Proceedings of the 4th Polish Symposium on Econo- and Sociophysics, Rzeszów, Poland, May 7–9, 2009

Dynamics of a Polish Internet-Based Social Network

K. Zmarzłowski[a], P. Mazur[a] and A.J. Orłowski[a,b]

[a]Katedra Informatyki SGGW, Nowoursynowska 166, 02-787 Warszawa, Poland

[b]Instytut Fizyki PAN, al. Lotników 32/46, 02-668 Warszawa, Poland

Dynamics of a number of new users registering for the first time to a Polish internet-base social network `Grono.net` is investigated via various regression models. Trends are estimated and the statistical significance of their forecasting is tested.

PACS numbers: 89.65.Gh, 89.65.Ef, 89.20.Hh, 02.50.−r

1. Introduction

`Grono.net` [1] is a web page (internet portal) hosting one of the largest (as for now it consists more than 2 millions of members) and well-integrated Polish internet-based communities. Service was launched in February 2004 and since its very inception till 2008 it was a close society — its membership was available "by invitation only". Within the society each member has a profile where any sort of information including photos and movies can be published and made available to other fellow members. The community can also be freely explored by registered members to find new friends and, possibly, add them to the personal lists of contacts. It is possible to join any of many subject forums or to create a new one. Plenty of information about various more or less interesting (not only cultural) events as well as a lot of advertisement can easily be found. More than 60 thousands posts per day are not unusual. Barely using the website is free of charge — only some extras are paid for. Business potential of such a forum can hardly be overestimated.

In this paper, using data coming from `Grono.net` as an illustrative example, we focus our attention on one essential aspect of any social network, namely the dynamics of its grow. Using a special Python script dedicated to logging in and collecting proper information we acquired a lot of precise daily data about new members registered on a given day to the `Grono.net`. We thoroughly investigate the dynamics of a number of new users registering for the first time to this network society. Here we restrict ourselves to practically complete time series spanning from January 1, 2006 to January 1, 2008. To quantitatively describe such a dynamics we employ different econometric models: from the classic least-squares-based linear regression to various ARIMA methodologies.

2. Econometric models

The linear regression model for the time series y_t is given by the following equation:

$$y_t = a_0 + a_1 t + \varepsilon_t \,,$$

where ε_t is a noise modeled by a stochastic process and parameters a_0 and a_1 have to be estimated. It is one of the most frequently used econometrical models of all times. Despite (or, perhaps, due to) its simplicity — estimation of trend parameters is based on the classic least-squares method — it is a very effective and successful model. Of course, the quality of the obtained model has to be thoroughly assessed, with various aspects being taken into account. In particular we have to assess: the overall significance of the model as measured by its determination coefficient R^2, statistical significance of model parameters (using Student's t-distribution statistics), and the properties of the model residuals (remainders), including their randomness (runs test, known also as Stevens or Wald–Wolfowitz test), normality (Shapiro–Wilk test), and autocorrelations (Durbin–Watson test), see e.g. [2].

The quality of the linear regression model strongly depends on the underlying assumptions. If the investigated time series does not satisfy them, we cannot use such a model for responsible inference. Therefore, anticipating some problem with the mentioned assumptions, we decided to apply also other, more sophisticated, statistical tools, namely autoregressive integrated moving average, i.e., ARIMA(p,d,q) models [3]. Nonnegative integers p, d, and q describe respectively the order of the autoregressive, integrated, and moving-average parts of the investigated model. To discriminate between various autoregressive models we assess their quality using Bayesian information criterion (BIC), also known as Schwarz criterion, based on the following statistics [4]:

$$\mathrm{BIC}(k) = n \ln\left(\frac{\mathrm{RSS}}{n-k}\right) + k \ln(n) \,,$$

where n is a number of observations (time moments), k denotes a number of model parameters (including the constant) to be estimated, and RSS is a sum of squares of model residuals.

To verify the quality of forecasting within this approach we use the mean absolute percentage error (MAPE) *ex post*, which can be computed according to the following expression [2]:

$$\mathrm{MAPE} = \frac{\sum_{T=n+1}^{n+h}\left|\frac{y_T - y_{TP}}{y_T}\right|}{h}100\%\,,$$

where y_T are empirical values of the time series under consideration, y_{TP} are (*ex post*) forecasted values for the same time moment, and h is the forecasting horizon.

3. Results

Even a rough inspection of Fig. 1 suggests that there should be an excellent linear fit to the observed data.

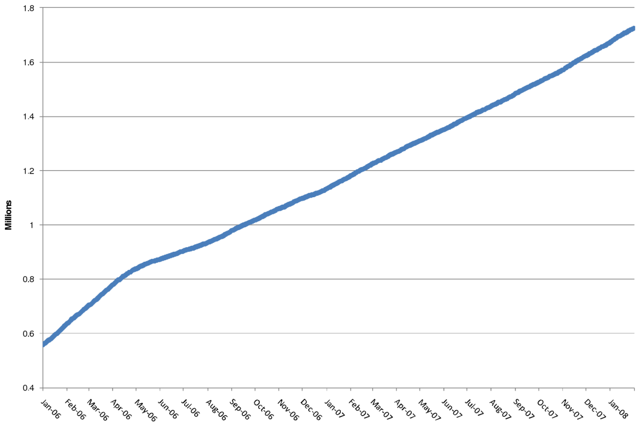

Fig. 1. Number of newly registered `Grono.net` members for the period of 01.01.2006–01.01.2008 as a function of time.

Direct application of the ordinary least-squares method to the linear model immediately gives us a trend that is almost perfectly fitted to the empirical data. For the investigated two-years period the obtained equation reads

$$\begin{aligned}&\text{number\_of\_members}\\ &\quad = 629784.6 + 1408.8\ \text{time\_moment}\,.\end{aligned}$$

Some of corresponding quality-measuring statistics are also impressive: overall significance of the model is very high as $R^2 = 0.99$ and both model parameters are statistically significant with very high probability (Student's $t = 463.8$, $p = 0.00$ for a_0 and $t = 437.8$, $p = 0.00$ for a_1). Unfortunately some properties of residuals, especially non-normality (Shapiro–Wilk test value 0.8457) and the presence of autocorrelations (Durbin–Watson test value 0.0006757), are too much different from those required to safely apply the linear model for time series forecasting. Despite these problems all our *ex post* predictions happened to be very accurate as can be seen from Table I.

Although the calculated forecasting terror MAPE = 0.427% is really very small, we have to be very careful because of the violated assumptions about model residuals.

Fortunately autoregressive models do not suffer from the above mentioned problems with additional assumptions. Applying Schwarz criterion we selected three the most promising ARIMA models as presented in Table II.

TABLE I

Ex post predictions for September–December of 2007.

Month	*Ex post* prediction	Actual number of members
September	1528645	1525397
October	1572320	1570172
November	1614586	1623086
December	1658261	1672260

TABLE II

Various ARIMA models ordered according to the Schwarz criterion.

ARIMA model	Informational Schwarz criterion
ARIMA(2,2,1)	9624.6
ARIMA(1,2,2)	9625.0
ARIMA(2,2,2)	9630.9
ARIMA(2,1,1)	9644.4
ARIMA(2,1,2)	9649.3
ARIMA(1,1,2)	9650.4
ARIMA(1,1,1)	9688.2

TABLE III

Predictions from various ARIMA models for September–December of 2007.

Month	ARIMA(2,2,1)	ARIMA(1,2,2)	ARIMA(2,2,2)	Members
September	1525531	1525530	1525522	1525397
October	1570289	1570292	1570307	1570172
November	1623179	1623179	1623170	1623086
December	1672246	1672250	1672278	1672260
MAPE	0.001%	0.006%	0.006%	

Results for *ex post* forecasting with ARIMA(2,2,1), ARIMA(1,2,2), and ARIMA(2,2,2) models are given in Table III. It can easily be seen from this table that forecasting errors are practically negligible (below 1%). Numerical values of the ARIMA-based predictions are very close to forecasts obtained previously from the linear regression model.

4. Brief summary

Although relatively new phenomenon, various internet-based social networks play an important and steadily increasing role in modern society. In this paper we have focused our attention on a dynamics of a number of members registered for the `Grono.net`, one of the largest Polish internet-based community. The standard linear regression model and well as some of ARIMA models have been successfully applied to trend estimation and forecasting for time series describing the number of new members. The most interesting period of 2006–2008 has been investigated. Both approaches provided very precise short-term (*ex post*) predictions producing similar numerical figures. Apparent linearity of the estimated trend for time series coming from presumably nonlinear dynamics deserves a deeper discussion. Such a discussion as well as more detailed analysis of other statistical aspects of `Grono.net` dynamics will be presented in a series of forthcoming papers.

References

[1] `http://grono.net` .

[2] G.S. Maddala, *Introduction to Econometrics*, 3rd ed., Wiley, New York 2002.

[3] T.C. Mills, *Time Series Techniques for Economists*, Cambridge University Press, Cambridge 1990.

[4] G.E. Schwarz, *Ann. Statist.* **6**, 461 (1978).