Who Uses gretl? An Analysis of the SourceForge Download Data

Riccardo (Jack) Lucchetti

Dipartimento di Economia - Università Politecnica delle Marche - Ancona, Italy
r.lucchetti@univpm.it

Abstract. This paper analyses the SourceForge download data to infer some characteristics of the population of gretl users. The rising number of downloads indicates Gretl's strong popularity as a teaching tool; however, despite the vast improvements in its features and performance, gretl's perceived status as a computational platform for research does not seem to be firmly established as yet, although this may change in the medium-long run.

1 Introduction

In the past few years, gretl has undoubtably come a long way in terms of features: thanks to constant feedback by a loyal user base and, most importantly, to Allin Cottrell's incredible committment and outstanding productivity, what used to be considered little more than a toy package now comprises a range of features which equal, and in some cases surpass, those found in commercial statistical programs. For example, the scope and efficiency of the routines for estimating GARCH-like models written for gretl by Balietti [2] are unrivalled by any other free software package.

The question I ask myself in this paper is: how has the gretl userbase evolved in response to the new features and the overall increase in usability of the package? In order to provide an answer, I will analyse the download data from SourceForge.

2 The SourceForge data

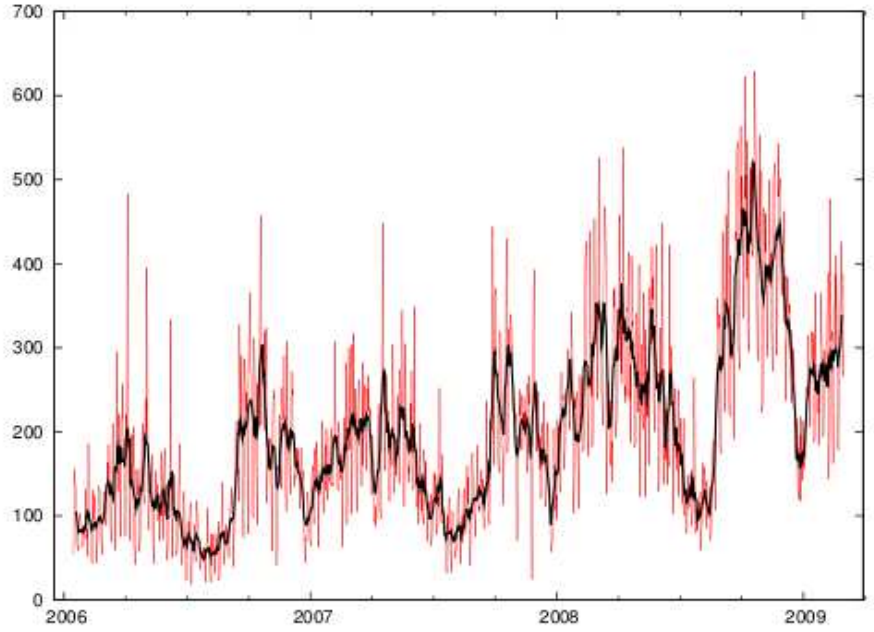
SourceForge is arguably the largest and most important hosting site for Free Software projects. It currently hosts tens of thousands of projects, among which hugely successful ones such as eMule and 7-zip.

Gretl ranks, at present, about 930th for number of total downloads, which amounts to about 300,000. However, it must be stressed that the number of downloads may not match the number of users for several reasons:

- The total number of downloads refers to all versions; a user who installed 10 versions of gretl on a computer counts as 10 downloads;

– some people may download the same version more than once, to install on different machines, possibly on different architectures;
– some people may download the "installer" once and give it to other people or make it available over a LAN;
– some gretl users may bypass SourceForge completely, especially those linux users who prefer to use the pre-packaged version of gretl offered by their linux distribution of choice (eg Debian and Ubuntu).

Fig. 1. Gretl daily downloads



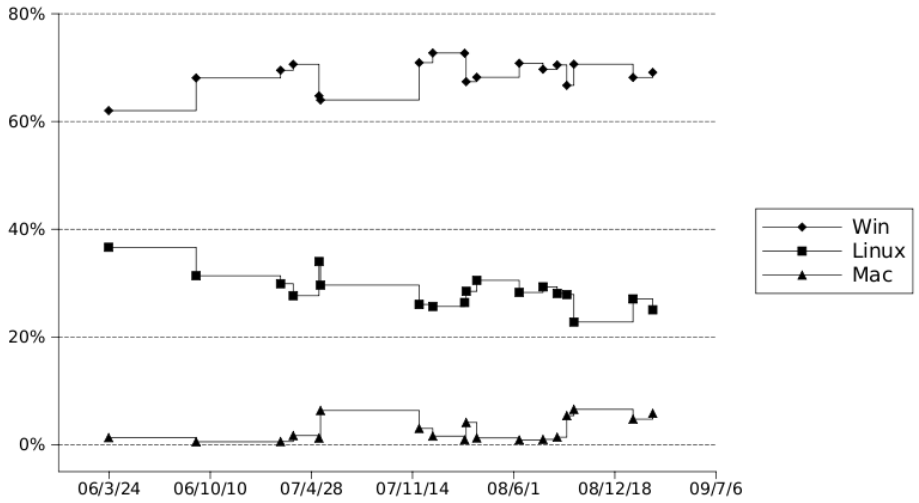
With this proviso, the number of daily downloads from SourceForge in the period 2006/1/15–2009/2/28 is shown in figure 1, together with a 7-day centered moving average. A few features are apparent:

– strong upward trend
– strong weekday effect
– strong seasonality

The upward trend is partly a consequence of the increase in popularity of Free Software at large, partly due to gretl's expanding user base (as confirmed by other anectodal evidence, such as the number of mailing list subscriptions

and so on); the weekday effect is unsurprising and of little interest in itself. The seasonal effect is, most likely, linked to the customary organisation of university courses. Roughly, the pattern is that downloads are low during the summer and the Christmas and Easter periods, while they spike up in September-October and in late February-March, which coincide with the beginning of terms in most universities, at least in the Northern hemisphere.

Fig. 2. Gretl downloads by architecture



It appears that gretl has reached, in three years, a much broader audience. This is, to some extent, confirmed by disaggregating the downloads of the various releases by platform.[1] Figure 2 shows the shares of downloads by platforms for each release from the beginning of 2006 to February 2009. The most striking feature of figure 2 is the decline of linux: I take this as evidence that gretl is now less perceived as a "geek" application than it used to be in 2006. Also notable is the increase of downloads for the Macintosh platform, which parallels the surge of its popularity among the general public[2].

Table 1. Weekly model

	Coefficient	Std. Error	t-ratio	p-value
const	2.82207	0.33705	8.3729	0.0000
y_1	0.38438	0.07337	5.2386	0.0000
tim	0.22222	0.03123	7.1154	0.0000
rel_1	0.11490	0.04764	2.4117	0.0171
a1	0.01668	0.02061	0.8093	0.4196
a2	0.13703	0.02479	5.5284	0.0000
s1	−0.01376	0.02257	−0.6094	0.5432
s2	−0.24451	0.03247	−7.5298	0.0000
q1	0.09449	0.02238	4.2213	0.0000
q2	0.03704	0.01959	1.8908	0.0606

Mean dependent var	5.168366	S.D. dependent var	0.499487
Sum squared resid	4.522887	S.E. of regression	0.172499
R^2	0.887399	Adjusted R^2	0.880732
$F(9, 152)$	133.1000	P-value(F)	1.99e–67
Log-likelihood	59.98606	Akaike criterion	−99.97212
Schwarz criterion	−69.09616	Hannan–Quinn	−87.43601
$\hat{\rho}$	−0.006229	Durbin's h	−0.216540

LM test for autocorrelation up to order 7 –
  Test statistic: LMF = 1.16168
  with p-value = $P(F(7, 145) > 1.16168)$ = 0.328463

Test for ARCH of order 1 –
  Test statistic: LM = 2.85511
  with p-value = $P(\chi^2(1) > 2.85511)$ = 0.091084

Koenker test for heteroskedasticity –
  Test statistic: LM = 7.84029
  with p-value = $P(\chi^2(9) > 7.84029)$ = 0.550318

Test for normality of residual –
  Test statistic: $\chi^2(2)$ = 1.94429
  with p-value = 0.378271

QLR test for structural break –
  Test statistic: $F_{\max}(10, 142)$ = 2.12638 (06/11/06)
  (10 percent critical value = 2.48)

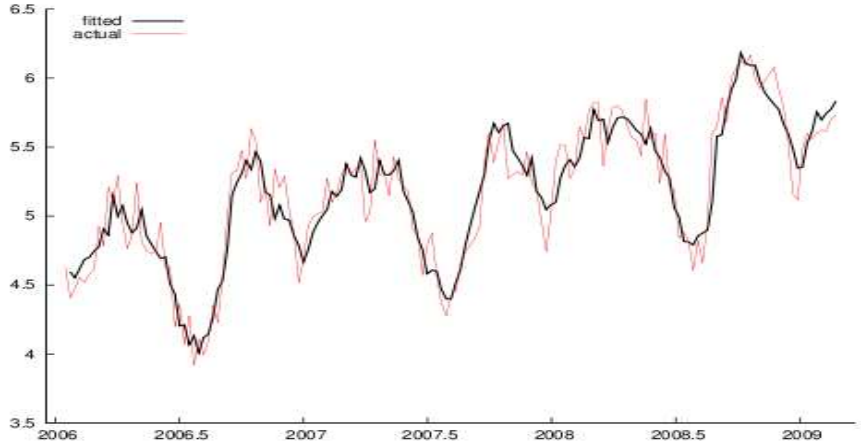
3 A model for weekly downloads

In this section, I will analyse downloads (in logs), after aggregating the data on a weekly basis, to get rid of the weekday effect. The explanatory variables are a time trend and a combination of sine-cosine terms with annual, semi-annual and quarterly period.[3] Additional regressors to account for short-term fluctuations are a dummy variable for the emergence of a new release on the previous week and one lag of the dependent variable. The model can therefore be represented as

$$(1 - \phi L)y_t = \beta_0 + \beta_1 t + \beta_2 r_{t-1} + \gamma' s_t + \varepsilon_t \tag{1}$$

where s_t is a vector of six trigonometric terms.

Fig. 3. Actual and fitted weekly downloads (logs)



OLS estimates of equation (1) are presented in table 1. As can be seen, the fit is excellent. The model predicts a rate of growth of about 43% per year[4] and the seasonal effect, as captured by the trigonometric terms, is highly significant.

[1] Downloads of the source package were assimilated to other linux packages.

[2] According to the Marketshare website (`http://marketshare.hitslink.com/`), the share of Mac users on the Net has risen from 6.09% in March 2007 to 9.61% in February 2009.

[3] Higher frequencies were also tried, but turned out to provide no significant contribution.

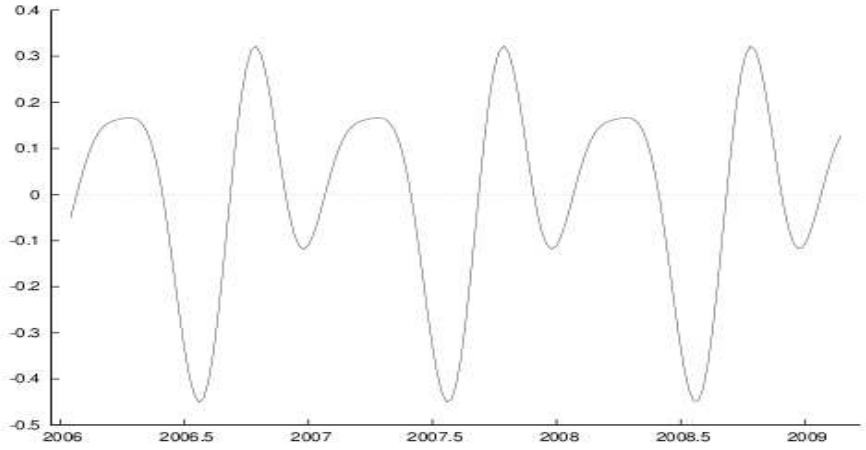
[4] This is computed as

$$\exp\left(\frac{\hat{\beta}_1}{1-\hat{\phi}}\right) - 1.$$

Moreover, the pure seasonal component (plotted in figure 4) shows clearly the summer and Christmas slowdowns I mentioned earlier. When a new version of gretl is released, downloads rise in the following week by about 12%.

Finally, the customary diagnostic tests show no sign of mis-specification, so we can conclude that the above model is adequate in summarising the data and that the two main stylised facts (the increase in downloads and its seasonality) are robust. The output of the Quandt LR test is especially important, since it indicates that the time pattern of gretl downloads has remained reasonably stable through our three-year sample.

Fig. 4. Seasonal component



4 Research or teaching?

The above model can be used for an indirect analysis of the composition of gretl users (or, more correctly, downloaders) between "researchers" and "teaching people" (which include students and teachers).

Clearly, this distinction is, to some extent, spurious: I, for one, belong to both categories. However, while gretl's aptness as a teaching tool is widely acknowledged (see for example Smith and Mixon [6] or Adkins [1]), there seems to be little recognition of gretl as a tool for applied research;[5] notable exceptions

[5] Under-reporting may be an issue here; research papers seldom cite the software used for their computations: for example, a recent paper of mine (Lucchetti and Palomba [4]) makes no mention of gretl whatsoever, despite the fact that gretl was almost exclusively used.

are Yalta and Yalta [7] (now slightly dated) and Rosenblad [5], who discusses gretl as a teaching tool but also highlights its potential for research use.

In my opinion, the composition of gretl's user base is of crucial importance for shaping gretl's future: people from the "teaching" community provide invaluable feedback and motivation to keep improving gretl's already excellent user interface. On the other hand, gretl's computing accuracy, scope and efficiency are put to the test (and pushed forward accordingly) primarily by those who use it for "real" applied statistical tasks. Moreover, if a true community of coders is to emerge, it is more likely that new coders come from the ranks of young, computer-literate researchers.[6] Therefore, it is important to ascertain the composition of gretl's user base if one wants to forecast, and possibly steer, the evolution of gretl as a computing platform.

Of course, a few assumptions are necessary here: I assume (a) that research activity is not seasonal and (b) that the share of "researchers" is a linear function of time. Needless to say, both hypotheses are a bit strong; we all have holidays in the summer and even atheists take a Christmas break. On the other hand, most of us do take advantage of periods when classes stop, to work on our research papers.

However, assumption (b) should be also only viewed as an approximation, meant to capture the overall trend, rather than a hard-edged fact and is in itself rather innocuous. As an indirect confirmation of assumption (a), I ran a simple seasonality extraction procedure on monthly download data for four popular scientific applications[7]: apart from gretl, I used gnuplot (the data visualistion tool that gretl also uses), maxima (a computer algebra system) and the GNU Octave software repository (a collection of additional modules for GNU octave, a free Matlab replacement). All these are mature projects, which are certainly used in teaching but also have a well-established reputation as research tools.

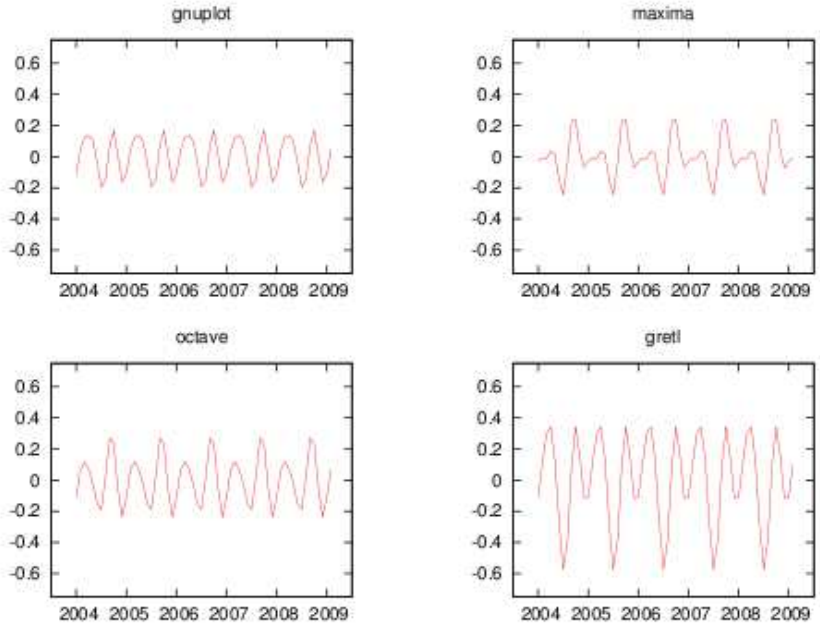
For each of these packages, the seasonality extraction routine is based on an OLS regression of the log of monthly downloads on a constant, one lag, a linear trend and the six trigonometric variables used above. The seasonality component is then reconstructed as the sum of each trigonometric variable weighted by its own estimated coefficient. Results are plotted in figure 5: it should be obvious that gretl's seasonal component is much larger than the other packages'.

In the light of assumptions (a) and (b), equation (1) can be generalised via a logistic smooth-transition regression model, similar in spirit to a STAR model

[6] It is true, however, that having been exposed to gretl as a student may encourage a researcher to study the source code and contribute original ideas and code.

[7] The source in all cases is, again, SourceForge. Unfortunately, I was unable to find download data for the R statistical project, which is not hosted by SourceForge and would have been extremely interesting to analyse.

Fig. 5. Seasonal component on monthly data for several free software projects



(see Granger and Teräsvirta [3]):

$$(1 - \phi L)y_t = \beta_0 + \beta_1 t + \beta_2 r_{t-1} + \frac{2}{1 + \exp(-\alpha t)}\gamma' s_t + \varepsilon_t \qquad (2)$$

Here, the α parameter measures the time variation of the importance of the seasonal component: if assumption (a) above is valid, then the basic idea is that α is a rough measure of how the weight of "teaching people" on the whole gretl ecosystem increases through time. Put another way, if α is greater than (less than) 0, then the share of people who download gretl for research decreases (increases). If α equals 0, the model reduces to (1).

Equation (2) was estimated by nonlinear least squares: the estimation results are shown in table 2. As is apparent, the coefficients are roughly the same as those in table 1. The estimate of α is negative, which suggests a reduction in time of the seasonal component, but is far from being significant. Hence, there is no compelling evidence of a reduction of the importance of the seasonality component in gretl downloads. If assumption (a) is valid, this means that in the period 2006-2008 the fraction of gretl downloads for teaching purposes has remained more or less stable.

Two considerations must be made at this point: first, the adoption of a statistical package as the tool of choice by applied economists and econometricians is

Table 2. Weekly nonlinear model

	Estimate	Std. Error	t-ratio	p-value
const	2.82425	0.33827	8.349	0.0000
y_1	0.38382	0.07364	5.212	0.0000
tim	0.11411	0.04787	2.384	0.0184
rel_1	-0.22265	0.03134	-7.104	0.0000
a1	0.01663	0.02139	0.778	0.4380
a2	0.14188	0.02995	4.738	0.0000
s1	-0.01456	0.02357	-0.618	0.5377
s2	-0.25353	0.04499	-5.635	0.0000
q1	0.09868	0.02715	3.634	0.0004
q2	0.03803	0.02071	1.836	0.0683
α	-0.04465	0.14856	-0.301	0.7642

Mean dependent var	5.168366	S.D. dependent var	0.499487
Sum squared resid	4.520373	S.E. of regression	0.173021
R^2	0.887462	Adjusted R^2	0.880009
Log-likelihood	60.03124	Akaike criterion	−98.06221
Schwarz criterion	−64.09892	Hannan–Quinn	−84.27276
$\hat{\rho}$	−0.005810	Durbin–Watson	2.001861

a long process: path-dependence and acquired habits may cause people to stick to obsolete tools for years, so, even if assumptions (a) and (b) are valid, it may just be the case that the sample I am using here is simply too short to capture this aspect adequately.

Moreover, the emergence of a community of gretl code contributors, well-versed in econometrics and programming at the same time, is unlikely to depend on the relative share of researchers on gretl's total user base, but rather on its absolute value. What counts is the number of "hackers" we have, not the percentage of users who are. In this sense, the stability of the share of "research people" on an increasing number of users allows us to be mildly optimistic.

5 Conclusions

Gretl has been so far a spectacular success story in terms of expansion of its user base. Obviously, the characteristic of free software that most people perceive as paramount (being "free as in beer") played its role, but this factor does not explain the whole story by itself: the Internet is full of gratis software which few people, if any, use. A large part of the merit goes to its intuitive and friendly interface, which makes it an ideal tool for teaching econometrics.

On the other hand, gretl's capabilities as a computing platform, which have also evolved dramatically, seem to have been overlooked by most practitioners, although lack of evidence may simply be a consequence of the limited time span of the data used here. Only time will tell gretl's future reputation as a solid and reliable computing nad there are reasons for moderate optimism. In any case, it is vitally important for the gretl community to work to advertise gretl's present capabilities to the widest possible audience and to work as hard as possible to extend and perfect them.

Bibliography

[1] ADKINS, L. (2009): *Using gretl for Principles of Econometrics, 3rd edition*. online.

[2] BALIETTI, S. (2008): "gig: a GARCH-like Implementation in Gretl," Master's thesis, Università Politecnica delle Marche.

[3] GRANGER, C. W., AND T. TERÄSVIRTA (1993): *Modelling Nonlinear Economic Relationships*. Oxford University Press.

[4] LUCCHETTI, R., AND G. PALOMBA (2009): "Nonlinear adjustment in US bond yields: An empirical model with conditional heteroskedasticity," *Economic Modelling*, forthcoming.

[5] ROSENBLAD, A. (2008): "gret1 1.7.3," *Journal of Statistical Software, Software Reviews*, 25(1), 1–14.

[6] SMITH, R. J., AND J. W. MIXON (2006): "Teaching undergraduate econometrics with GRETL," *Journal of Applied Econometrics*, 21(7), 1103–1107.

[7] YALTA, A. T., AND A. Y. YALTA (2007): "GRETL 1.6.0 and its numerical accuracy," *Journal of Applied Econometrics*, 22(4), 849–854.