UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE COMPUTAÇÃO

A Two-Phase Model for Wikipedia Growth

Jorge Stolfi

Technical Report - IC-09-45 - Relatório Técnico
November - 2009 - Novembro

The contents of this report are the sole responsibility of the authors.
O conteúdo do presente relatório é de única responsabilidade dos autores.

A Two-Phase Model for Wikipedia Growth

Jorge Stolfi
Institute of Computing, State University of Campinas
Caixa Postal 6176, 13081-970 Campinas, SP.
stolfi@ic.unicamp.br

2009-11-30 03:03:59 by stolfi

Abstract

The number of articles $N(t)$ in Wikipedia is quite accurately modeled as a function of time t by two exponential regimes or phases, with a relatively sharp transition over a one-year period centered on Janary 2006. The first regime has a positive rate constant $R_1 = +0.00217\,\text{day}^{-1}$, corresponding to a doubling time of about 10.5 months. The second regime has a negative rate constant $R_2 = -0.000407\,\text{day}^{-1}$, corresponding to a halving time of about 4.5 years. The model predicts that $N(t)$ will tend to a finite limit, a little over 8 million articles. We advance some possible explanations and implications of the negative rate.

1 Introduction

The English Wikipedia site [3] opened on January 15, 2001, and by the end of that month it had only 617 articles. By November 2009 it had grown to over 3,100,000 articles, which have suffered almost 350,000,000 edits. See figure 1.

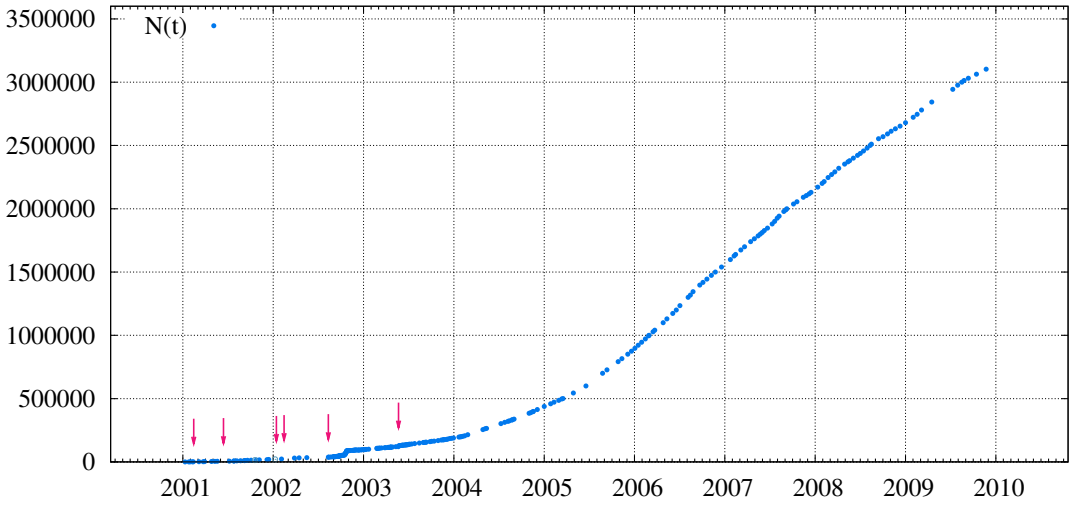

Figure 1: Number of articles in the English Wikipedia, as provided in their site. Open light dots are unreliable estimates. The arrows indicate changes in the article counting software (see text).

By November 2009 Wikipedia had over 150,000 volunteer editors who had contributed at least one edit in the past 30 days; and 4900 editors who made at least 120 edits each in the same period (4 edits per day, on average).

Wikipedia editors come from extremely varied backgounds and are widely spread around the globe. They are almost all self-appointed volunteers, with no pressure to contribute, and work in a largely uncoordinated way. While Wikipedia has machanisms to undo edits and delete or merge inappropriate articles, these tools are used almost exclusively against malicious edits. Creating a new article is still a very task.

Considering these features, and the large numbers of people and articles involved, it is not surprising that the size of Wikipedia has evolved in time in a rather smooth way. However, attempts to model and forecast that growth have been only partly successful and/or lacked a plusible justification.

Here we propose another model, which, in spite of being quite simple and plausible, fits the the English Wikipedia article counts of the past 8 years, and particularly those after 2006, with errors of only a few percent. The model also shows that the large short-term fluctuations of the growth rate, which were generally assumed to be random, are in large part due to a regular seasonal modulation factor.

2 Notations and definitions

The primary quantity that we analyze here is the number $N(t)$ of articles in the English Wikipedia, as a function of the time t since its creation.

The quantity $N(t)$ is defined implicitly by a Wikipedia program that supposedly counts only articles which have some informative contents [4]. The count therefore excludes non-content Wikipedia pages, such as the so-called *redirects* (analogous to the symbolic links of Linux file systems), *disambiguation pages* (menus of articles with similar names), *templates*

(akin to C language preprocessor macros) and the like. The count $N(t)$ also excludes articles that were created but deleted before time t.

2.1 Irregular events

The aspects of Wikipedia growth that can be meaningfully analyzed are those that result the combined efforts of a large number of human users. However, there were two isolated "anomalous" events in the history of Wikipedia that had a disproportionate effect on the article count, and which must be removed before analyzing those "normal" trends. See figure 2.

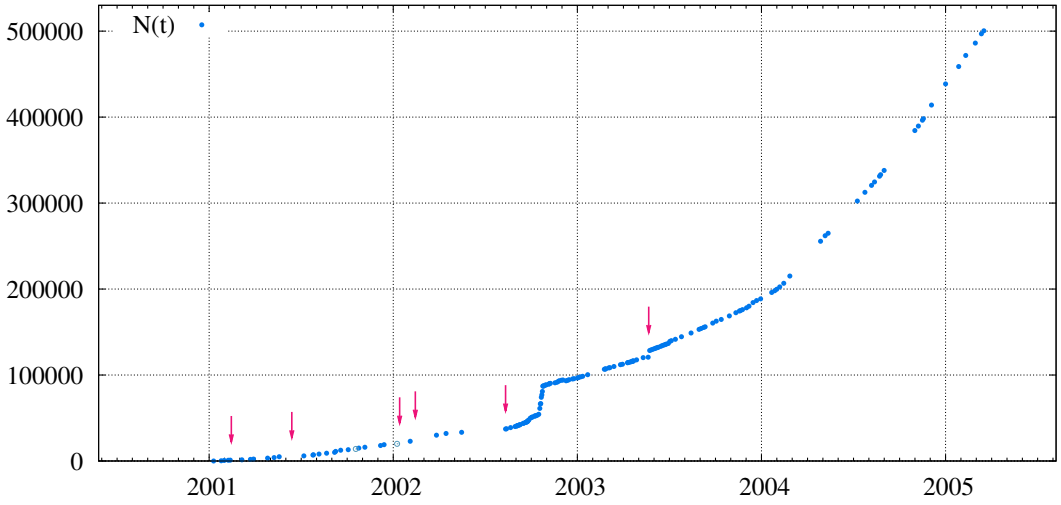

Figure 2: Number of articles in the English Wikipedia, 2001 to 2005 (magnified part of figure 1).

The first major anomalous event was the creation of about 33,000 new articles by an automated script (`rambot`), over an 8-day period centered on 2002-10-20. The articles provided basic Census data (such as location and population) for all US cities and towns. This resulted in a conspicuous jump in $N(t)$ around $t = 22$. See figure 3 (left).

The other significant "anomalous" event was the replacement of Wikipedia's article counting program, on or about 2003-03-23. That software had been replaced several times in previous years (on the approximate dates marked with arrows in figure 1), but the effect of those changes on the count $N(t)$ were indistinguishable from normal statistical flutuations. However, it appears that the counting program used between 2002-08-10 and 2003-03-23 (`mpacIII`) was missing a small fraction of the valid articles. So, when that program was replaced by an improved version (`mpag3.1`), the article count $N(t)$ experienced a sudden jump of about 6%. See figure 3 (left).

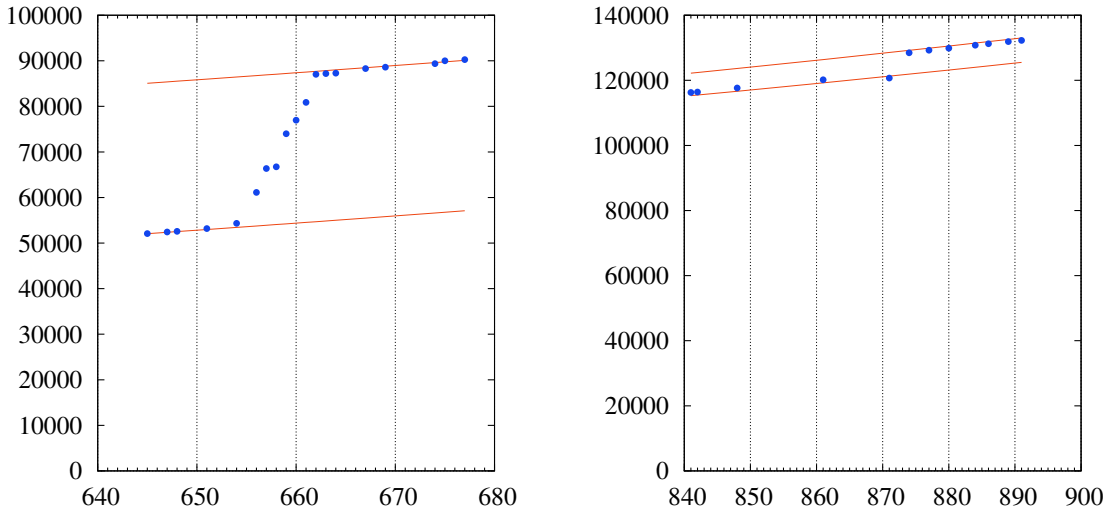

Figure 3: Major glitches in the English Wikipedia article count. The numbers in the horizontal axis are days elapsed since 2001-01-01. The jump at left, starting on day 654 (2009-10-17) is the work of the `rambot` script. The jump at right, on day 874 (2003-05-25) is a change in the article counting software.

2.2 Data cleanup and interpolation

In order to properly analyze the "normal" growth of Wikipedia over its whole history, it was therefore necessary co adjust the raw data so as to correct those two major glitches. The correction entailed subtracting the 33,000 `rambot` articles from all counts after 2002-10-20; and multiplying the counts between 2002-08-10 and 2003-03-23 by 1.06.

To simplify subsequent processing, we also resampled the $N(t)$ data, corrected as above, at regular intervals. For the sampling interval we chose the *lunisolar month* (mo) defined as precisely 28 calendar days. Thus one calendar year, in our analysis, is a little over 13 lunar months (more precisely $365.25/28 = 13.04$ mo). This period is reasonably close to the calendar months used in previous analyses, but has constant length and comprises a whole number of weeks.

This last feature is desirable because of the possibility of the growth rate $N'(t)$ having a regular variation according to the day of the week — as editors may be more likely to create new articles on weekends than on weekdays, or vide-versa. If such fluctuations with 7-day period do exist, the resampling of $N(t)$ at 30-day intervals (say) would turn them into regular variations in the estimated growth rate $N'(t)$, with a 7-*month* period. By sampling $N(t)$ at whole weeks apart, any fluctuations with 7-day period are largely averaged out.

After excluding the dubious points (open dots in figure 1), we interpolated them to obtain the value of $N(t)$ at the close of each 28-day period. See figure 4.

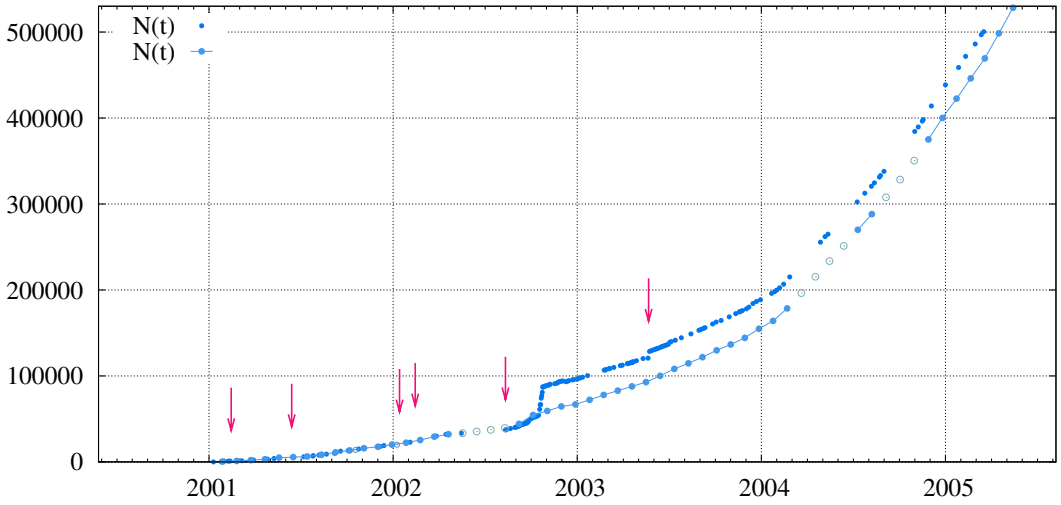

Figure 4: The English Wikipedia article counts $N(t)$: raw data (dark blue dots) and corrected and interpolated values (light blue solid dots). The open dots are unrelibale interpolated values. From 2004 on, the correction merely subtracted 33,000 from every count.

The interpolation was performed in logarithmic scale, which is to say, by the formula $N_0^{1-\alpha} N_1^{\alpha}$, where N_0, N_1 are two successive data points and α is the relative position of t betwen the corrsponding times t_0, t_1. The difference between this method and plain linear interpolation is noticeable only when interpolating across wide gaps (spanning two or more sampling periods); but in that case the interpolated value is unreliable anyway, and should be ignored in the analysis.

Since $N(t)$ is a cumulative count, this resampling did not lose any articles, even with this simple linear interpolation. Its main negative effects were to blur any small-scale features that might exist (such as weekly periodic fluctuations, or acivity peaks tied to specific external events) and to randomly shift the function values in time by up to a month. Neither of these drawbacks is relevant to our analysis.

3 Growth rate

The value of $N(t)$ reflects largely the past history of Wikipedia. To sense the current state and behavior of Wikipedia, a more relevant quantity is the time derivative $N'(t)$, the rate at which new articles are being created around time t. The unit we will use for measuring the growth rate is "articles per lunar month," which is fairly close to the "articles per calendar month" used in previosu studies. See figure 5.

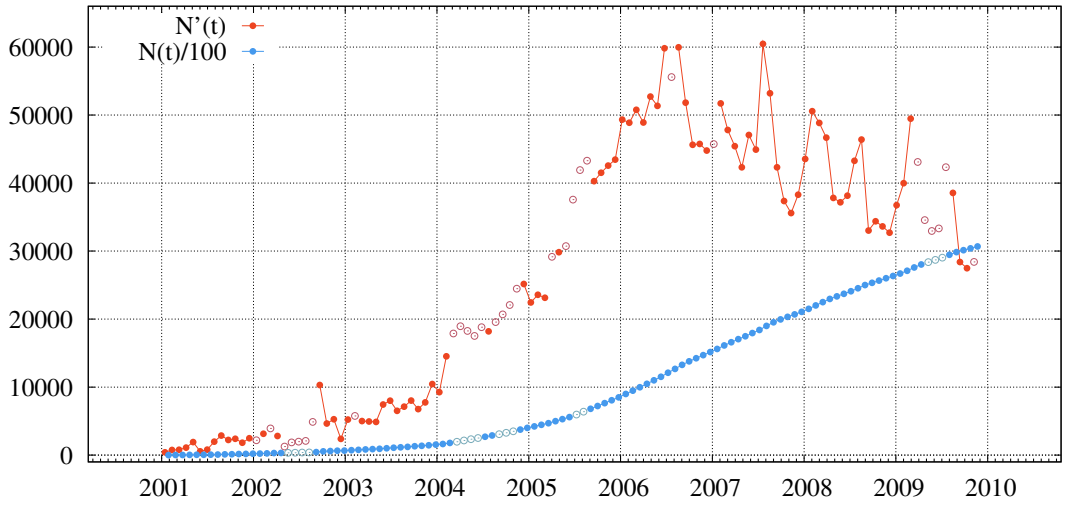

Figure 5: The net number $N'(t)$ of articles added to the English Wikipedia per lunar month (higher red curve). The cumulative article count $N(t)$ is also shown (lower blue curve). Grayish open dots indicate uncertain interpolated values of $N(t)$ or $N'(t)$.

The derivative N' was estimated by finite differences; namely, $N(t+28)-N(t)$ was taken to be the value of $N'(t+14)$, where t is measured in days. We flagged as "dubious" any values of $N(t)$ that had to be interpolated between measured data points that were more than 56 days apart; as well as any derivatives $N'(t)$ computed from those dubious sizes. The dubious values are grayed out in figure 5

4 Modeling Wikipedia's growth

4.1 Previous attempts

The growth of the English and other Wikipedias has been extensively analyzed by Wikipedia user *HenkvD*. [1]. Early models based on simple exponential growth failed to account for an inflection point in the plot, around 2006, that changed the shape of the curve from concave to convex. Subsequent attempts at fitting a logistic curve or Gaussian integral [1] were unable to match the shape of the growth rate N'.

A model that is similar ours was sketched by another Wikipedia editor, user *Wikid77*. [2]. He focused on the behavior after 2006, looking at the relative growth rate N'/N averaged over full years. He observed a steady decline of about 10% per year, but did not quite produce a formula. Instead he discussed whether this decline would continue and even argued that it should stabilize eventually as new articles would continually be needed.

4.2 Our model

Our model for the growth of the English Wikipedia divides its history in two phases, roughly from 2001 to 2005 and from 2006 to the present. Within each phase k, we find that the growth rate $N'(t)$ fits rather well an exponential function. See figure 6.

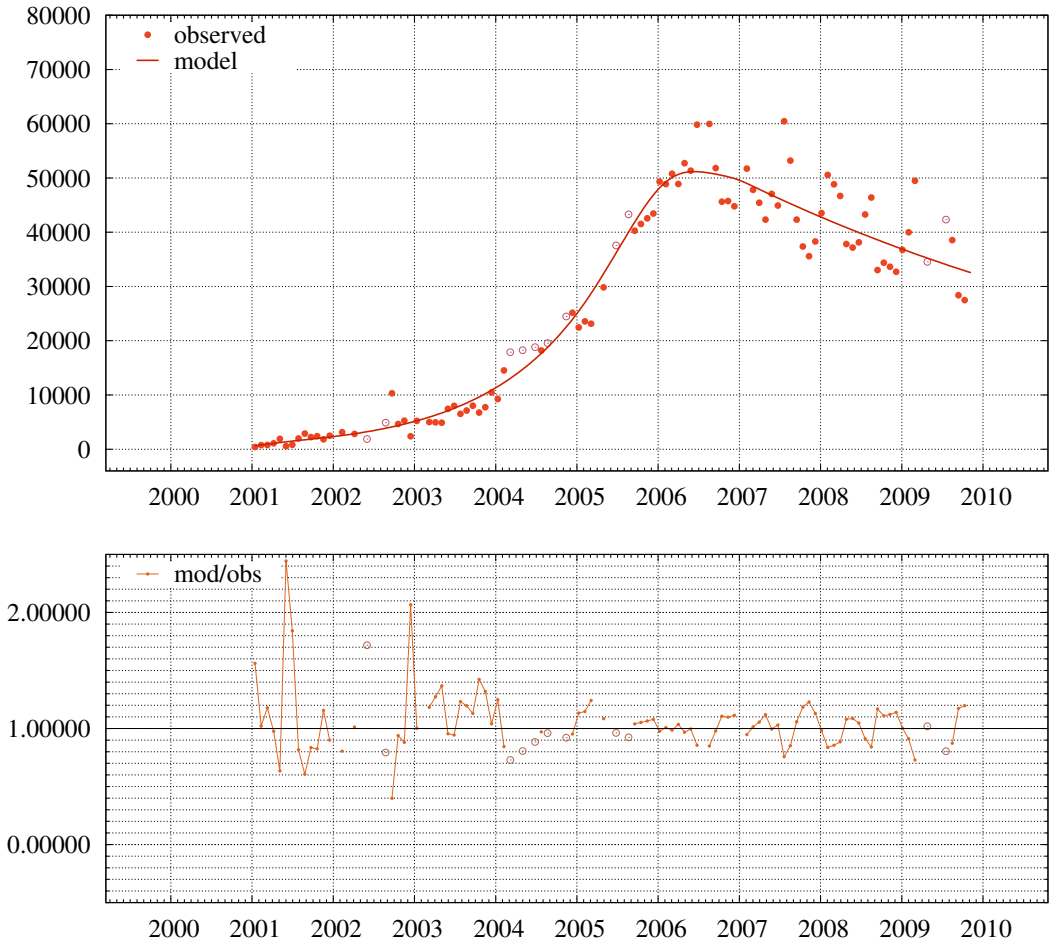

Figure 6: Plots of the Wikipedia growth rate $N'(t)$ (new articles per lunar month), showing the observed values (dots) and the proposed two-phase mathematical model (solid line). The bottom plot is the ratio between the model and observed values.

The division of history in two phases is fairly evident when we plot $N'(t)$ in logarithmic scale, where exponentials become straight lines. See figure 7.

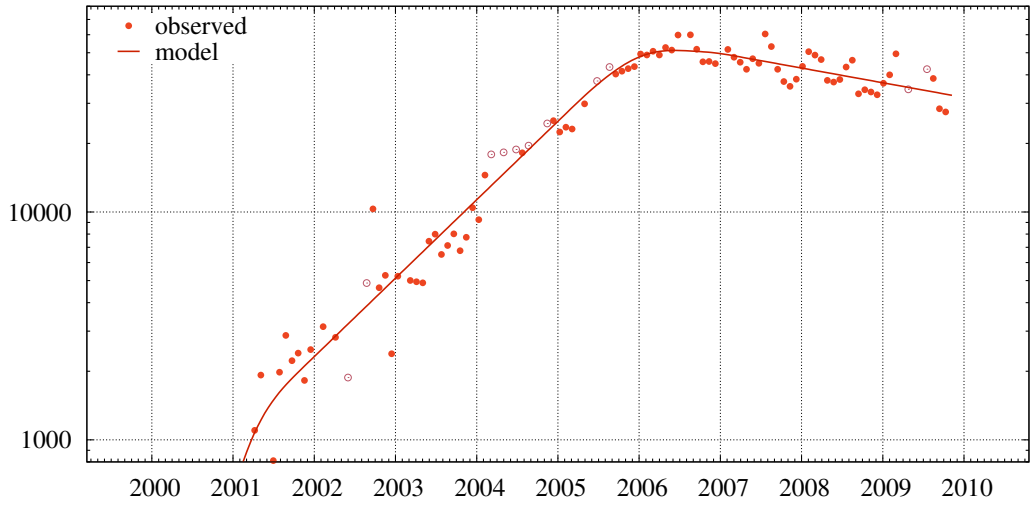

Figure 7: Log-scale plot of observed (dots) and modeled (solid line) values of the Wikipedia growth rate $N'(t)$ (new articles per lunar month).

As can be seen in figures 6 and 7, the transition between the two regimes was gradual but was complete within a period of about one year, between mid-2005 and mid-2006.

4.3 Mathematical formulas

For convenience, we modeled each phase as a horizontally shifted exponential $K_i \exp[R_i(t - S_i)]$, where S_i is an arbitrary reference date, K_i is the value of the exponential on that date, and R_i is the rate of increase. The first phase has rate $R_1 = +0.00217\,\mathrm{day}^{-1}$, which means that $N'(t)$ was doubling every 10.5 months or so. The second phase has $R_2 = -0.000407\,\mathrm{day}^{-1}$, which means that $N'(t)$ is now decaying to about one half every 4.5 years. The other parameters we chose are $S_1 = 0$, $K_1 = 1050\,\mathrm{art/day}$, $S_2 = 1826$, $K_2 = 57670$.

To model the transition, between the two regimes, we add the two exponentials, each multiplied by an appropriate windowing function. Namely, our model is

$$N'(t) = \sum_{i=1}^{2} K_i \exp[R_i(t - S_i)]\omega(B_i, E_i, t) \tag{1}$$

where B_i and E_i are the nominal beginning and ending times of phase i, and $\omega(B, E, t)$ is the windowing function for a given interval $[B \_ E]$. Namely, $\omega(B, E, t)$ is 1 when t is well inside that interval, is 0 when t is well outside that interval, and makes a smooth transition at either end. See figure 8

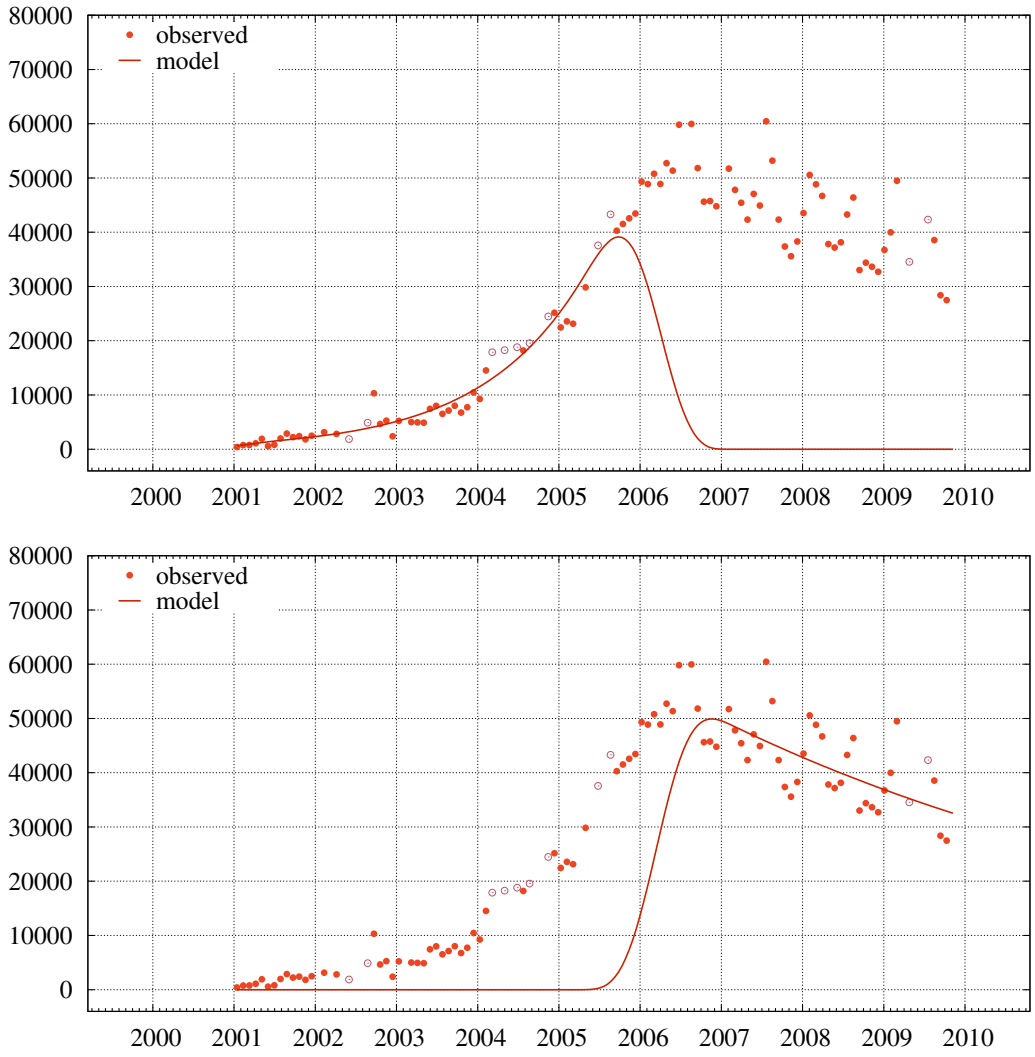

Figure 8: The two phases of the proposed model for $N'(t)$ (solid lines), compared to the observed values of $N'(t)$ (dots).

By integrating the model of the growth rate $N'(t)$, we get a model for the size $N(t)$ of the English Wikipedia. See figure 9.

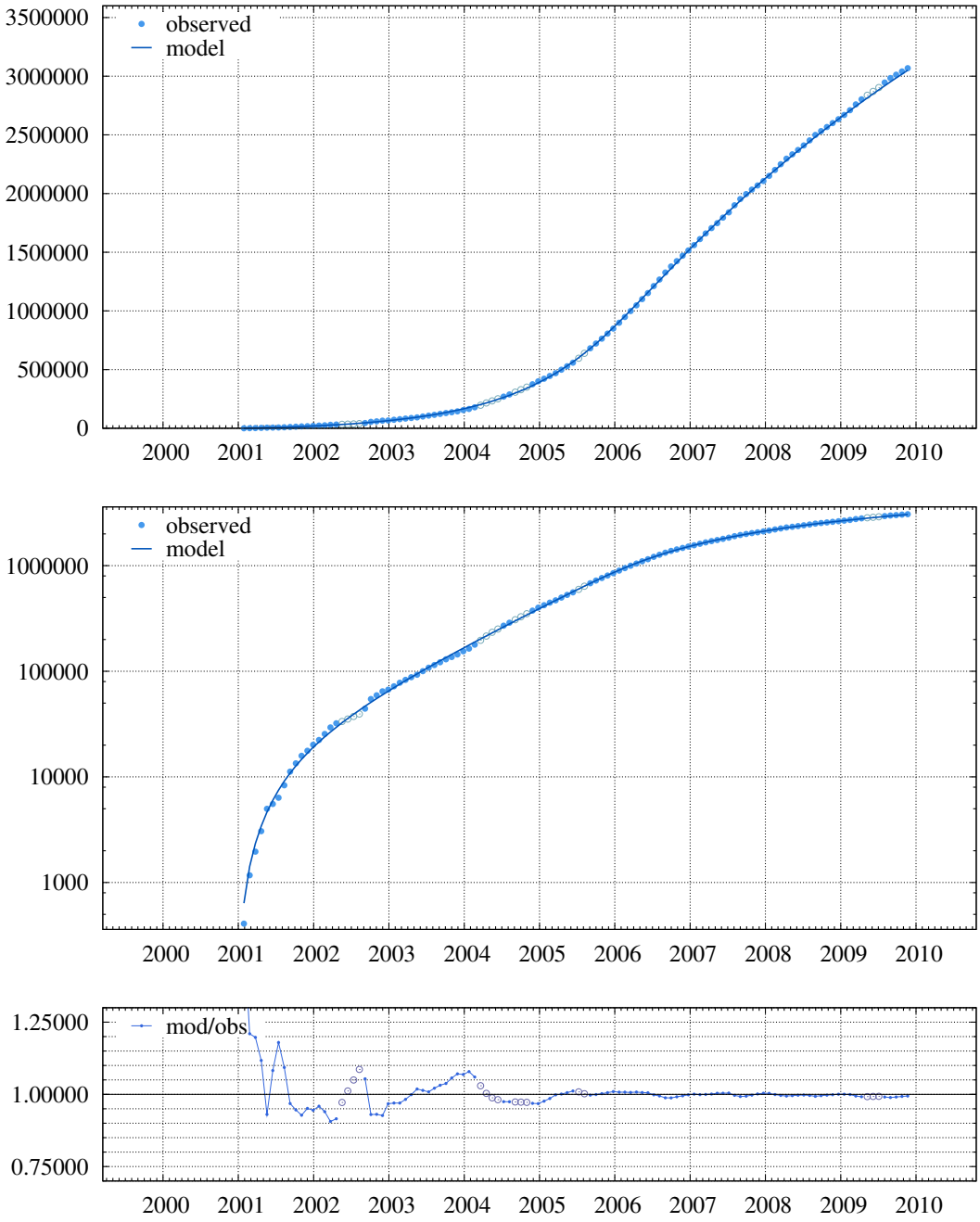

Figure 9: The number of articles $N(t)$ in the English Wikipedia, comparing the observed values (dots) and the values obtained by numerical integration of the proposed model of $N'(t)$ (solid line). the top and middle plots show $N(t)$ in linear and logarithmic scale, respectively. The bottom plot is the ratio between the predicted and observed values.

Note that within phase i, the article count $N(t)$ is the integral of the exponential $K_i \exp[R_i(t-S_i)]$, and is therefore a vertically shifted exponential $K_i/R_i \exp[R_i(t-S_i)] + C_i$ where C_i is a interpolation constant. Note also that a shifted exponential is *not* a straight line in log-scale plots; that is why the the exponential character of phase 2 is not apparent in figure 9.

4.4 The windowing function

For completeness, we describe here the windowing function ω that we used; even though our choice has no particular justification, and is probably not ideal.

The smoothness of the transition is controled by a duration parameter W, which we set at one calendar year (365.25 days). The function $\omega(B, E, t)$ is 0 if $t \le B - W$ or $t \ge E + W$, and 1 if $B + W \le t \le E - W$. (We assume that $E - B \ge 2W$.) When $B - W \le B + W$, then $\omega(B, E, t)$ is defined as $\frac{1}{2}(1 + \sigma((t - B)/W))$, where σ is defined below. Finally, when $E - W \le E + W$, $\omega(B, E, t)$ is $\frac{1}{2}(1 + \sigma((E - t)/W))$.

The auxiliary function $\sigma(z)$ is a sigmoid that varies from -1 to $+1$ as z varies over the same interval. It is defined as $\sin(\frac{\pi}{2}\sin(\frac{\pi}{2}z))$. See figure 10.

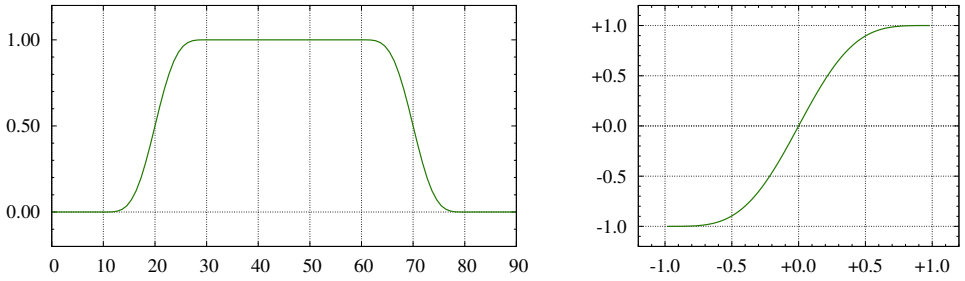

Figure 10: The windowing function $\omega(20, 50, t)$ for $W = 10$ (left) and the sigmoid $\sigma(z)$ (right).

5 Seasonal modulation

A distinctive feature of phase 2 is that the monthly-averaged growth rate $N'(t)$ fluctuates widely about the exponential trend. Examination of the data suggests that the fluctuations are clearly seasonal, with two two peaks per year — roughly in Spring and Fall of the northern hemisphere.

To include the seasonal fluctuations in our model, we multiply each exponential by a simple periodic modulation factor:

$$N'(t) = \sum_{i=1}^{2} K_i \exp[(R_i(t - S_i)] \exp[A_i \cos(2\pi(t/T_i - D_i))]\omega(B_i, E_i, t) \tag{2}$$

where A_i determines the amplitude of the fluctuations, T_i is their period, and D_i is a delay parameter between 0 and 1 that defines the relative position of the first maximum within a period. A good fit (not necessarily optimum) was obtained with $T_2 = 182.6$ day (one semester), $A_2 = 0.17$ (meaning a variation of approximately $\pm 17\%$ from the mean rate), and $D_2 = 0.15$ (corresponding to a peak in mid-February and another in mid-August).

No such fluctuations are discernible in phase 1, although they may have been masked by the natural random fluctuations. Therefore we set $A_1 = 0$, and the parameters T_1 and D_1 are immaterial. See figure 11

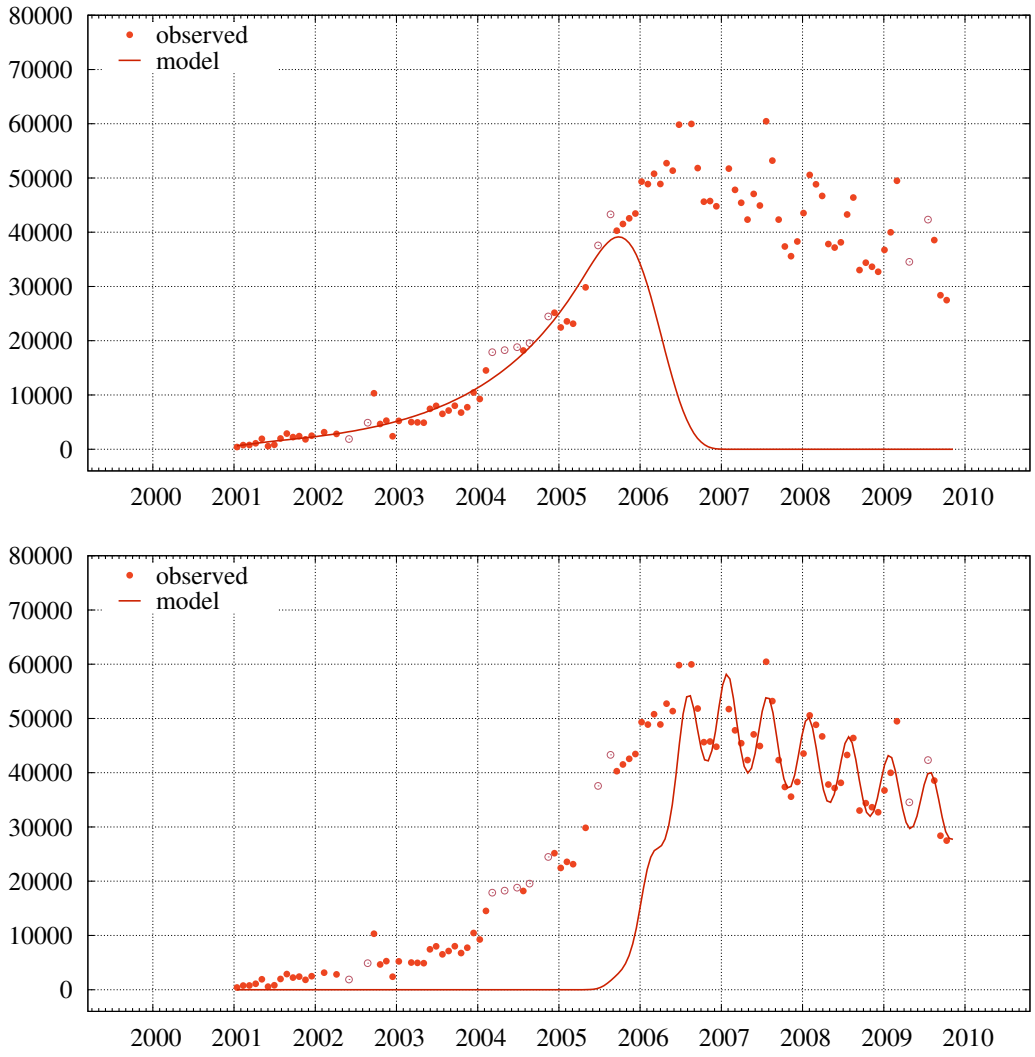

Figure 11: The two phases of the proposed model for $N'(t)$, augmented with seasonal factors (solid lines) and compared to the observed data (dots).

This seasonal correction improves the fit between the model and the observed data. See figure 12.

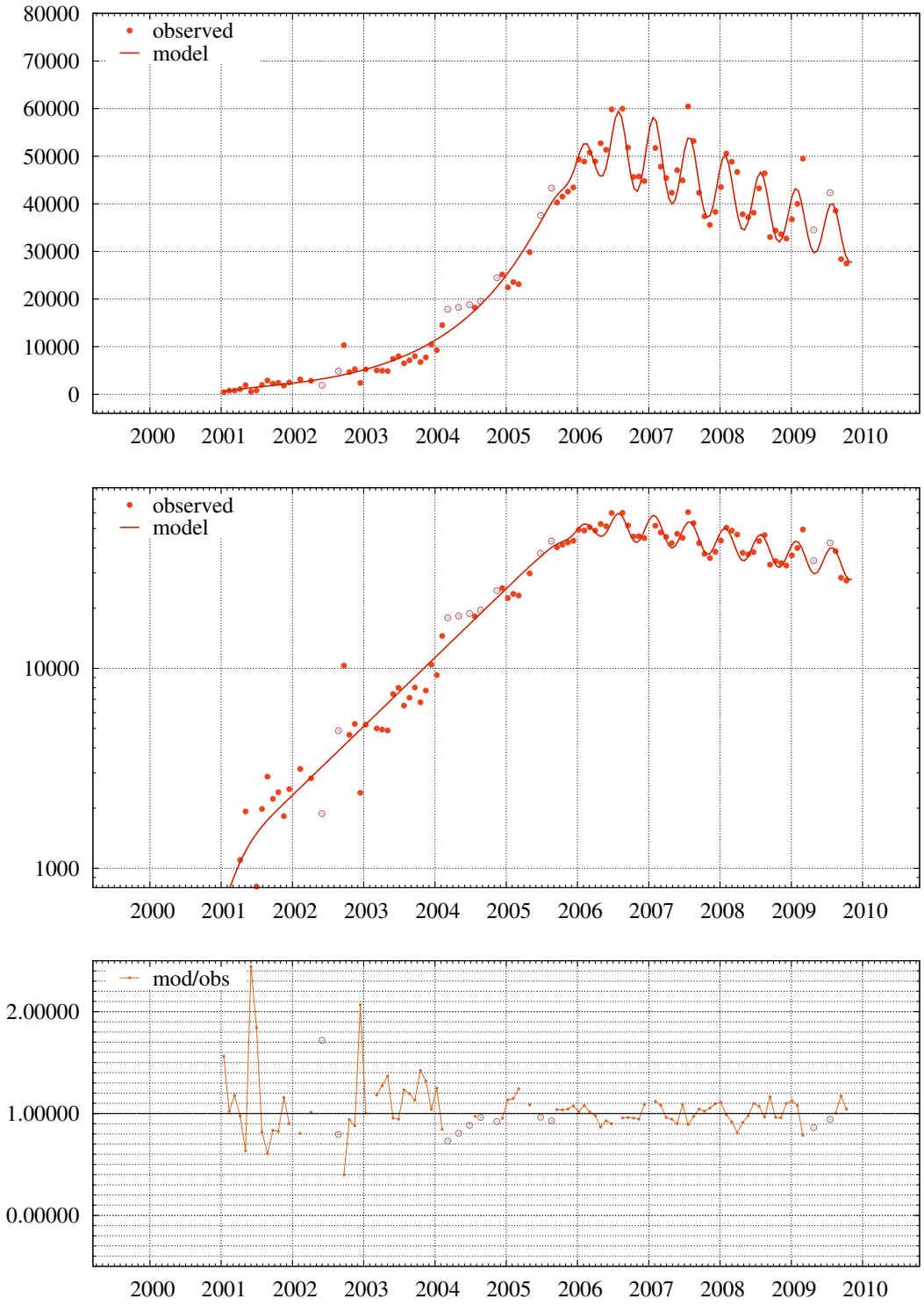

Figure 12: The monthly article creation rate $N'(t)$ in linear scale (top plot) and logarithmic scale (middle plot), showing the observed values (dots) and the values computed by the seasonally modulated two-phase mathematical model (solid lines). The bottom plot is the ratio between the modeled and observed values.

This modification also improves the accuracy of the model for the cumulative count $N(t)$.

See figure 13.

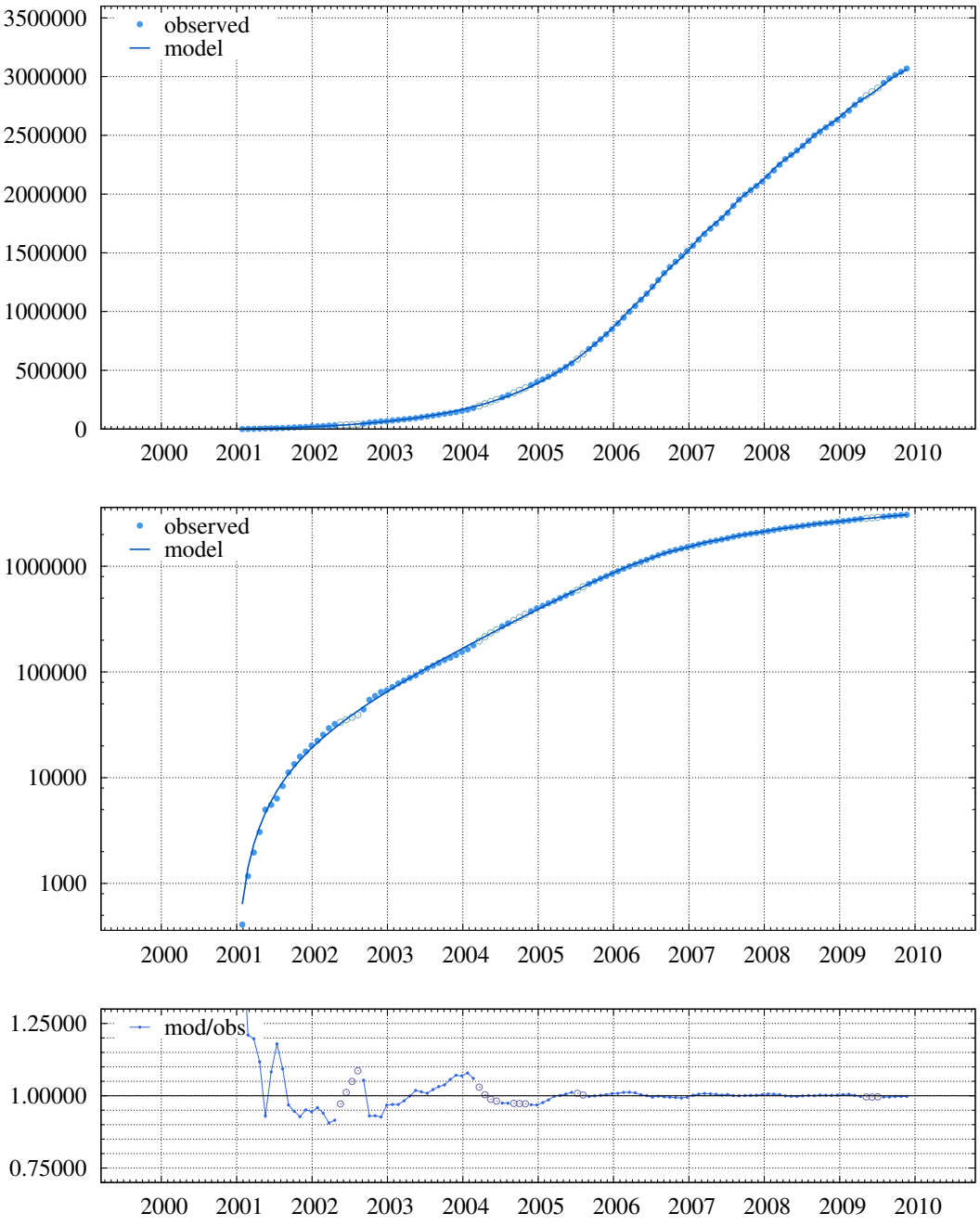

Figure 13: The cumulative article count $N(t)$ in linear scale (top plot) and logarithmic scale (middle plot), showing the observed values (dots) and the values obtained by integrating the seasonally modulated two-phase mathematical model (solid lines). The bottom plot is the ratio between the modeled and observed values.

6 Discussion

6.1 Accuracy of the model

With only 7 effective parameters $(K_1, R_1, E_1, W, B_2, K_2, R_2)$, the basic two-exponentials model can describe the evolution of the Wikipedia article count $N(t)$ and its derivative $N'(t)$ with surprising accuracy, especially in the period dominated by the second phase (middle of 2006 to the end of 2009). During this period, the prediction errors in N' are about $\pm 20\%$ and those of N are about $\pm 2\%$. Inclusion of the seasonal factors in phase 2 adds another three parameters (A_2, T_2, D_2) and reduces the errors in N' to $\pm 10\%$ and those of N to $\pm 1\%$.

The model is less accurate during phase 1 (2001 to 2005), but the remaining variation in N' seems largely random.

6.2 Adequacy and justifications of a two-phase model

The use of a two-phase model seems well-justified in view of plots like figure 7.

In the contexts of previous analyses, it has generally been assumed that the observed fall of N' since 2006 was a "natural" phenomenon, due solely to the evolution of the contents of Wikipedia and the editors's reaction to that. One optimistic conjecture was that, as the coverage of Wikipedia kept expanding, the editors felt less need or opportunity to create new articles, because all important ones had been created already. In other words, N' was falling because Wikipedia was aproaching completeness. Based on that conjecture, several (single-phase) "saturation curve" models were tried; but, as more data became available, they became harder and harder to fit.

However, the sharp phase transition visible in the plots (and confirmed by our model) makes that explanation seem unlikely. The drastic switch — from a steady doubling every 11 months to a steady halving every 4 years — is too abrupt to be due to the evolution of Wikipedia's article base, or to events exernal to Wikipedia. A more likely explanation is that a single event in Wikipedia support structure, between 2005 and 2006, drastically changed the habits of the editors — in particular, their "fertility" for article creation. The event may have been a software or hardware change, or (morelikely) a change in general Wikipedia policies — such as the erquirement for references, or the article assessment and award structure.

6.3 Conjectures about the mechanism of the switch

The article creation rate of an individual editor is not likely to grow epponentially as years go by. After a short initial period, it usually falls or remains bounded. Therefore, the exponential growth seen from 2001 to 2005 can only be explained by a similar exponential increase in Wikiedia's army of editors. To explain the latter, we conjecture that the growing numeber ad quality of articles led to an exponential increase in its readership. As more people spent more time reading and exploring Wikipedia, more people felt tempted to contribute. So, the two-phase model above suggests that the fall in N' since 2006 is probably due to a decrease in the *number* of active editors, not just in their fertility.

One conjecture that is suggested from this model is that, after that event, article creation became much harder. As a consequence, those editors who were connected to academic institutions — which may be the majority of them — had to limit their Wikipedia editing to school vacations or the months just after them. At the same time, the increased cost has been causing many editors to drop off, and has deterred new editors from joining.

7 Implications for the future of Wikipedia

As long as the growth rate $N'(t)$ can be modeled by the phase-2 decaying exponential $K_2 \exp[R_2(t - S_2)]$, the article size N will continue to grow but at an ever slower pace, as

$$N(t) = L + \frac{K_2}{28R_2} \exp[R_2(t - S_2)] \tag{3}$$

where L is the ultimate limiting size. (Note that R_2 is negative, and $N'(t)$ is 28 times the derivative of N.) We will ignore the seasonal factor since it will have little effect.

The limit L is readily computed as

$$L = N(t^*) - \frac{K_2}{28R_2} \exp[R_2(t^* - S_2)] \tag{4}$$

where t^* is any date after January 2007, when phase 2 was fully established. If we pick $t^* = 2922$ (2009-01-01) we have $N(t^*) \approx 2,680,000$, and we get that the limiting size of Wikipedia will be $L \approx$ 5.9 million articles — or only 2.8 million more than its present size.

This is a worrisome conclusion. As noted above, the falling trend in N' is not due to lack of missing articles, but rather to a shrinking of the pool of article creators — if not of editors. There are easily many millions of potential articles that are missing from Wikipedia but which ought to be created. There are many categories of things which are still poorly covered by wikipedia: living species, chemical products, companies, politicians, rivers, mountains and towns of the world, schools, churches and other historical buildings, books and their authors, paitings, plays, movies, ... In each of those categories, and of many others, there is easily a million items that still lack a Wikipedia article, but should have one. So, if the current trend persists, Wikipedia will eventually freeze out in a still very incomplete state.

8 Conclusions and future work

The two-phase model is simple, accurate and admits natural explanations. It strongly indicates that a major change occurred in Wikipedia between 2005 and 2006, which (1) turned a steady positive exponential trend into a steady negative (linear or exponential) trend, and (2) introduced a semestral variation of ±15% in the article creation rate.

This model can surely be improved in many ways. The parameters given in this report were adjusted by hand; a non-linear least-squares optimization, with the proper weight assigned to each datum, is likely to yield a better-fitting model.

The data seem to suggest that phase 1 is actually two sub-phases, with slightly different growth rates — a lower one from 2001 to mid-2003, and a slightly higer one from mid-2003 to 2005.

The simple sinusoidal factor in formula (2) could be replaced by a more complicated but still periodic seasonal factor, e.g. $\exp[\sum_{r=1}^{n} A_{ir} \cos(2\pi(rt/365.25 - D_{ir}))]$, for some $n \leq 12$. (The current model has only the term $r = 2$). Also, the windowing function ω that we used has no logical justification.

One would expect the growth rate N' to exhibit strong periodic fluctuations with a 7-day period. It would be interesting to know whether there was any change in the presence or intensity of such weekly rythms between 2005 and 2006.

Further insights are likely to come from the analysis of other measurements of Wikipedia, such as the number and size of edits (other than article creation events), the distribution of article sizes, the number of active editors and their editing patterns, etc.. All those quantities should be examined to see whether they too would fit a simple two-phase model.

Acknowledgements

This research was supported in part by CNPq (grant 301016/92-5).

References

[1] User:HenkvD. Wikipedia:Modelling Wikipedia's growth. English Wikipedia internal article at `http://en.wikipedia.org/wiki/Wikipedia:Modelling_Wikipedia's_growth`. Acessed on 2009-11-29. See also `http://en.wikipedia.org/wiki/User:henkvD`., 2009.

[2] User:Wikid77. Wikipedia:modelling wikipedia extended growth. English Wikipedia internal article at `http://en.wikipedia.org/wiki/Wikipedia:Modelling_Wikipedia_extended_growth`. Acessed on 2009-11-29. See also `http://en.wikipedia.org/wiki/User:Wikid77`., 2009.

[3] James Donal "Jimbo" Wales. Wikipedia, the free encyclopedia. Site at `http://en.wikipedia.org/wiki/` accessed on 2009-11-29, 2001.

[4] Wikipedia editors. Wikipedia:Size of Wikipedia. English Wikipedia internal article at `http://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia`. Acessed on 2009-11-29., 2009.