Data Analytics to Improve School Libraries

Dr. Lesley S. J. Farmer
California State University Long Beach
1250 Bellflower Boulevard, Long Beach CA 90840
USA
Lesley.Farmer@csulb.edu

Abstract

To improve programs, school librarians can analyze and apply data. Standards can help with the process, such as California's model school library standards. To meet those standards, it is important to recognize and focus on variables to improve library programs. Data analytics can help librarians identify which data to collect, how to organize and analyze the data, and make informed recommendations for library improvement. Data analytics based on the California school library survey offer a predictive model for school library program effectiveness. Data analytics based on the American Association of School Librarians survey offers a longitudinal look at school library programs. A sample scenario that addresses one of the key variables provides the basis for employing data analytics to improve services.

Keywords: data analytics, standards, predictive models, surveys, program improvement

Introduction

According to the 2015 *IFLA School Libraries Guidelines*,"the goal of all school libraries is to develop information literate students who are responsible and ethical participants in society" (p. 7). The guidelines also acknowledge the importance of evaluation; "Successful evaluation leads to renewal of programs and services, as well as development of new programs and services" (p. 7). To that end, school librarians should systematically collect and analyze data in order to make informed decisions. This process is called data analytics. One process within data analytics is data analysis: organizing, evaluating and finding patterns in the data, oftentimes using statistical tests.

Library Standards

Data analytics and library standards go hand in hand to help librarians improve school library programs. It should be noted, though, that school library standards exist, mainly because each community has unique characteristics and needs. Guidelines such as the IFLA's and the American Association of School Librarians (AASL) of 2009 provide criteria for assessing school libraries as a basis for program improvement. Nevertheless, some states in the USA have created standards, including California. Furthermore, California saw a need to underpin the standards with research, an effort spearheaded by Dr. Farmer. She identified contributing variables that appeared consistently in the literature:

- staffing (*full-time credentialed school librarian, full-time paraprofessional*)
- access (*flexible access to the library throughout the day for groups and individuals*)
- services (*instruction, collaboration, reading guidance and promotion, reference, interlibrary loan*)
- resources (*large current diverse and relevant materials that are well organized*)
- technology (*Internet connectivity, online databases, online library catalog, library web portal*)
- management (*budget, administrative support, documented policies and procedures, strategic plan with assessment*).

The presence of the specific variables (shown in italics) became the basis for the California school library baseline standards. The variables that are quantitative in nature (e.g., budget size, currency of collection) were calculated to determine adequate levels of support, which also constituted part of the baseline library standards (California State Department of Education, 2011).

Collecting and Cleaning Data

As school librarians identify possible variables that impact school libraries, they need to make sure their data are valid and reliable. For instance, circulation readings do not measure the amount that students read because a person might read a small portion of the document – or not at all. Self-reporting can be unreliable since a person might write what is expected rather than the truth. Even standardized tests measure only some aspects of reading. While holistic and somewhat hard to measure, students' discussion about materials they have read might well constitute more valid and reliable data about reading habits.

Collected data also need to be organized and "cleaned" before they can be analyzed. Data are typically inputted into a spreadsheet, with each variable constituting a column, and each "observation" or respondent constituting a row. To facilitate analysis, variables might be coded (e.g., elementary school = 1, middle school = 2, high school = 3); a code book maintains the coding scheme. Sometimes a new variable, composed of existing data, can be created to provide useful information, such as calculating the number of books per student. Data might have missing values or typos, which have to be addressed. Some data need to be anonymized to protect individuals' privacy. Only after the data are well structured and identified can meaningful analysis be conducted.

Matching Quantitative Data with Data Analysis Tools

Sometimes the most difficult decision is matching quantitative data with the appropriate data analysis tool. Descriptive statistics (e.g., mean, median, mode, range, variation) can be used with most all quantitative data. However, inferential statistics provide more powerful analysis. Here are a few such statistical methods:

- Correlation analysis. This statistical method quantifies linear relationships between two variables. The analysis generates a correlation coefficient r that tells the strength of the relationship and whether that relationship is positive or negative. For example, what is the correlation between the number of ebooks and print book circulation? This analysis of continuous values uses a Pearson correlation measure; a chi square measure is used for categorical data (e.g., gender).
- Regression analysis. This statistical method tries to generate a model (i.e., a line or curve that fits the data best) that shows the relationship between pairs of numerical data. Unlike correlation analysis, the regression line may be nonlinear. Several variables might affect one dependent variable, which would use multiple regression. When such models are established, they can be used to predict outcomes (the dependent variable). For instance, a regression analysis might inform librarians about optimum length of times for training.
- ANOVA (analysis of variance). Variability of data determines if greater differences in data exist between groups than within a group. For instance, do KM-trained engineers and physicists get more relevant website "hits" than their untrained counterparts; does more variance occur within or across subject domains or training?
- Cluster analysis. This statistical method uses distances between variables to group observations together. Those with smaller distances between them are assumed to be similar, so looking closer at the individual clusters can potentially determine important characteristics. For example, how might information behaviors cluster for successful retrieval of information?
- Failure modes and effects analysis (FMEA). This step-by-step approach identifies all possible failures (e.g., defects, errors) of a product or process, such as cataloging a resource. Failure mode refers to the ways that something might fail. Effects analysis refers to the effects or consequences of a failure. Failures are prioritized by their frequency, the seriousness of their consequence, and the ease of their detection. The analysis informs further determination of probable causes of failure, and potential ways to reduce failure.

Other data analysis tools are useful for visualizing data.

- Histogram (bar graph). Bar graphs visualize categorical quantities, such as usage of different types of resources.
- Pie charts illustrate distributions as percentages of a whole, such as collection mapping.
- Pareto chart. This kind of bar graph represents quantities from the longest on the left to the shortest on the right, which visualizes the most significant factor. For instance, what is the average turn-around time for different kinds of interlibrary loan requests?
- Run chart. This line graph shows the measure of process performance (y-axis) over time (x-axis), and is used to compare performance before and after an intervention. Usually the goal is consistency or improvement. For example, a run chart could show the amount of time it takes to code a resource.
- Control chart. This graph shows how a process changes over time. The chart has a central line for the average, an upper line for the upper control limit, and a lower line for the lower control limit. The chart shows the variability over time, with the intent to decrease that variability. Control charts are useful for baseline data, and to determine the impact of an intervention. A variable control chart displays variable data; which are subgrouped (e.g., desktops vs. laptops); averages and ranges for each subgroup are plotted on separate charts.

Data Analytics to Develop a Predictive Model

As the California school library program standards were being approved, Farmer and Safer (2010) wanted to determine if a significant difference existed between those school libraries that met the standards and those who did not. Using the state's most recent school library data set (2007-2008), the researchers applied descriptive statistics to identify standards variables; to be so designated, at least half of the survey respondents had to meet that specific baseline standard variable (that is, the library didn't have to meet all of the factors' standards). Next, the researchers divided the data set into two groups: one that met all baseline standards, and one that did not meet all baseline standards. A t-test determined that the two groups were significantly different relative to resource and service standards; the most significant difference relative to the baseline standards was the presence of a full-time school librarian.

The researchers also used logistic regression to predict a response based on input data; that is, what is the effect of changing one predictive variable on the response variable, holding the other predictive variables constant (e.g., flexible scheduling and planning with teachers). The logistic regression analysis found that several variables related to resources and services further differentiating the two groups: number of subscription databases, library web portal presence, information literacy instruction, Internet instruction, flexible scheduling, planning with teachers, book and nonbook budget size, and currency of collection.

This research led to the effort to develop a statistical model to predict which school libraries could meet state standards, based on a set of variables (Farmer, Safer & Leack, 2015). The researchers used 2007-2008 and 2011-2012 data sets from the California school library survey to construct the model. Two main statistical techniques are recognized for developing decision procedures: logistic regression and decision trees. Decision trees are used to predict a categorical response, such as meeting standards or not; a decision tree diagram looks like a flow chart because it is essentially a sequenced set of if-then decisions based on questions. When comparing the decision tree and logical regression model, the decision tree resulted in a model that fit the data better (e.g., fewer misclassifications). The important predictive variables that emerged from the decision tree included (in order of importance): budget for non-book materials, evening access, book budget, number of books, level of library, availability of DVDs, having classified staff, having online subscriptions (including streaming), and providing textbook service.

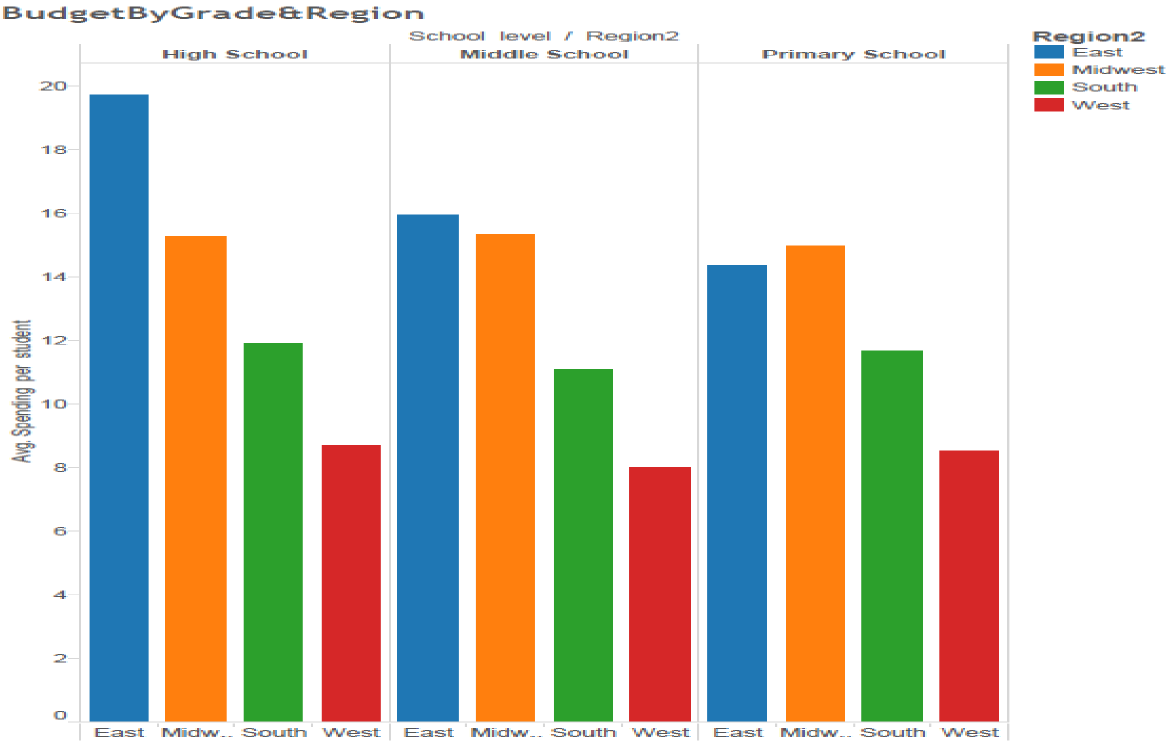
Funding overall, and use of a variety of funding sources, were major factors in school library program status, and form the basis of a predictive model for effective school library programs. The findings pointed out the need for librarians to be aware of these different funding streams, and to take advantage of them, which may require pro-active communication and negotiation with decision-makers. Resources and access to them constituted another important “leg” of school library programs. Books, nonprint and online resources are all needed, and some analysis seemed to indicate that not only physical access but intellectual access through instruction were needed in order to make a difference. In general, there seemed to be a sizable gap between the vast majority of school libraries who are providing basic resources and services and those stellar libraries who have rich collections, innovative services, and expanded access.

American Association of School Librarians Longitudinal Data

For several years, AASL surveyed librarians about their school library programs: School Libraries Count! Researchers Farmer and Safer wanted to see what patterns emerged over time, between 2009 and 2012, in this national data set. They used the statistical packages MiniTab, SAS, SPSS, and the data visualization tool Tableau to analyze fifteen key variables: spending per student, total expenditures, enrollment, books per student, books, average copyright year, audio items, video items, periodical subscriptions, hours open, hours worked by all staff, number of teacher librarians, number of individual visits per week, number of group visits per week, and number of computers in the library. The trends for the important variables were considered by school level, AASL member, and region.

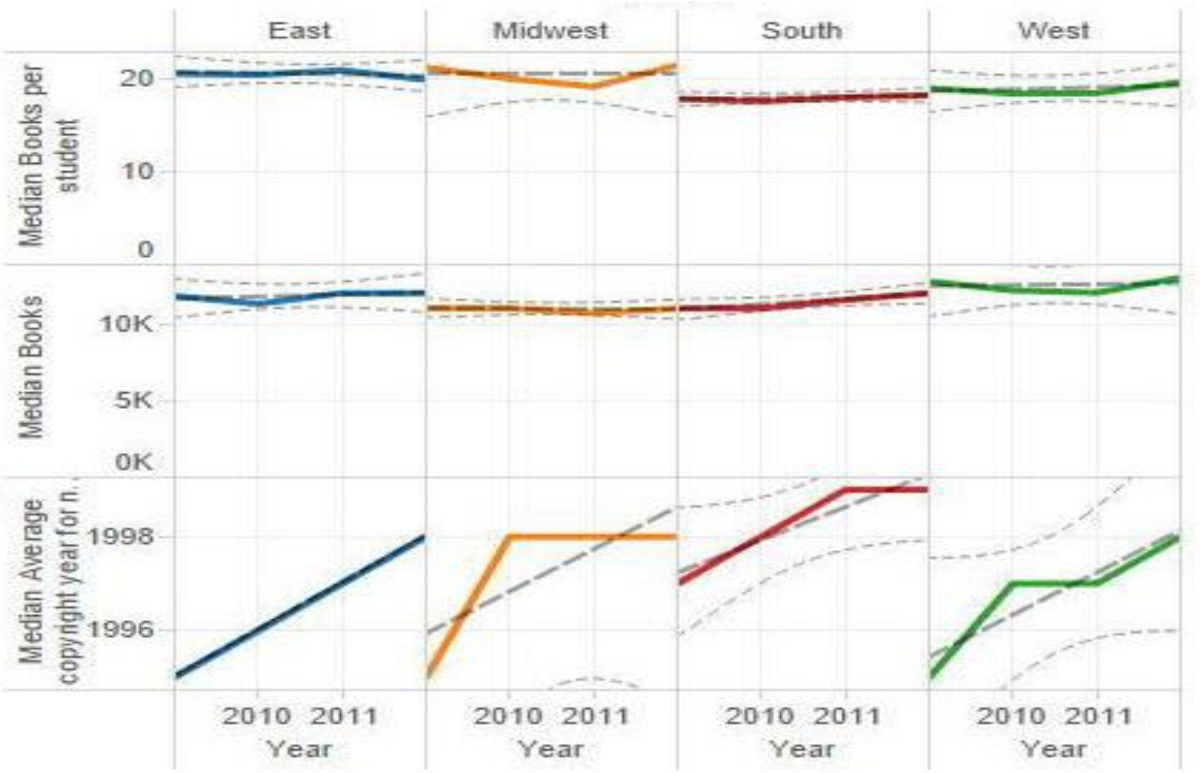
However, clear positive trends existed for the average copyright date of books and the number of computers in the library. Descriptive statistics provided a useful picture of trends, especially if median (the 50% point) values were used instead of mean (average) values; this decision was made to address outlier values (e.g., new core collections or budgetary windfalls). For the continuous variables (spending per student, total expenditures, audio items, video items, and periodical subscriptions) clear negative trends occurred as evidenced by school level and particularly by geographic region. This is likely due to the differences between regions observed in spending per student and total expenditures.

Figure 1: Budget per Student by School Level and Region



However, clear positive trends existed for the average copyright date of books and the number of computers in the library. On the whole, these trends confirm that school library resources and services were negatively affected during the years from 2009 to 2012.

Figure 2: Median Number of Books per Student, Median Book Collection Size, and Median Book Copyright Date by Year and Region



Several correlational statistics tests were applied to determine possible relationships between variables. A couple of variables were especially telling. Using ANOVA, it was found that he number of hours that the school librarian met with teachers correlated positively with:

- Spending per student (.037 p=.02),
- Number of Group Visits per Week (.062 p<.001)
- Number of Instructional Hours (.128 p<.001)
- Number of Computers (.154 p<.001)
- Currency of Books (.067 p<.001)

The number of books per student correlated negatively (-.075, p <.001). Similarly, more computers correlated significantly with fewer books per student. On the other hand, the number of computers correlated positively with the currency of books (.120, p <.001).

Longitudinal data are useful because they mitigate changes that might occur just for one year; they provide more stable patterns in school library resources and services. Longitudinal data can also be used to predict future trends. For instance, AASL data indicated how collections decreased as computers increased. This finding can be reinforced by asking about subscriptions to ebooks and the use of digital reference sources, which could account for fewer print purchases – and may predict future library budgeting decisions.

Sample Scenario for Data Analytics

One of the core functions of the school library is to provide physical access to its collection of information resources. Along the way several bottlenecks can occur along the way, from processing and cataloging to shelving, the end result that the user cannot access the wanted information. Because errors can happen at any of these points, a failure analysis is an important aspect of quality control and effective workflow. A good first step is to collect data about unavailability: what was the reason for not finding the item; and how was that problem solved (McGurr, 2007). A sample table helps keep track of problems on a daily basis at the service desks. Several rows are left open for staff to identify other possible problems. Once the recording has stabilized, the table can be reviewed to add possible often-reoccurring problems, such as:

	MON	TUES	WED	THUR	FRI	RESOLVED	NOT RESOLVED
Missing							
Withdrawn							
Being Catalogued							
Being Processed							
Being Repaired							
Being Shelved							
Misshelved							
Being Held In Reserves Or Other Location							
Being Transferred							
Miscatalogued							
Wrong Barcode							

Even a cursory look at the above data grid can reveal several potential patterns. It takes staff considerable time to track down items that the user cannot find. Data might show that misshelving might occur more on certain days; follow-up observation and interviews can determine if the shelver is less productive or else needs training (the fact that the weekend person takes a long time to find how that the item is being shelved might point to a time management issue). The fact that it takes 12 minutes to note that an item is mislabeled, but that it occurs just once, can lead one to conclude that mislabeling seldom occurs but that it can significantly cost time when it does happen. Patterns also exist in terms of making local notes in the MARC record; it is apparently done consistently for cataloging and repairs, but not for processing, which calls for further investigation and probably a reminder to that unit to record their work more consistently. A high number of items being repaired might also signal further investigation: what is causing the need for repairs; are some staff over-identifying the need for repairs, could some minor repairs such as fixing a slight tear be done at the circulation desk, is the turn-around time for repairs reasonable? These data should also take into account the total circulation figures; if the daily checkout is over a thousand, then the unavailability figures are reasonable; if the daily checkout is less than a hundred, serious follow-up is needed. Some of these problems may be hard to identify, or take considerable time to figure out, which is all the more reason to identify frequent problems and resolve the basis for them so as to provide better service. It is obvious that collecting such data regularly can inform operations, and facilitate their targeted improvement.

References

American Association of School Librarians. (2009). *Empowering learners: Guidelines for school library programs*. Chicago, IL: American Library Association.

California Department of Education. (2011). *Model school library standards for California public schools kindergarten through grade twelve*. Sacramento, CA: California Department of Education.

Farmer, L., & Safer, A. (2010). Developing California school library media program standards. *School Library Media Research, 13.*

Farmer, L., Safer, A., & Leack, J. (2015). Using analytic tools with California school library survey data. *Evidence-Based Library and Information Practice, 10*(2), 90-107.

International Federation of Library Associations and Institutions. School Libraries Section. (2015). *IFLA School Libraries Guidelines.* The Hague, Netherlands: International Federation of Library Associations and Institutions.

McGurr, M. (2007). Improving the flow of materials in a cataloging department. *Library Resources & Technical Services, 52*(2), 54-60.

Biographical note

Professor Lesley Farmer coordinates the Teacher Librarian Program at California State University. She received her doctorate in Adult Education from Temple University and her MSLS from the University of North Carolina at Chapel Hill. Dr. Farmer has worked in public and private K-12 school libraries as well as in public, academic and special libraries. Dr. Farmer served as IASL VP for Organization Relations, and chairs the IFLA School Libraries Section. She is a Fulbright Scholar, and has won several professional awards. A frequent presenter and writer, Dr. Farmer's research interests include data analytics, information literacy, and digital citizenship. Her latest books are Information and Digital Literacies: A Curriculum Guide for Middle and High School Librarians (Rowman & Littlefield, 2015) and *Introduction to reference and information services for today's school libraries* (Rowman & Littlefield, 2014).