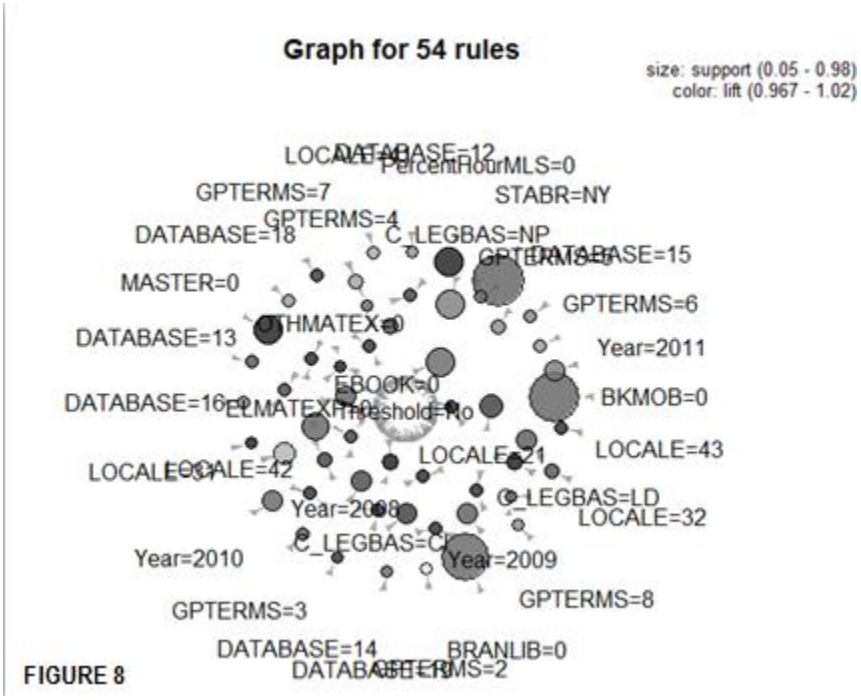

FIGURE 8

Figure 8 The support is visualized by the size of the circles, the color of the circles shows the lift, and the distance from the center represents the strength of the rules. When "Threshold=No" the closest circles (spatially) are EBOOK=0, OTHMATEX=0, and ELMATEX=0.

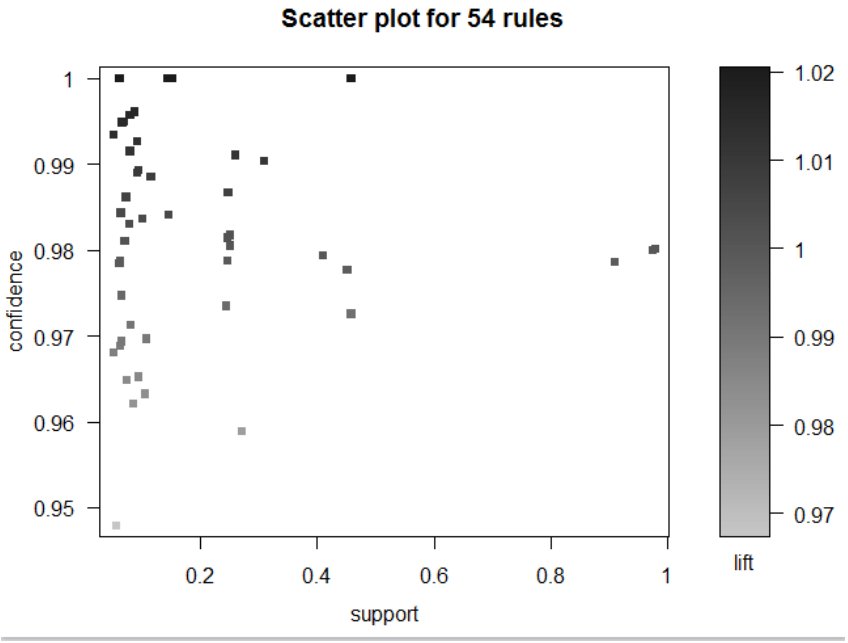

Figure 9 The strength of the top 54 non-redundant rules. Darker shades represent the third value dimension of the lift, in addition to support and confidence. The top strongest rules appear in the upper left quadrant of the scatter chart.

Limitations and Solutions

An objection was leveled by a NY3Rs Big Data in Libraries (2013) conference attendee, who argued that the analysis was faulty because it showed some NYS public libraries in (2008-2011) with 0 hourly employed MSLIS librarians. The objection was based on the claim that there is a law that requires NYS public libraries to hire at least one MS LIS librarian. In rebuttal, the authors conducted research and found that indeed, NY State has education requirements for its libraries. However, not all are required to hire ALA MLS librarians. In fact, depending on the size of the county population, a library may not be required to hire personnel with a professional or library degree (NYCRR TITLE 8-EDUCATION: §90.8 Appointment of Library Personnel, 2010). The law states that New York state public libraries are required to "employ a paid director with qualifications based on the population served" (Aldrich & Nichols, 2010).

Second, the scope of this research does not include offering steps for handling institutional data, such as internal patron usage records. Publicly available data (such as that of the IMLS) is stripped of identifying markers that might threaten library patron or employee privacy. Strategies

for warehousing and anonymizing patron data is addressed by bibliomining research (Nicholson, 2006). For example, work has been done on systematizing the process of de-identifying individuals' usage data looking to the Health Insurance Portability and Accountability Act (HIPAA) as a framework (Nicholson & Smith, 2007).

Discussion and Conclusion

Overall, analysis indicates an overall increase in the yearly trend of hourly employment rate of ALA MLS librarians in NY public libraries. The highest rates of hourly employment of ALA MLS librarians for each year were in the NY locale 42, and patterns of library resource collection show paradigm shifts from the years 2008-2011.

Next steps in this research will include scaling by two dimensions: First, the number of years included in analysis and second, the number of states. This expansion to a wider window of time in the multi-year analysis will allow the capturing the temporal dynamics in library hiring practices and collection development and programming. Including other U.S. states will make comparative analysis possible. Finally, future research will work toward the development of a dedicated toolkit tailored to enable the analysis of public library and open data so that public, academic, special, and corporate libraries alike can continue to harness data science techniques for institutional insight and community action.

References

Aldrich, R. S., Nichols, J. (2010). *Handbook for New York Public Library Directors.* New York Library Association. Retrieved from: http://www.nysl.nysed.gov/libdev/trustees/handbook/handbook.pdf

Ekbia, H., Mattioli, M., Kouper, I., Arave, G., Ghazinejad, A., Bowman, T., . . . Sugimoto, C. (2014). Big data, bigger dilemmas : A critical review. *Journal of the Association for Information Science and Technology*. doi:DOI: 10.1002/asi.23294

Institute of Museum and Library Services. (2013). Data File Documentation Public Libraries Survey: Fiscal Year 2011. PL 111-340. Retrieved from: http://www.imls.gov/assets/1/AssetManager/fy2011_pls_data_file_documentation.pdf

Krol, J. J. (2013, July 2). Plot Addresses on a Map Using R. *jjkrol.pl.* Retrieved from: http://jjkrol.pl/plot-addresses-on-a-map-using-r/)

Marchi, M. (2013, January 24). Maps in R: choropleth maps. *Milano R Net*. Retrieved from: http://www.milanor.net/blog/?p=634.

Nicholson, S. (2006). The basis for bibliomining: Frameworks for bringing together usage-based data mining and bibliometrics through data warehousing in digital library services. *Information Processing & Management* 42(3), 785-804.

Nicholson, S. & Smith, C.A. (2007). Using lessons from health care to protect the privacy of library users: Guidelines for the de-identification of library data based on HIPAA. *Journal of the American Society for Information Science and Technology* 58(8), 1198-1206.

Mathew, P., Dunn, L., Sohn, M., Mercado, A., Custudio, C., & Walter, T. (2014). Big-data for building energy performance: Lessons from assembling a very large national database of building energy use. *Applied Energy*, 140(2014), 85-93.

NYCRR TITLE 8-EDUCATION: §90.8 Appointment of Library Personnel. (2010, March 15). *New York State Education Department*. Retrieved from: http://www.nysl.nysed.gov/libdev/excerpts/finished_regs/908.htm

Stanton, J. (2012). An Introduction to Data Science. (p 196). Retrieved from: https://docs.google.com/file/d/0B6iefdnF22XQeVZDSkxjZ0Z5VUE/edit?pli=1

Yau, N. (2011). *Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics*. Wiley. Ebook.

Zhao, Y. (2014). Association Rules. *Rdatamining.com.* Retrieved from: http://www.rdatamining.com/examples/association-rules