Evidence of Wikipedia usage during the school year: A numerical study

Johnny Snyder
josnyder@coloradomesa.edu
Department of Business
Colorado Mesa University
Grand Junction, CO 81501 USA

Abstract

Wikipedia. One word, many opinions. This is especially true among academics, who hold strong opinions about the use of Wikipedia for scholarship. Are students using Wikipedia no matter what their faculty say? Numerical evidence suggests that American students are using Wikipedia. A mathematical model for the number of unique visitors to the English language Wikipedia site is developed and discussed. Results indicate that students are using Wikipedia during the school year, with usage dropping during the summer months. Future predictions and refinements to the model are addressed.

Keywords: Wikipedia, mathematical model, population model, web analytics

1. INTRODUCTION

In terms of web site popularity, Wikipedia has maintained a top ten position for the past six years (Alexa, 2014). This list includes such well-known sites such as Google, Facebook, and YouTube, illustrating the global reach (which equals influence) of the English language site, en.Wikipedia.org. Business Insider valued Wikipedia at $4 billion if it ran advertisements, demonstrating its potential economic influence in web culture (Wikipedia, 2014d). Many issues with the content of a Web 2.0 (Web 2.0 being defined as "user generated content") site can be called into question. These include the correctness of the information as well as the completeness of the information. Wikipedia has had its successes and failures over its lifetime (Jan. 15, 2001 – present) (Wikipedia, 2014a), but its popularity has never waned. In fact, the growth of Wikipedia from 2001 to 2011 has been incredible, with a peak usage of over 100 million unique visitors to the site in October, 2011 and again in March and October, 2014 (Compete, 2013; Compete, 2014). Current usage statistics estimate Wikipedia as having 18 billion page views per month (Wikipedia, 2014d). Wikipedia made it into the "Global Top Ten" in terms of website popularity in January, 2007 and has never left (Wikipedia, 2007d).

Today, traditional aged college students were entering elementary school when Wikipedia was emerging into the cyber world. In other words, today's students have never been without Wikipedia (as well as Google) during their academic lifetimes. One reason students might use Wikipedia for research is because it appears at the top of a page rank list when using a search engine with academic terms entered into the search utility. This paper explores the numerical evidence that students have been using Wikipedia and continue to use Wikipedia during the school year. What they are using it for and how faculty should react are separate issues.

2. LITERATURE REVIEW

Many papers, conference presentations, and conference themes center on the wiki-way and Web 2.0 technologies (Snyder, 2007). The most popular of these wikis is Wikipedia, the free online encyclopedia. Along with Wikipedia's

popularity comes Wikipedia's controversy. Issues with editorial oversight (the Essjay controversy) (Wikipedia, 2014b) and inaccurate content (the Seigenthaler incident) (Wikipedia, 2014c) gave Wikipedia negative publicity, but also enabled Wikipedia to be aware of, discuss, and improve weaknesses in its editorial structure.

The academic world has been both critical and supportive of Wikipedia. Some of the critics deride Wikipedia's editorial policies (Waters, 2007; Denning, Horning, Parnas, & Weinstein, 2005), while others advocate for its use as an educational tool (Moy, Locke, Coppola, & McNeil, 2010; Crovitz & Smoot, 2009). No matter which side of the debate one stands on, the fact remains: Students are using Wikipedia (Murley, 2008; Snyder, 2013). In fact, Head and Eisenberg (2010) reported that 82% of the college students in their study group used Wikipedia, while Snyder (2013) found that 77% of students surveyed reported Wikipedia usage. It has also been shown, in the academic environment, that faculty (70%) are using Wikipedia as well, adding to the school year usage patterns of the Wikipedia site being explored in this paper (Snyder, 2013).

3. METHODOLOGY

The data set for this study contains the number of unique visitors to the English language Wikipedia site from August, 2002 until November, 2014. Wikipedia was formally launched in January, 2001, but took a number of months before users started visiting the site in large numbers, and before usage patterns started to become evident.

Data illustrating the number of unique visitors to the site, en.wikipedia.org, will be analyzed to evaluate usage patterns throughout the calendar year. Multiplicative models including a seasonal component will be explored to illustrate the seasonal usage patterns in Wikipedia. Multiple forms of trend functions will be analyzed for best fit (minimize error) to find the best trend component for the model. Numerical results for the best fit model and a forecast for usage patterns into 2015 will be presented.

4. RESEARCH HYPOTHESIS

Wikipedia is used more during the traditional US school year than in the winter and summer breaks, indicating that the academic community might be utilizing the site.

5. RESULTS AND ANALYSIS

Data for the number of unique visitors to the English language site, https://en.wikipedia.org/wiki/Main_Page, has been collected for a number of years by harvesting from the web analytics sites alexa.com and compete.com (Alexa, 2007; Compete, 2013; Compete, 2014). A time series plot illustrating the usage pattern of Wikipedia is given in Graph 1. This graph shows an increasing trend (exponential) in usage patterns, with seasonal components appearing in 2005, and continuing throughout the remainder of the time series. The seasonal components (seasonal index values) will be computed using the ratio-to-moving average method presented in Groebner, Shannon, and Fry (2014). The model used to fit to the data will be the multiplicative model, due to the underlying non-linear growth trend with constant variability seasonal components (Groebner, et al., 2014). The model is given as equation (1), where T_t represents the trend component, S_t the seasonal component, C_t the cyclical component, and I_t the irregular or random component at time index, t.

$$Y_t = T_t\, S_t\, C_t\, I_t \qquad (1)$$

Seasonal components in the data series begin to appear in 2005, indicating that the data from June, 2005 through November, 2014 should be used for computation of the seasonal index values. Pronounced declines in usage appear in the summer months as well as December, indicating that when schools are not in session Wikipedia usage declines. After computing a 12-point moving average, the ratio-to-moving average method was used to compute the seasonal index values (Groebner, et al., 2014). These values are reported in Table 1 and plotted in Graph 2.

Graph 2 indicates a conspicuous trend in Wikipedia usage – it drops off a bit in December (holiday break from school) and significantly during the summer months (summer break from school). The negative values in Graph 2 indicate that the usage has dropped by 7 – 9% during the months of summer. This annual usage pattern (the data span the last ten years) indicates that the academic community has been using Wikipedia – arguably to do school work or school related research.

Once the index values have been constructed, the moving average is used to compute the underlying trend component. As can be seen

from Graph 1, the trend component is non-linear and sigmoidal. This makes physical sense, slow growth at the beginning while a name and reputation are being established, rapid (exponential) growth in the middle as Wikipedia's popularity soared, slow growth at the end, as the upper limit (carrying capacity) of the Internet's English speaking user population is reached. Table 2 lists the types of functions attempted and the resulting mean absolute deviation (MAD) value from each. The parameter values were computed using Excel's Solver function, with appropriate initial values for the parameters and the GRG Nonlinear solver routine (Excel, 2013).

Table 2 reveals that the Gompertz functional form works best (minimizes the MAD) for describing the underlying trend component (constructed from the 12 point moving average) of the curve in Graph 1. An illustration of this is given as Graph 3. Other functional forms, not presented in Table 2 were also tried (arctangent, hyperbolic tangent, four parameter logistic, rational), but the resulting MAD values were much higher than those in Table 2, and the graphical "fit" was noticeably inferior. Many of the functional forms in Table 2 originate from population dynamics, which is what is being modeled, population growth of a user group.

With the trend component and seasonal component of equation (1) determined, forecasting for the coming year can be predicted. The forecast is illustrated in Graph 4.

The forecast presented in Graph 4 continues the trend along with the seasonal patterns exhibited by Wikipedia users. The forecast predicts that Wikipedia will eclipse 100,000,000 unique visits per month in April, 2015 – under-estimating (in term of time) the actual peak usage observed. The actual peak usage was 100 million users in October, 2011, March 2014, and October of 2014 (during the academic semesters!). However, the multiplicative model does capture the behavior of the usage pattern and can be used to predict future user loads on the English language Wikipedia site. This model also illustrates that usage of Wikipedia.org is approaching a horizontal asymptote, indicating that (currently) one hundred million users will visit the English language Wikipedia site per month for the foreseeable future.

6. THE IRREGULAR COMPONENT, I_t

By observing Graph 4, one can visually identify a departure of the data set from the model. This usage spike occurs in October, 2011. This traffic spike took a bit of digging to uncover, but the discrepancy between the data and the model is 12,060,002 users. It turns out that Steve Jobs passed on October 5, 2011 (Wikipedia, 2016a), and his Wikipedia page took 7.4 million views on October 6, 2011 and 1.6 million views on October 7, 2011, accounting for 75% of the magnitude of the traffic spike (Wikipedia, 2016b). Incorporating this irregular component into the model, the error is reduced by 0.25%, to 1,520,184 (compare to Gompertz in Table 2). Other irregular components could be sought out and incorporated into the model to attempt a reduction of the error.

7. ASYMPTOTIC ENDING BEHAVIOR

As the usage pattern becomes less "Gompertz" and more "linear" (see Graph 3), another multiplicative model (equation 1) can be used with a linear trend component to analyze the current traffic patterns to the English language Wikipedia site. The data from January 2011 to the present is selected for this analysis, as this is where the data becomes more linear in its behavior (actually asymptotic, approaching a horizontal asymptote of user load). Doing so yields Table 3, seasonal index values for current usage patterns.

Table 3 illustrates similar behavior as the overall model, above average usage during the semesters, with below average usage during the summer months. This is shown in Graph 6.

The underlying trend component, T_t, is given by equation 2.

$$\hat{y} = 88{,}319{,}743 + 191{,}288t \qquad (2)$$

The slope in equation 3 indicates that, on average, Wikipedia can expect 191,288 new users per month to visit their site. This slope is significant with a p less than 0.01.

8. DISCUSSION

Wikipedia has been in the global top ten visited sites for a decade, indicating that people use and value the site (Alexa, 2007). In this, the information age, a Google search for an academic term will, more often than not, return a Wikipedia link in the first page of results. With

information (and misinformation) filling the Web, the academic community must be aware of students' usage of web resources, and explain the concept of information literacy in the Internet era.

Wikipedia is being referenced in textbooks on databases, e-commerce, and networking. There are conferences and academic papers covering the topic, illustrating the popularity of Wikipedia, and academic interest in the topic. Wikipedia, and the Wikimedia Foundation, have emerged as leaders in the Internet era, and must be acknowledged and managed as a classroom resource.

The model in this paper captures usage patterns on the English language Wikipedia site. Seasonal variations following US academic calendars were explored, measured, and recorded, illustrating that students and faculty seem to increase the usage load on Wikipedia during the academic year (and decrease it during the summer months).

The additive model was also explored (see equation (3)), under the assumption that the seasonal variations were constant over time (not an incorrect assumption; however, the growth in seasonal variations was not very pronounced).

$$Y_t = T_t + S_t + C_t + I_t \qquad (3)$$

Equation (3) also predicted lower usage patterns in the winter and summer months, as illustrated in Graph 5. However, the MAD for the additive model was 2,405,432 compared to a MAD of 2,243,533 for the multiplicative model (with Gompertz trend components). Both models work reasonably well and predict usage patterns for 2015 similarly.

Both models illustrate that usage of Wikipedia drops significantly in the summer and also in December. Thus, both models indicate that students and faculty (and perhaps others in the academic community) might be using the English language Wikipedia.

9. LIMITATIONS OF STUDY

Usage patterns for websites have been said to be overstated due to the users' clearing of cookies and other tracking software on personal computers. This could lead to an over-estimate as to the number of unique users of any particular site.

The observation that Wikipedia usage declines in the summer is not necessarily due to English speaking students utilizing the site. Further analysis of traffic patterns from top level domains would be necessary to confirm this hypothesis.

10. CONCLUSION

Usage patterns on the English language Wikipedia site indicate that students (and others in the academic environment) might be using Wikipedia. Previous studies (survey research) also indicate that students and faculty are using Wikipedia, if not for scholarly works, at least as a general introduction to a subject, or for fun!

The academic community needs to realize that Wikipedia might be emerging into the research realm, and thus management of this information source becomes critical. The fact that Wikipedia is one of the most visited websites in the world indicates that academics need to consider the information literacy effects of Wikipedia usage in the academic realm. Perhaps instead of forbidding its usage, academics could, instead, educate students on how to effectively use it as a research tool. Some topics for effective use could include exploiting the reference inks to the primary sources for the Wikipedia article, or utilizing the article ranking feature (one must be a registered Wikipedia user to enable this tool) to evaluate the quality of the article.

The Wikipedia management team could use this data model to schedule fund drives (March), system maintenance (July), and upgrades involved with running a global top-ten website.

11. REFERENCES

Alexa (2007). Wikipedia. Retrieved July 3, 2007 from:

http://www.alexa.com/data/details/traffic_details?url=wikipedia.org

Alexa. (2014). Wikipedia.org. Retrieved from: http://www.alexa.com/siteinfo/wikipedia.org

Compete. (2013). Wikipedia.org. Retrieved from: https://siteanalytics.compete.com/wikipedia.org/

Compete. (2014). Wikipedia.org. Retrieved from:

https://siteanalytics.compete.com/wikipedia.org/#.VJ1xJF7uwC0

Crovitz, D., & Smoot, W. (2009). Wikipedia: friend, not foe. English Journal, 98(3), 91-97.

Denning, P., Horning, J., Parnas, D., & Weinstein, L. (2005). Wikipedia Risks. Communications of the ACM, 48(12), 152.

Excel. (2013). Computer software. Microsoft: Redmond, WA.

Groebner, D., Shannon, P., & Fry, P. (2014). Business Statistics a Decision Making Approach (9th edition). Pearson: Boston.

Head, A., & Eisenberg, M. (2010). How today's college students use Wikipedia for course-related research. *First Monday*, 15(3).

Moy, C., Locke, J., Coppola, B., & McNeil, A. (2010). Improving science education and understanding through editing Wikipedia. Journal of Chemical Education, 87(11), 1159-1162.

Murley, D. (2008). In defense of Wikipedia. Law Library Journal 100(3), 593-607.

Snyder, J. (2013). Wikipedia in the Academic Environment: Faculty and Student Perspectives, *International Journal on E-Learning*. 12(3), 303-327.

Snyder, J. (2007). It's a wiki-world: utilizing Wikipedia as an academic reference. Proceedings of the 2007 Mountain Plains Management Conference.

Waters, N. (2007). Why you can't cite Wikipedia in my class. Communications of the ACM. 50(9), 15-17.

Wikipedia. (2014a). Wikipedia. Retrieved from: https://en.wikipedia.org/wiki/Wikipedia

Wikipedia. (2014b). Essjay controversy. Retrieved from: https://en.wikipedia.org/wiki/Essjay

Wikipedia. (2014c). Wikipedia biography controversy. Retrieved from: https://en.wikipedia.org/wiki/Wikipedia_biography_controversy

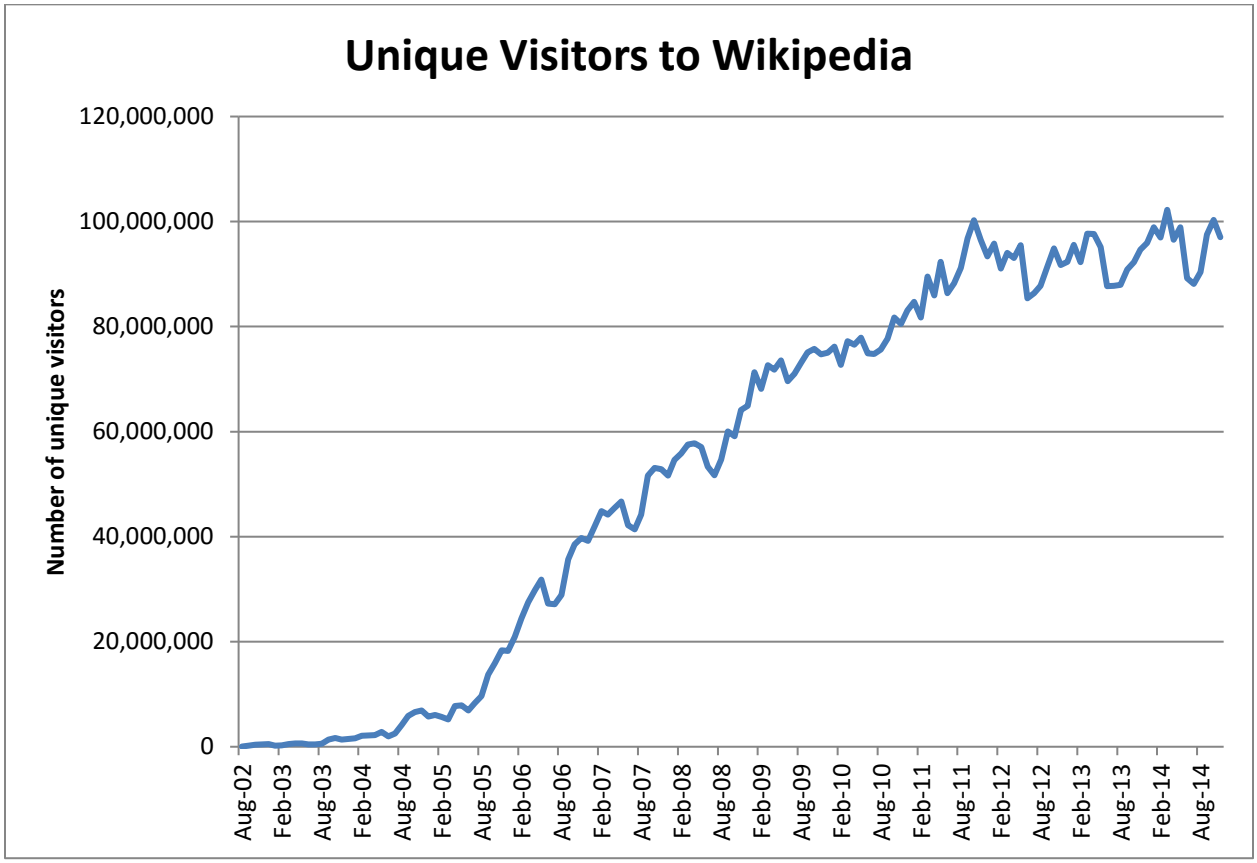
Wikipedia. (2014d). Wikipedia. Retrieved from: https://en.wikipedia.org/wiki/Wikipedia

Wikipedia. (2016a). Steve Jobs. Retrieved from: https://en.wikipedia.org/wiki/Steve_Jobs

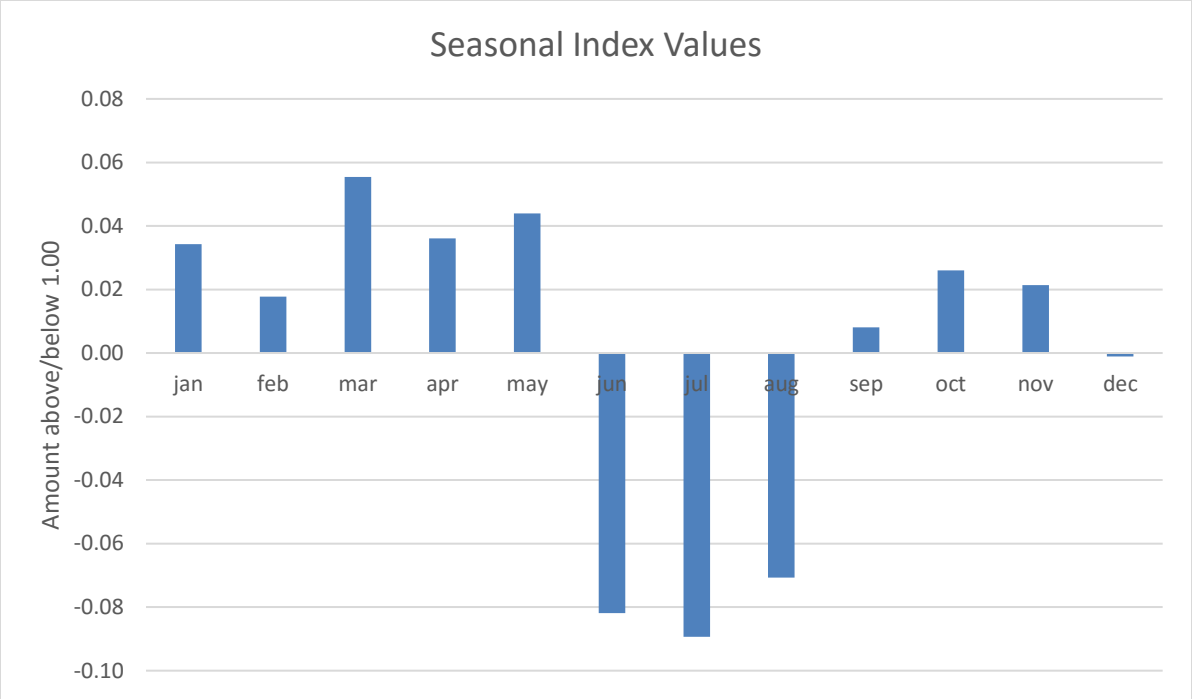
Wikipedia. (2016b). Wikipedia: Article Traffic Jumps. Retrieved from: https://en.wikipedia.org/wiki/Wikipedia:Article_traffic_jumps

ISSN: 2167-1508
v9 n4263

Appendix – graphs and tables



Graph 1
Number of unique users per month to the English language Wikipedia page



Graph 2
Seasonal index values for Wikipedia monthly usage

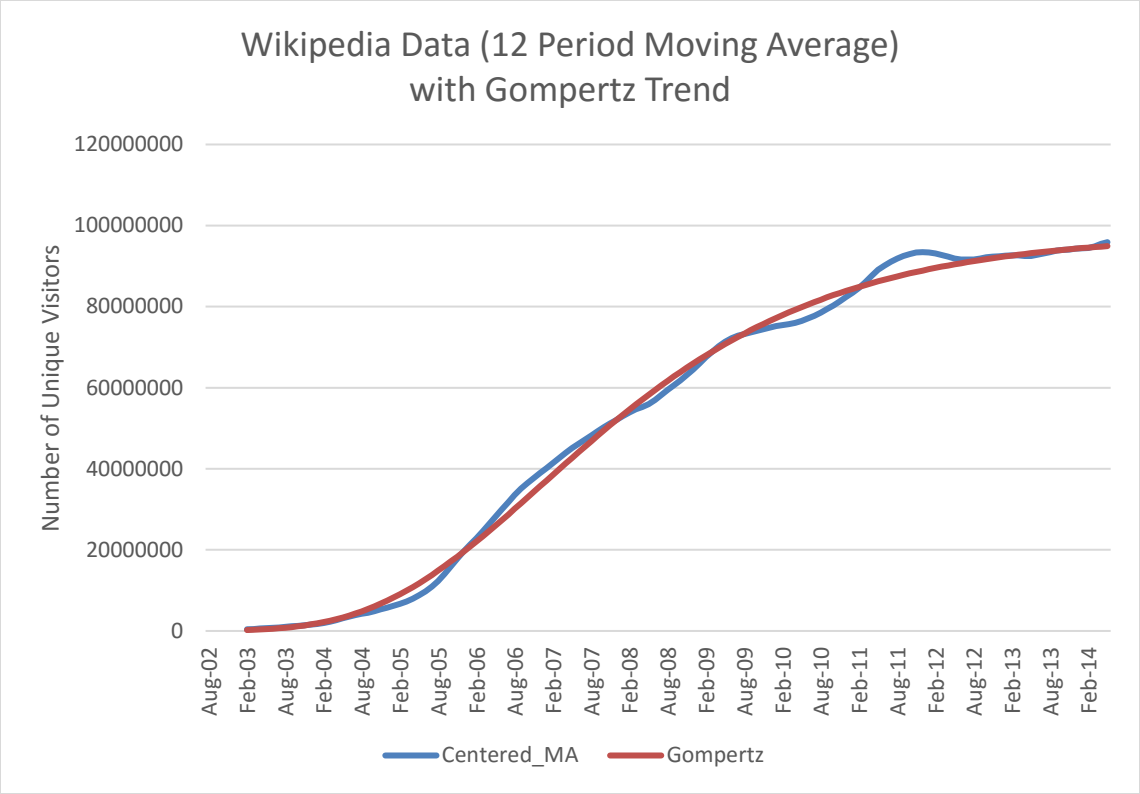
Month	Jan.	Feb.	March	April	May	June	July	Aug.	Sep.	Oct.	Nov.	Dec.
Index	1.03	1.02	1.06	1.04	1.04	0.92	0.91	0.93	1.01	1.03	1.02	1.00

Table 1
Seasonal index values for Wikipedia usage

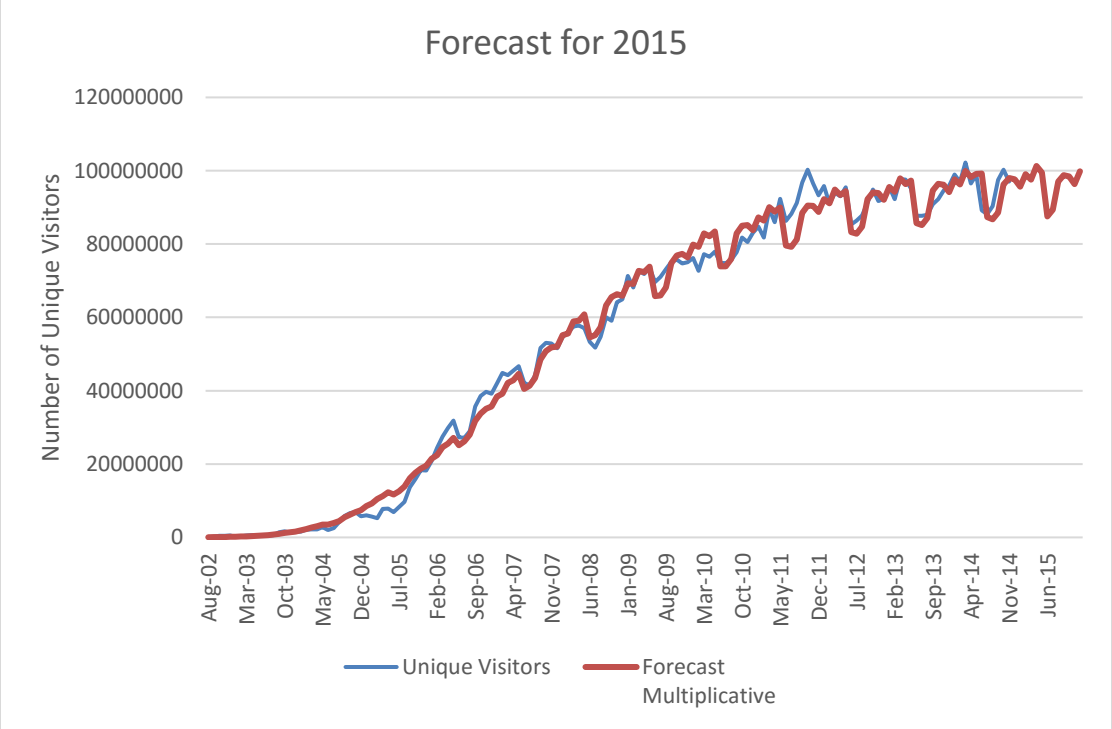
Function Name	**Functional Form**	**Parameters for Best Fit**	**MAD Value for Trend**
Logistic	$y = \frac{A}{1 + e^{-B(t-C)}}$	A = 94,878,486.34 B = 0.0594518 C = 62.58387	2,684,868.52
Generalized Logistic	$y = \frac{C - A}{[1 + B\ e^{-D(t-E)}]^{\frac{1}{\nu}}}$	A = 49.996 B = 0.011309 C = 97,763,000.08 D = 0.0400296 E = 81.60058 ν = 0.0356433	1,553,831.28
Algebraic	$y = \frac{At}{\sqrt{B + C\ t^2}}$	A = 6,324,926.92 B = 56.4970 C = 0.0008646	6,754,468.33
Gompertz	$y = Ae^{-Be^{-Ct}}$	A = 97,950,477.41 B = 7.6774389 C = 0.0390255	1,524,019.28

Table 2
Types of trend functions, parameters, and MAD values

ISSN: 2167-1508
v9 n4263



Graph 3
Wikipedia data set (moving average) with Gompertz trend component

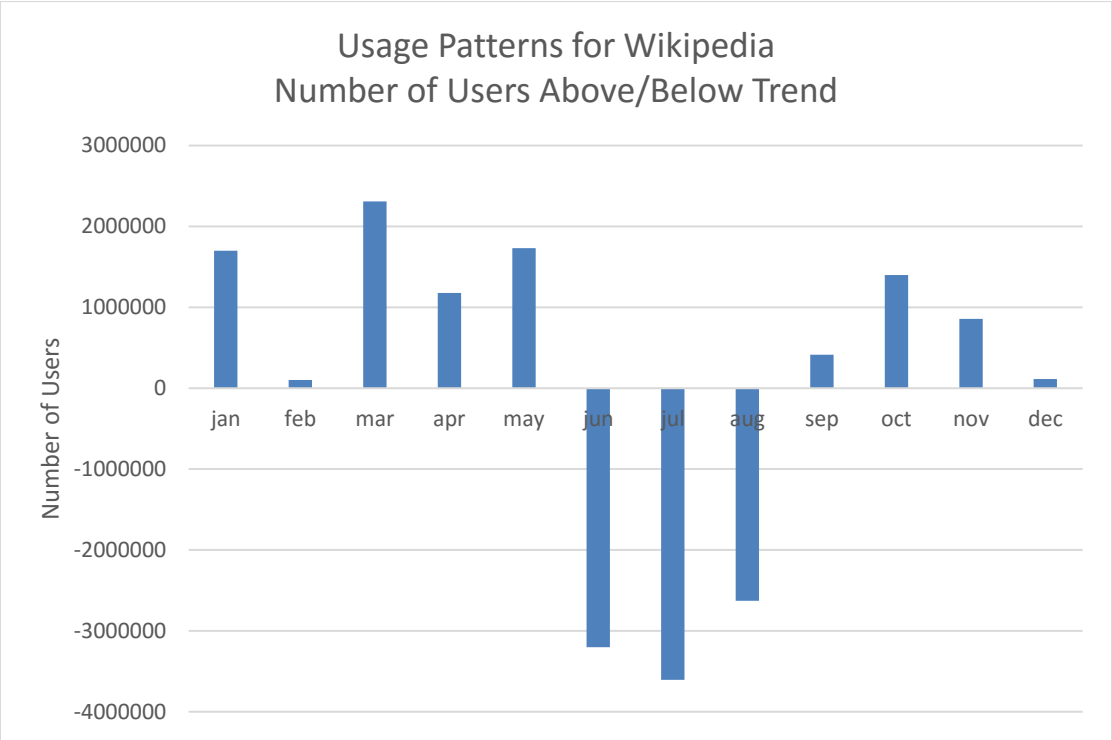


Graph 4
Forecasting Wikipedia usage for the coming year

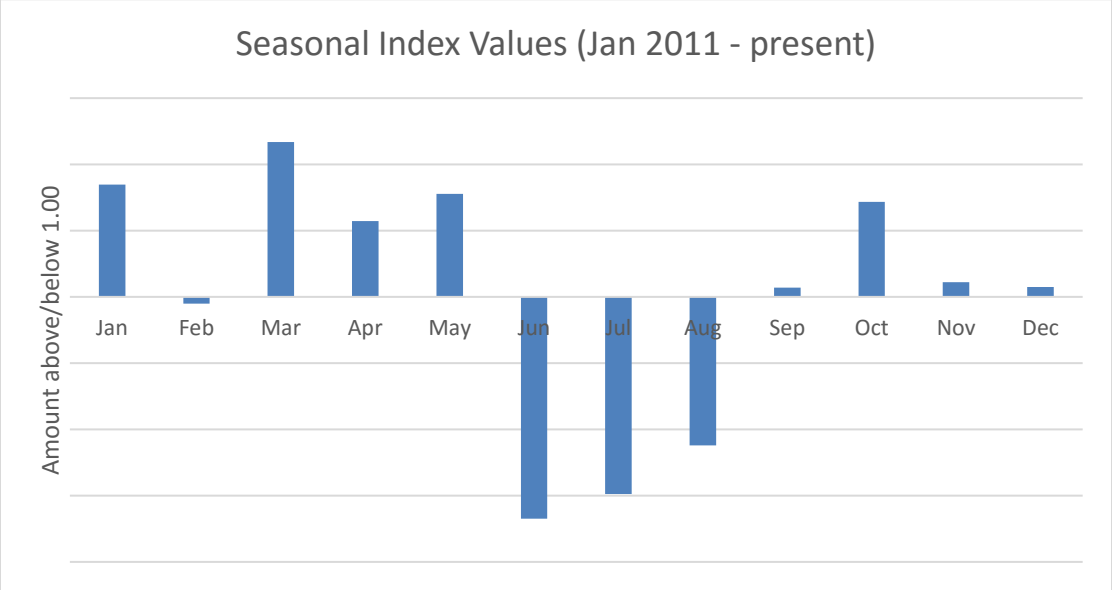
ISSN: 2167-1508
v9 n4263

Month	Jan.	Feb.	March	April	May	June	July	Aug.	Sep.	Oct.	Nov.	Dec.
Index	1.03	1.00	1.05	1.02	1.03	0.93	0.94	0.96	1.00	1.03	1.00	1.00

Table 3
Seasonal index values for Wikipedia usage (2011 – present)



Graph 5
Seasonal fluxuations under the additive model



Graph 6
Seasonal fluxuations under the trend model (Jan 2011 – present)

