Library Circulation as Interaction Between Readers and Collections: the Square Root Law

Paul B. Kantor, Wonsik Shim
Alexandria Project Laboratory
School of Communication, Information and Library Studies
Rutgers University

Abstract

We present a model of the relation between the circulation at a large library, the collection at that library, and the size of the user population. Based on analysis of data at Academic Research libraries in the United States we find a “square root law” which describes this relation quite well. We explore the theoretical foundations for such a law, in a geometric analogy, and speculate on bibliometric implications.

1. THREE MODELS OF LIBRARY CIRCULATION.

It is reasonable to suppose that the circulation at a particular type of library increases as the size of the collection increases. It is also reasonable to suppose that circulation increases as the number of readers increases. The simplest model of these two relationships is a linear one.

But circulation involves interaction between the collection and the readers. If there are no readers, there will be no circulation, whatever the size of the collection. Similarly, if there is no collection, there will be no circulation, whatever the number of readers.

A very important model of this kind of interaction was given more than 60 years ago by Lotka and Volterra (independently). The essential idea is to model some phenomenon which requires an interaction between two populations of different types. According to the Volterra-Lotka (VL) model, the rate at which such a phenomenon occurs should be proportional to the product of the sizes of the two populations. This model has been applied successfully to model many natural systems (Murray, 1989), ranging from fish to wolves. It has recently been applied to human activity as well (Zangwill & Kantor, 1998). We therefore propose that it should be applicable to the specific interaction between readers and books that gives rise to the phenomenon known as circulation.

Even more generally, phenomena which vanish unless two kinds of entities are both present can be described by log-linear models.

To sum up, representing circulation by $C$, reader population by $R$, and collection size by $B$, we have three possible models which might describe the relation between these variables at a set of similar libraries. [We suppress an index $i$ that labels individual libraries.]

Linear Model (L):

$$C = \alpha R + \beta B + C_0 + \epsilon$$

Product or Volterra-Lotka Model (VL):

$$C = C_{11} RB + \epsilon$$

Log-Linear Model

$$\ln C = \ln C_{11} + \alpha \ln R + \beta \ln B + \epsilon$$

In these models, $C_0$, $C_{11}$, and $C_{11}$ are normalizing constants. $\alpha$ and $\beta$ are coefficients, and $\epsilon$ represents the random variation not accounted for by each model. The random variation is assumed to follow a normal distribution.
Note that if each model holds exactly, then the log-linear model includes the VL model as a special case in which $\alpha = \beta = 1$. However, the two models differ in the behavior of the random deviations which must be present in a real situation. In the Product, or VL model, the quantity $(C - C_1 R B)$ has a normal distribution, while in the log-linear model, the quantity $(\ln C - \ln C_1 - \ln R - \ln B)$ has a normal distribution. These are different assumptions, which can be tested by examination of the residuals.

2. PREVIOUS WORK

While there has been a substantial amount of research seeking to explain the costs of library operations in terms of various measures of the inputs and the outputs of those libraries (Cressanthis, 1995; Kantor, 1986) we have been unable to find any studies which seek to explain the levels of output in terms of library inputs and contextual variables or constraints, such as the size of the potential user populations.

Conceptually, our model is particularly close to the work of Orr (1973), who, while not presenting any sort of a mathematical formulation, did note that the utilization of any library is the result of both the capability of the library (as measured by usual indicators of library quality), and the demand for its services. From this perspective, the present work can be seen as an effort to use the size of the user population as a surrogate for the demand, and to use circulation as a convenient measure of utilization. We are well aware that it is only one of several possible measures of utilization, and in that regard the present work marks only the first steps in an effort to develop a comprehensive theory of the levels of utilization of libraries of all types.

Central to our analysis is the idea that when each of two factors is necessary for some third effect to take place, the product model is the most natural way to represent that relationship quantitatively. The product relationship captures the idea that the total utilization can go up in proportion to either of the variables, if the other is held fixed. As we shall see, the situation for library utilization does not quite correspond to this most direct and simple model.

3. DATA AND DEFINITIONS

We have analyzed data on members of the Association of Research Libraries, obtained from their web site (gopher://arl.cni.org:70/11/stat/machine/94-95 for the machine readable files or http://www.arl.org/stats/Statistics/arlstat/arlstat.html for the human readable form). For circulation ($C$) we use total circulation (TOTCIRC, initial circulation plus renewal). For the size of the reader population ($R$) we use total number of full time students (TOTSTU). For the collection size ($B$) we use total volumes held (VOLS).

We selected academic libraries in the United Stated for study, as we suspect that whatever underlying rule may emerge might well be different for the non-academic, or the Canadian libraries included in the ARL. For the same reason, we have analyzed the U.S. Academic ARL libraries together, and also separated into groups according to whether they are publicly or privately supported.

4. RESULTS OF THE ANALYSIS

4.1 Three models compared.

All three models have been tested against 9 sets of data (Public, Private, All) for (1995, 1996, and 1995-96), resulting in 27 analyses. The resulting goodness of fit measure $R^2$, together with the degrees of freedom, $F$ statistic, and associated statistical significance, as shown in Table 1. Each model is presented in a form which permits us to use linear regression. In the first and second cases the circulation is the dependent variable. For the linear model, the number of readers, and the size of the collection, are independent variables. For the VL product model, there is a single auxiliary independent variable, $X = BR$, and the intercept is required to be zero. Finally, for the log-linear model, the natural logarithm of $C$ is the dependent variable, and the logarithms of $B$ and $R$ are the independent variables. Regressions were done using the SPSS statistical package for Windows (Version 6), and the Microsoft Excel Spreadsheet (97 Version). Detailed results are available in the related Technical Report (Shim & Kantor, 1998).
Comparing the three models (see Table 1) we see that the linear model, when applied to pooled data, does quite poorly. It is inferior to the others, in terms of R-squared, and, as discussed above, lacks an intuitive justification. After all, under a strict linear model, one will predict that there is non-zero circulation at libraries with no books, or libraries with no readers!

4.2 From log-linear to log-product models

We now concentrate on the log-linear models. They combine good performance with the interactive property which should, intuitively, be a feature of correct descriptions of circulation. The relevant data on this model, for each of the 9 cases considered, is shown in Table 2.

There are two striking features of this data. First, the parameters $\alpha$ and $\beta$ are very nearly constant, but not close to each other. Note also that the values of $\alpha$ are within two standard deviations of 0.5, which corresponds to a square root dependence on $R$. 

<table>
<thead>
<tr>
<th></th>
<th>Linear Model</th>
<th></th>
<th></th>
<th>VI-Model</th>
<th></th>
<th></th>
<th>Log-Linear Model</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R^2$</td>
<td>d.f.</td>
<td>F</td>
<td>Sig</td>
<td>$R^2$</td>
<td>d.f.</td>
<td>F</td>
<td>Sig</td>
<td>$R^2$</td>
</tr>
<tr>
<td><strong>1995</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Public</td>
<td>0.628</td>
<td>62</td>
<td>52.228</td>
<td>***</td>
<td>0.846</td>
<td>64</td>
<td>351.78</td>
<td>***</td>
<td>0.628</td>
</tr>
<tr>
<td>Private</td>
<td>0.710</td>
<td>26</td>
<td>31.839</td>
<td>***</td>
<td>0.831</td>
<td>28</td>
<td>137.38</td>
<td>***</td>
<td>0.670</td>
</tr>
<tr>
<td>All</td>
<td>0.624</td>
<td>91</td>
<td>75.585</td>
<td>***</td>
<td>0.842</td>
<td>93</td>
<td>495.20</td>
<td>***</td>
<td>0.653</td>
</tr>
<tr>
<td><strong>1996</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Public</td>
<td>0.575</td>
<td>62</td>
<td>41.967</td>
<td>***</td>
<td>0.816</td>
<td>64</td>
<td>284.10</td>
<td>***</td>
<td>0.667</td>
</tr>
<tr>
<td>Private</td>
<td>0.797</td>
<td>24</td>
<td>47.133</td>
<td>***</td>
<td>0.907</td>
<td>26</td>
<td>252.10</td>
<td>***</td>
<td>0.753</td>
</tr>
<tr>
<td>All</td>
<td>0.586</td>
<td>89</td>
<td>63.015</td>
<td>***</td>
<td>0.830</td>
<td>91</td>
<td>444.27</td>
<td>***</td>
<td>0.705</td>
</tr>
<tr>
<td><strong>1995-96</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Public</td>
<td>0.598</td>
<td>127</td>
<td>94.546</td>
<td>***</td>
<td>0.830</td>
<td>129</td>
<td>631.03</td>
<td>***</td>
<td>0.648</td>
</tr>
<tr>
<td>Private</td>
<td>0.750</td>
<td>53</td>
<td>79.547</td>
<td>***</td>
<td>0.867</td>
<td>55</td>
<td>358.80</td>
<td>***</td>
<td>0.700</td>
</tr>
<tr>
<td>All</td>
<td>0.603</td>
<td>183</td>
<td>139.010</td>
<td>***</td>
<td>0.836</td>
<td>185</td>
<td>941.04</td>
<td>***</td>
<td>0.677</td>
</tr>
</tbody>
</table>

*** $p < .001$
This inequality suggests that the reasoning underlying the product model is at best partially correct. For, if the coefficients $\alpha$ and $\beta$ were equal we could represent them both by a single parameter $\gamma$. The model then becomes a “log-product” model:

$$\ln C = \ln C_i + \gamma \ln R + \gamma \ln B + \varepsilon$$
$$= \ln C_i + \gamma (\ln R + \ln B) + \varepsilon$$
$$= \ln C_i + \gamma \ln(RB) + \varepsilon$$

In other words, the equality of these parameters would mean that both readers and collection enter the determination of circulation only through their product.

We have tested this idea, and the results are shown in Table 3.

Table 3. Parameters of the Log Product Model for ARL Circulation Data as a function of the Reader Population and the Collection Size

<table>
<thead>
<tr>
<th>Year</th>
<th>Group</th>
<th>$C_i$</th>
<th>StdDev</th>
<th>$\gamma$</th>
<th>StdDev</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1995</td>
<td>Public</td>
<td>-1.157</td>
<td>0.679</td>
<td>0.645</td>
<td>0.063</td>
<td>0.624</td>
</tr>
<tr>
<td></td>
<td>Private</td>
<td>-0.291</td>
<td>0.893</td>
<td>0.566</td>
<td>0.085</td>
<td>0.623</td>
</tr>
<tr>
<td></td>
<td>All</td>
<td>-0.725</td>
<td>0.508</td>
<td>0.606</td>
<td>0.047</td>
<td>0.639</td>
</tr>
<tr>
<td>1996</td>
<td>Public</td>
<td>-1.569</td>
<td>0.659</td>
<td>0.683</td>
<td>0.061</td>
<td>0.665</td>
</tr>
<tr>
<td></td>
<td>Private</td>
<td>-0.891</td>
<td>0.759</td>
<td>0.62</td>
<td>0.072</td>
<td>0.750</td>
</tr>
<tr>
<td></td>
<td>All</td>
<td>-1.298</td>
<td>0.484</td>
<td>0.658</td>
<td>0.045</td>
<td>0.703</td>
</tr>
<tr>
<td>1995-96</td>
<td>Public</td>
<td>-1.362</td>
<td>0.47</td>
<td>0.664</td>
<td>0.044</td>
<td>0.645</td>
</tr>
<tr>
<td></td>
<td>Private</td>
<td>-0.547</td>
<td>0.584</td>
<td>0.589</td>
<td>0.055</td>
<td>0.678</td>
</tr>
<tr>
<td></td>
<td>All</td>
<td>-0.995</td>
<td>0.35</td>
<td>0.63</td>
<td>0.033</td>
<td>0.670</td>
</tr>
</tbody>
</table>

Comparing to Table 1, we see that there is a much poorer fit. Hence the data do not strongly support the log-product model, which would have been preferred on grounds of logical simplicity, and for its relation to the Volterra-Lotka arguments.

5. THEORETICAL UNDERPINNINGS OF THE SQUARE ROOT MODEL.

Given the effectiveness of the two-parameter model we must look for explanations of how it could arise. We examine two theoretical pictures, a geometric picture, and a bibliometric picture. In each case we concentrate on the parameter $\alpha$, describing dependence on the number of readers and for clarity, take the value to be $\alpha = 0.5$. This is within the 95% confidence interval of all of the observed values, and has conceptual simplicity.

5.1 Bibliometric Picture

Consider the reader population. Let us suppose that the activity of books follows a Lotka distribution with exponent 3. This means that the fraction of all readers who will check out a specific number of books $n$ is given by:

$$R = \int_{n_{\text{min}}}^{n_{\text{max}}} \frac{D}{n^3}$$
the requirement that the total number of readers is given by a sum (or integral) over all levels of activity from the lowest, \( n_{\text{min}} \) to the highest, \( n_{\text{max}} \).

\[
R = \int_{n_{\text{min}}}^{n_{\text{max}}} \frac{D}{n^3} \, dn
\]

Since the fraction \( 1/n^3 \) falls sharply for large \( n \), we can replace \( n_{\text{max}} \) by \( \infty \).

\[
R = \int_{n_{\text{min}}}^{\infty} \frac{D}{n_{\text{max}}^{n/3}} \, dn
\]

\[
= \frac{D}{2n_{\text{min}}^2}
\]

And so:

\[
D = 2n_{\text{min}}^2 R
\]

In this situation the circulation activity due to these readers is the sum of the number of readers at each activity level, multiplied by the number of books \( n \) that each such reader circulates. This is given by the integral:

\[
R_A = \int_{n_{\text{min}}}^{n_{\text{max}}} \frac{D}{n_{\text{max}}^{n/2}} \, dn
\]

We suppose that the only difference between larger and smaller user populations is that for larger collections the value of \( n_{\text{min}} \) is lower. This says that \( D \) is a (more or less) universal constant.

Solving for the activity of the readers, \( R_A \) we have, for the total number of readers, \( R \):

\[
n_{\text{min}} = \sqrt[3]{D/(2R)}
\]

While for the activity of the readers, \( R_A \) we have:

\[
R_A = \frac{D}{n_{\text{min}}}
\]

Thus, substituting for \( n_{\text{min}} \) we have:

\[
R_A = \frac{D}{\sqrt[3]{D/(2R)}}
\]

\[
= \sqrt{2D \sqrt{R}}
\]

In other words, if (1) the exponent 3 holds in a Lotka law describing the reader activity, and (2) all populations of readers at similar universities are described by a single universal law, then the activity of all readers combined will always be proportional to the square root of the actual number of readers.

5.2 Geometric Picture

Suppose, hypothetically, that the reader population of a library grows like a colony of bacteria in a Petri dish, retaining a roughly circular disk shape. When the reader population is at a size \( R \), this hypothetical disk will have a “radius \( r \)” related to \( R \) by the equation:

\[
R = \pi r^2
\]

So that:
\[ r = \sqrt{R/\pi} \]

Suppose also that, in this expanding disk representing the readers, only the new readers (at the outer edge or rim) read a lot. The number of readers in this rim will be almost exactly \(2\pi r t\), where \(t\) is the hypothetical width or “thickness” of the zone of active readers represented by the rim. Combining these two equations we see that the active part of the readership, \(R_A\) is given by:

\[ R_A = 2\pi t \sqrt{R/\pi} \]

\[ R_A = 2t \sqrt{\pi} \sqrt{R} \]

Hence the geometric picture can account for the appearance of the factor \(\sqrt{R}\) in the Square Root model. Only the active part of the readership interacts with the books.

6. CONCLUSION

We have established an empirical rule which describes the relation between circulation, regarded as a dependent variable, and reader populations, regarded as independent variables. For the data from which we have derived this rule, the description is remarkably accurate.

This rule, which we propose to call the “Square Root Law,” predicts circulation proportional to the product of the square root of the reader population and fractional power of the collection size. The appearance of a product is justified by the Volterra-Lotka argument. The appearance of the square root operation has required further analysis.

We have shown that the square root law follows rigorously if either of the following two properties applies to the readers.

- **Geometric property.** Institutions of different size are related like disks in a plane, whose active readers are confined to a thin layer near the outermost edge.\[ 6. \]

- **Bibliometric property.** Larger populations and smaller populations at similar institutions are described by a common universal law, and differ only in that larger institutions have an added population which is less active than those at the smaller institutions.

To sum up: the Volterra-Lotka approach, coupled with either a geometric picture or a bibliometric picture provides the theoretical foundation for a remarkably effective empirical law describing circulation at large academic libraries. The law has only two free parameter, which can be adjusted to describe specific groups of libraries. This law states that:

Circulation is proportional to the square root of the product of reader population, and fractional power of the collection size.
powers of variables, generalizing the kind of relation required by a Volterra-Lotka analysis. It is even more impressive to find not one, but two theoretical pictures which can explain the precise “square root” part of the law.

In the geometric picture, a rim of new growth accounts for the activity of the reader population. The alternative model, that the “latest additions” to readership represent a lower activity, is more plausible. Thus we prefer a bibliometric picture of the square root law, modeling reader activity as due to expansion by the addition of less active readers.

The proposed “Square Root Law” of library circulation unifies a bibliometric picture of readership and a widely observed rule governing interaction phenomena. We believe that this proposed “Square Root Law” can be tested, in its empirical form, on data describing other libraries and other time periods. The two adjustable parameter, $C_1$ and $b$ can be studied to explore its dependence on library type, and on time. In addition, the theoretical models can be extended, with suitable changes, to other uses of a library collection, and to other activities of the readers.

ACKNOWLEDGMENTS.

This research has been supported in part by the Council on Library Resources under Grant Number 6607.

REFERENCES


