Modelling trends in the digital sphere – a comparison of two methods

Stéphane Gauvin, Université Laval (CANADA)
Jérôme Coutard, Compurangers (CANADA)
Isabelle Dornic, Compurangers (CANADA)[1]

Abstract:

Corporate governance involves the management of conflicting interests amongst relevant stakeholders. Proximate stakeholders, such as shareholders, customers, suppliers and personnel, have relatively clear means at their disposal to influence the decision process; distant stakeholders and society at large rely on voice (e.g. public opinion) or coercion (e.g. regulation/litigation) to influence corporate governance.

We explore the feasibility of formalizing the analysis of such measures by gathering data in the public digital sphere and modelling the formation of voice using two models of non-linear trend analysis. The ultimate goal is to be able to make early predictions on the prevalence and timing of a dominant voice. In this paper we will introduce data exemplars generated by SWAMMER and preliminary modelling results obtained through these models.

To be published in the proceedings of the *Ecommerce and Internet Governance 2007* conference

June 2007

[1] All correspondence should be addressed to stephane.gauvin@fsa.ulaval.ca

1 INTRODUCTION

The Internet is having a profound impact on business. What has started fifteen years ago as a race for pre-eminence in cyberspace for select few and the rapid transformation of business processes for most, is taking a new twist with the advent of the so-called web 2.0. Business organizations are currently busy trying to find ways to profit from the work of others (to wit, the enormous success of sites like youTube where key assets are so-called User Generated Content).

Today the digital sphere is morphing into the most important governance tool ever available. The Internet has become the übermedia. What has been the province of mass media, reluctantly but inexorably drawn to post their content over the Internet in order to avoid becoming irrelevant, has become a sphere of generalized exchange of opinions by an ever growing group of ordinary people who feel compelled to be heard, would it be by just a few.

Recent statistics suggest that over 70 million blogs have been created, with roughly 14 million being considered active by Technorati. In the political sphere, the Internet has overtaken traditional media as the primary area of concern precisely because blogging has changed the rules of the game – it is believed that blogs have become the dominant source of news material for media organizations (Drezner and Farrell 2004, Howard 2005).

What this revolution implies is that the digital sphere may have become an accurate reflection of public opinion, as opposed to a reflection of what educated observers think about public opinion. We are therefore witnessing several important developments in the area of information gathering methodologies targeting the digital sphere (Baumgartner et al. 2005, Chau and Chan 2005, Marjanovic 2007, Maynard et al. 2005, Soper 2005, Surman and Reilly 2003).

One such example is Swammer, a company which relies on the content of the digital sphere to report on the visibility of diverse subjects. Their work has started to raise eyebrows when it became apparent that monitoring the digital sphere yielded electoral predictions that were just as accurate as those of the best survey organizations, at a fraction of the cost. If scanning for the prevalence of concepts in the digital sphere can replace election surveys, it is certainly possible to envision alternative applications that will profoundly impact corporate governance.

Consider the case of a company in the food business. As recent developments have shown, there is an increasing awareness of the exposure to litigation for corporate decisions that do not look, prima facie, unreasonable. Consider lawsuits by people who argue that they became obese due to the negligence of fast food operators (McCann and Haltom 2004). What is not clear at this point is whether lawsuits will be dismissed as frivolous or whether courts of law will find for the plaintiffs and award significant damages. For the moment public opinion considers that the lawsuits are frivolous and the lawyers are largely depicted as trying for another money grab. But as McCann and Haltom (2004) noted, this is not unlike what happened when the first cases were brought to justice against tobacco companies. This is why organizations involved in this sector must keep a close eye on the interaction between public opinion, tort lawyers and court decisions in order to ascertain their exposure and the likelihood that the regulatory environment will change (see Strnad 2005 for an exhaustive argument on “fat taxes”).

Whereas organizations dutifully track their customers' satisfaction in an obvious effort to manage their revenue stream, tracking the concerns of distant stakeholders is rarely done because the cost would be prohibitive given the low likelihood that a given concern will have material impact. Yet as society becomes more complex, distant stakeholders are learning to make their points and gain compensation though a variety of means. We are all familiar with the case of the tobacco industry whose fate has taken a turn for the worse, globally, when the public at large became sympathetic to the idea of the hazard of second hand smoke. Individual rights of smokers were deemed less important than those of the non smokers, as soon as smokers were a minority.

Searching the digital sphere quite in the same way as when we search public opinion would open a whole new perspective on responsible governance as organizations would no longer be reacting to overt attempts to impact their processes, as they could see threats and opportunities over a much broader horizon than what they can usually manage.

Our purpose in this paper is to explore time series extracted from the digital sphere. We wish to address two specific questions. First, we ask whether such time series have a good signal-to-noise ratio. In other words, we ask whether one can make reasonably accurate predictions based on digital sphere data, or is instead limited to recover the obvious that would not need sophisticated search to begin with.

Second, we ask what types of models provide good tools to investigate such data. On the one hand there is the convenience of well known parametric models like a simple regression, on the other, such series may not be amenable to simple techniques and could require more elaborate tools. Since it is unlikely that such series will turn out to be normally distributed around a straight line, we want to explore the merits of robust regressions. We also wish to introduce spectral analysis to the field of management. Spectral analysis is becoming the leading methodology in several fields of science where some of the best minds try to uncover cycling processes and trends.

The next section introduces the data. Time series modeling is an inductive discipline that greatly benefits from close encounters with the raw data. Then we introduce our models. We first present two parametric models, one linear and the other S-shaped, and briefly explain the concept of *robust estimation*, a procedure that is not frequently used in marketing or management, but is nonetheless invaluable in cases like ours. We also introduce the essential ideas that stand behind singular spectral analysis. Third, we present our results. We keep technical details to an absolute minimum and rely heavily on graphs to allow a good understanding our key findings. We conclude on a hopeful note -- our findings suggest that digital sphere data can be modeled quite accurately -- directions that we intend to take to continue our work and a few caveats.

2 Data

The digital sphere is invisible. Even though we know that it is composed of billions of pages of information, we cannot see the general picture. Instead, we rely on search queries to explore its content, as if we were blind men asking questions in an attempt to see. The nature of these questions and our abilities to integrate answers into a coherent whole are critical to the process of understanding the digital sphere.

2.1 Overview

Swammer is a recent upstart specialized in the automated generation of meaningful search queries. In a typical application[2], a team of specialists will study their client's needs and elaborate a set of queries with the purpose of extracting relevant information. But unlike many search specialists, Swammer has no interest in the content that is returned by the queries – the information that is sought is the prevalence of a given concept in the relevant space of the digital sphere. The key index that is generated is a visibility score.

What may seem trivial at first is more complicated than it seems, in actual practice. For instance, if we were to try to build a visibility index for, say, "John Smith," we would have to generate queries for various forms by which Internet pages refer to that individual. A query on "Smith" will return hits referring to targets that are largely irrelevant. A query on "John H. Smith" or "John Smith" will return hits that are more appropriate, but also largely contaminated by hits related to the professional hockey player, to the accountant, to the entrepreneur, i.e. to individuals who share a name yet are not the one we wish to track. If we add "marketing" we can prune a large proportion of irrelevant hits. We can also use another query that will add "ABC University." We can examine what happens if we launch queries on "Smith AND marketing" and add whatever NOT clauses are required to zero in on our target.

In some cases, queries can be further refined in order to distinguish between positive and negative content. Swammer thus generates indexes of positive, negative and neutral references to a target.

Eventually, a set of queries is designed that satisfactorily captures our target. It is then possible to run these queries every day and compute an index that will weigh hits in proportion of their visibility on the Internet.

In this paper we use two datasets graciously provided by Swammer. The first dataset pertains to the French presidential race of 2007. We will report on the two leading candidates, Mrs. Ségolène Royal and Mr. Nicolas Sarkozy.

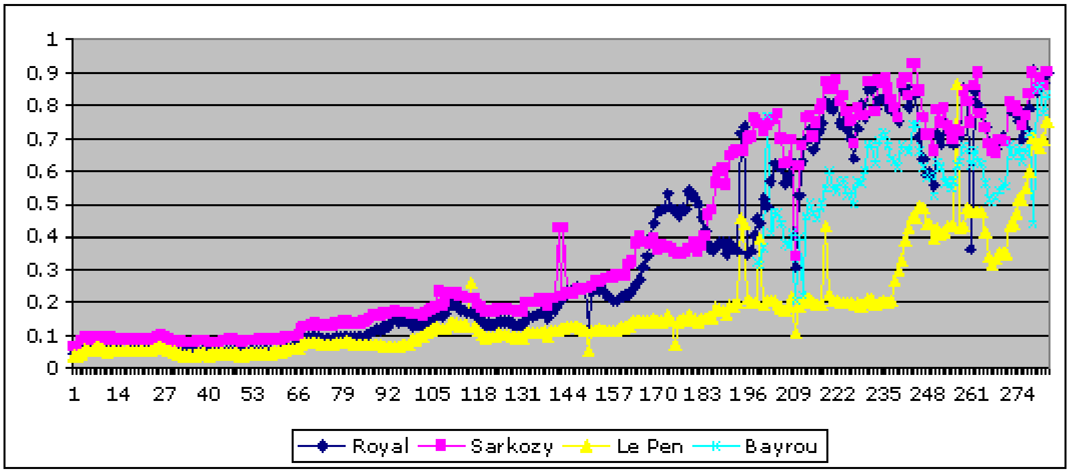
The second dataset was generated at the request of an industry association in order to track several items. Indexes were designed to capture the visibility of individual organizations (publicly traded companies) in the context of several issues (ex: environment, corporate performance, etc.). Because of the sensitive nature of the data, we cannot reveal the type of industry or the exact nature of the measurements – however, the data will illustrate the nature of the estimation problem.

[2] The example is for illustrative purpose only. The actual process is a trade secret.

2.2 *French presidential data*

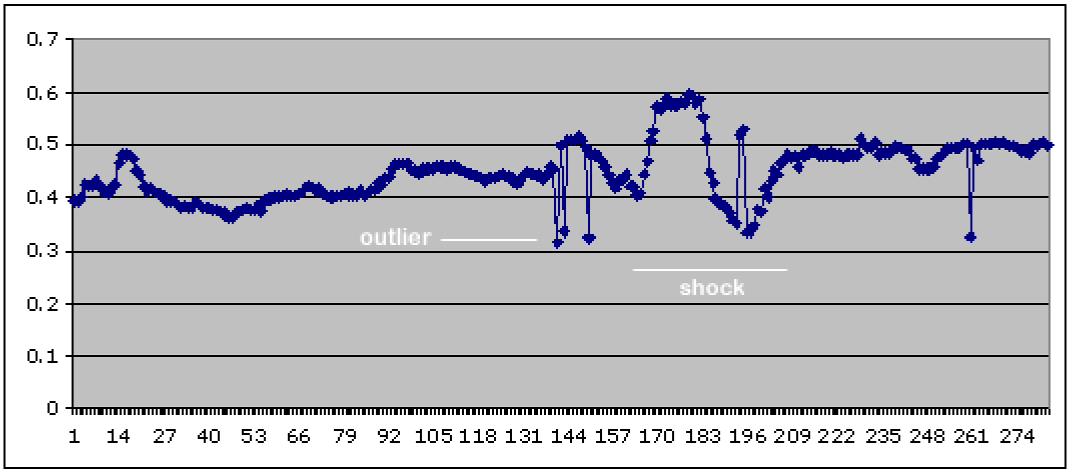
The dataset reports visibility indexes for the top four candidates, starting on July 3rd 2006 until May the 1st 2007.

Figure 1: Visibility scores for the 4 leading candidates to the 2007 presidential French election



The data clearly show two dominant features. First, there is an obvious trend upwards as, presumably, official media, bloggers and political parties became more and more focused on the topic of the presidential elections. We can remove this underlying trend by computing visibility shares. We have focused on the two leading candidates who made it to the elections' second turn, and report the visibility share of Mrs. Royal (figure 2).

Figure 2: Visibility share of Mrs. Royal (Royal + Sarkozy = 100%)

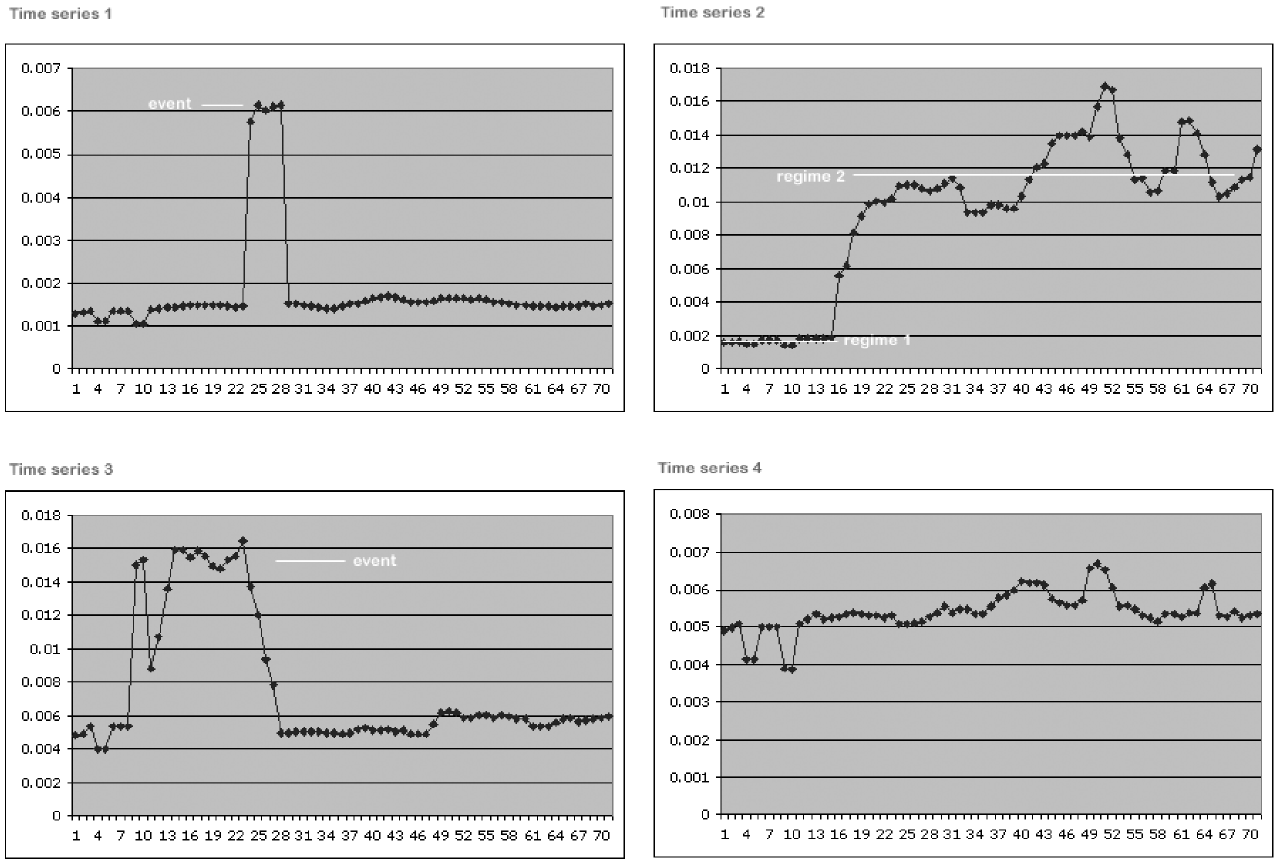


Several isolated, single period, extreme values are readily apparent (days 142, 143, 150 and 261 in particular), and a shock extending over a longer horizon occurs in the vicinity of day 180. The time series climbs quickly from a low point of 0.4 on day 158 to a high point of 0.6 on day 179 before falling to a (normal) low of .35 on day 196 and returning to the apparent trend line around day 206. Yet, overall, the series is fairly well behaved and we should be able to extract trend data without too much difficulty.

2.3 *Industry data*

The dataset tracked 7 distinct concepts such as environment, management, R&D or performance, for 9 large publicly traded corporations and for the industry as a whole. Indexes of positive and negative value were computed for each enterprise / topic combination, for a total of 210 distinct series. The project has been relatively brief such that we have only 71 data points per series. In this paper we will limit our presentation to four randomly selected series, presented in figure 3.

Figure 3 Industry data


Time series 1

Time series 2

Time series 3

Time series 4

This first series shows a major event starting at period 23 and lasting for 5 periods. Such event could be a news item reported by mainstream media or an entry on a high traffic site such as Digg. The series is otherwise rather well behaved.

The second series shows a different situation. Here it seems that we are witness to a change in regime, with a low visibility phase extending up to period 15, followed by a transition period during which the series climbed to a new and much higher visibility regime. Peaks and valleys occur around periods 50 – 63 and may or may not signal a third, oscillatory regime.

The third series is somewhat similar to the first, with a highly visible event. In this case however the climb shows a marked drop after two periods of high visibility. The series returns to the high-visibility plateau for 10 more periods after which it falls back to the base-level regime over the course of five evenly spaced visibility levels.

The fourth series has no remarkable feature. Like the other series that we have briefly introduced, it shows what appears to be a fairly robust signal and in this case the signal is not perturbed by a visible event or a change in regime.

3 MODELS

The vast majority of the literature in marketing and management sees trend more as an annoyance rather than as a signal worthy of analysis. As Dekimpe and Hanssens (2000) indicated in their thorough review of this question, researchers in those fields are generally concerned with theory building and will therefore rely on various techniques such as Vector Auto Regressive (VAR) models to exclude trend data and isolate the apparent causality of one variable on another. (see Dekimpe et al. 2006 for an update. See Franses 2005, Fok et al 2005, Hyung and Franses 2005, Nielsen 2003, Pauwels et al. 2004 for illustrative examples). The fact is that if one is interested in structural models, underlying trends may artificially inflate the impact coefficients.

Our goal is more modest, perhaps, in that we simply wish to be able to capture trends in a univariate time series in order to make statements about its eventual plateau. To this end, we can rely on two families of models: parametric models where the functional shape of the trend is pre-specified, and non-parametric models where the model may yield an arbitrary trend line.

3.1 Parametric models

Two questions must be considered. First, we must choose a model specification. The simplest trend model is the ubiquitous linear regression of the form $Y = \alpha + \beta T + e$ where T will be a time index. While such parsimonious model will tell if a series is moving up or down over time, it suffers from two important drawbacks. First, since it does not put bounds on the value of Y, it will generate implausible predictions as one moves along the time horizon. Second, our cursory presentation of the data strongly suggests that a linear model will fail to appropriately capture regime changes such as the one visible in the second industry time series.

An s-shaped response model should be a better alternative. We will therefore fit the data using:

$$Y(T) = \alpha + \beta\ (T^{\delta}/(T^{\delta} + \gamma)) + e$$

where:

T is the time index,
Y is the value of the visibility score at a point in time
α is the intercept
β is the difference in regime (i.e. $\alpha + \beta$ yields the new plateau)
δ and γ are shape parameters, >= 0

This model is probably familiar to many as it uses the same functional form as ADBUDG (Little 1970). This model can accommodate a wide range of phenomena, including concave, quasi-linear and up- or downward sloping s-shaped responses depending on the relative magnitudes of the parameters.

The second question pertains to the appropriate estimator. As we have seen in the previous section, our time series are often perturbed by extreme values, shocks or other major events. Two alternatives are conceivable. We can either introduce additional variables to explicitly, a priori, model shifts in the series and then use an OLS estimator, or we can use a robust estimator that will weigh observations in inverse proportion of the magnitude of the error term (see Frances et al 1996 or Fried 2004 for examples, Western 1995 and Huber 1981 for a description of the methods). Both approaches have their limitations. Explicit modeling requires extensive work from the analyst and some theoretical foundations in order to distinguish between extreme values that are explainable and can be explicitly modeled and those that should fall in the "error" category. The robust approach is, as its name aptly says, very robust to extreme values. It makes implicit assessments of each observation's contribution to the trend estimation, but yields optimistic fit statistics. Robust models are often associated with data mining, a practice frowned upon by theory building academics (see Franses 2005 p. 10 for instance).

Robust estimators can be written:

$$Y = \phi\, XB + e$$

where ϕ is a vector of weights applied to the observations. Each individual weight is iteratively computed based on the probability that a given observation belongs to the trend line. At each iteration, mean error and variance are computed. Extreme values, with extremely low probability of being on the trend line, are given negligible weights in the estimation of the parameter. Calculations iterate until we have satisfied convergence criteria of the maximum likelihood estimation. This robust regression is called M-regression.

In this paper, we will estimate both a simple linear model as well as the non-linear model introduced above, with an M-regression procedure.

3.2 *Non parametric models: SSA*

Several non parametric models are being used to investigate digital sphere time series (Bagchi and Mukhopadhyay 2006, Papagiannaki et al. 2005) Yet one of the most promising technique is still virtually unknown in our field even though it has become dominant in several areas of science: singular spectrum analysis. Singular Spectrum Analysis (SSA) is, according to Golyandina et al. (2000), a technique of multivariate geometry rather than a statistical methodology. It is also a somewhat bewildering array of concepts: "[i]n addition to statistics and multivariate geometry, the theory of SSA comprises the elements of signal processing, linear algebra, nonlinear dynamical systems, the theory of ordinary differential and finite-difference equations, and functional analysis (p. 9)"

In recent years, SSA models have become increasingly popular, especially in the fields of earth sciences (ex: Ghil et al 2002) where the goal was to capture the signal of climactic time series, biology (Balazs and Chaloupka 2004) and epidemiology (Koelle and Pascual 2004). The relatively similar wavelet approach has been used to model Internet related phenomena (ex: Antoniou et al. 2003, Edwards 2006).

In essence, SSA decomposes a time series in a set of frequencies that are grouped into three broad classes: trend signal (low frequencies), cycling/seasonal signal (well structured frequencies that coincide with a theoretically meaningful periods) and noise (arbitrary frequencies). Unlike other techniques, SSA can provide useful insights on short series.

The SSA procedure proceeds in four steps. First, we create a matrix of lagged vectors (X). For a time series of length L, the maximum number of underlying functions that can be identified is L/2. This length is called the window and while it could be less than L/2, it should be long enough to capture cyclical phenomena. Each vector of the X matrix lags the time series by one period. The second step involves the extraction of eigen vectors of the matrix XX^T. The third step involves the grouping of the eigen values according to a variety of criteria depending on the objective of the analysis. In our case, eigen values exhibiting slow varying changes in the autocorrelation with the original series will be kept. If our purpose were to identify cycling/seasonal patterns, we would have searched for systematic/short frequency patterns (we will present an illustration in the next section).

It is possible to automate the process using simple extraction and reconstruction rules. An automated trend extraction software, autoSSA, has been developed by Theodore Alexandrov (2006), a former doctoral student of Golyandina. This is the procedure we will use initially to extract the trend for our time series. For the moment, recall only that the trend, as it is defined by the SSA methodology, refers to a slow varying change in the underlying series, which can take any shape whatsoever.

4 RESULTS

In this section we present a comparison of the results obtained by four models. First, as a benchmark, we fit a standard OLS linear trend model. Second, we fit the same model using a robust estimator. Third, we fit a non-linear robust model and finally we compare with the trend component extracted with the SSA trend extractor. For each series we report the parameter estimates and R^2 or pseudo R^2 values and related statistics. We provide results in tabular and graphical form, first for the presidential series and then for the four industry series.

Figure 4 – Results: presidential time series[3]

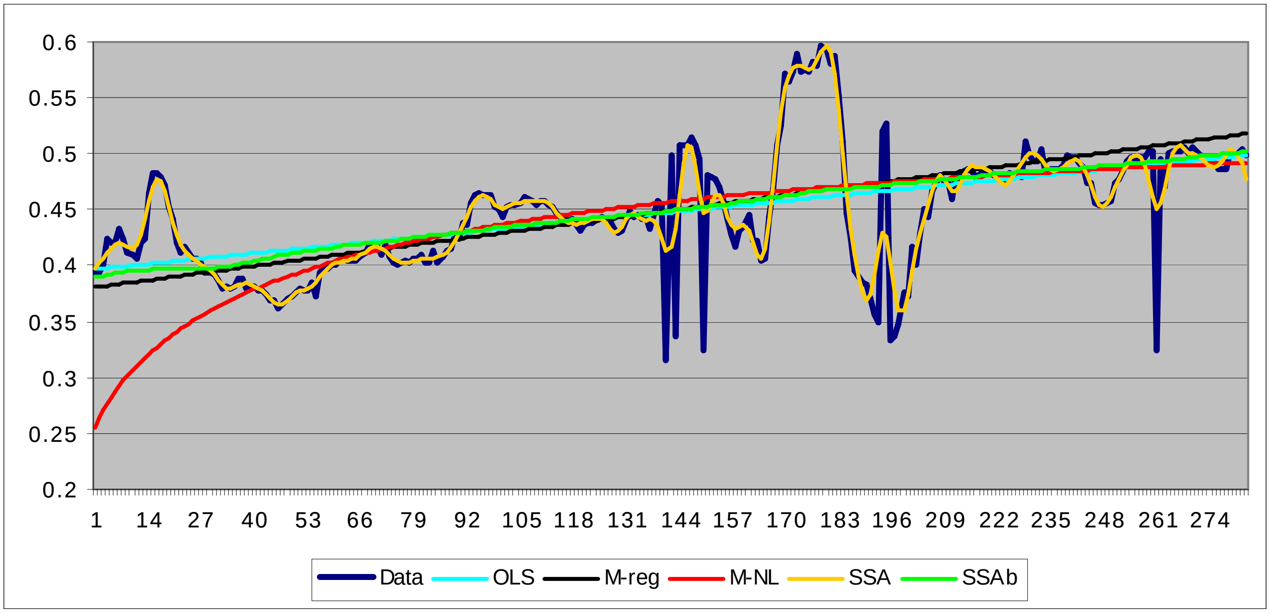


Table 1 – Results: presidential time series

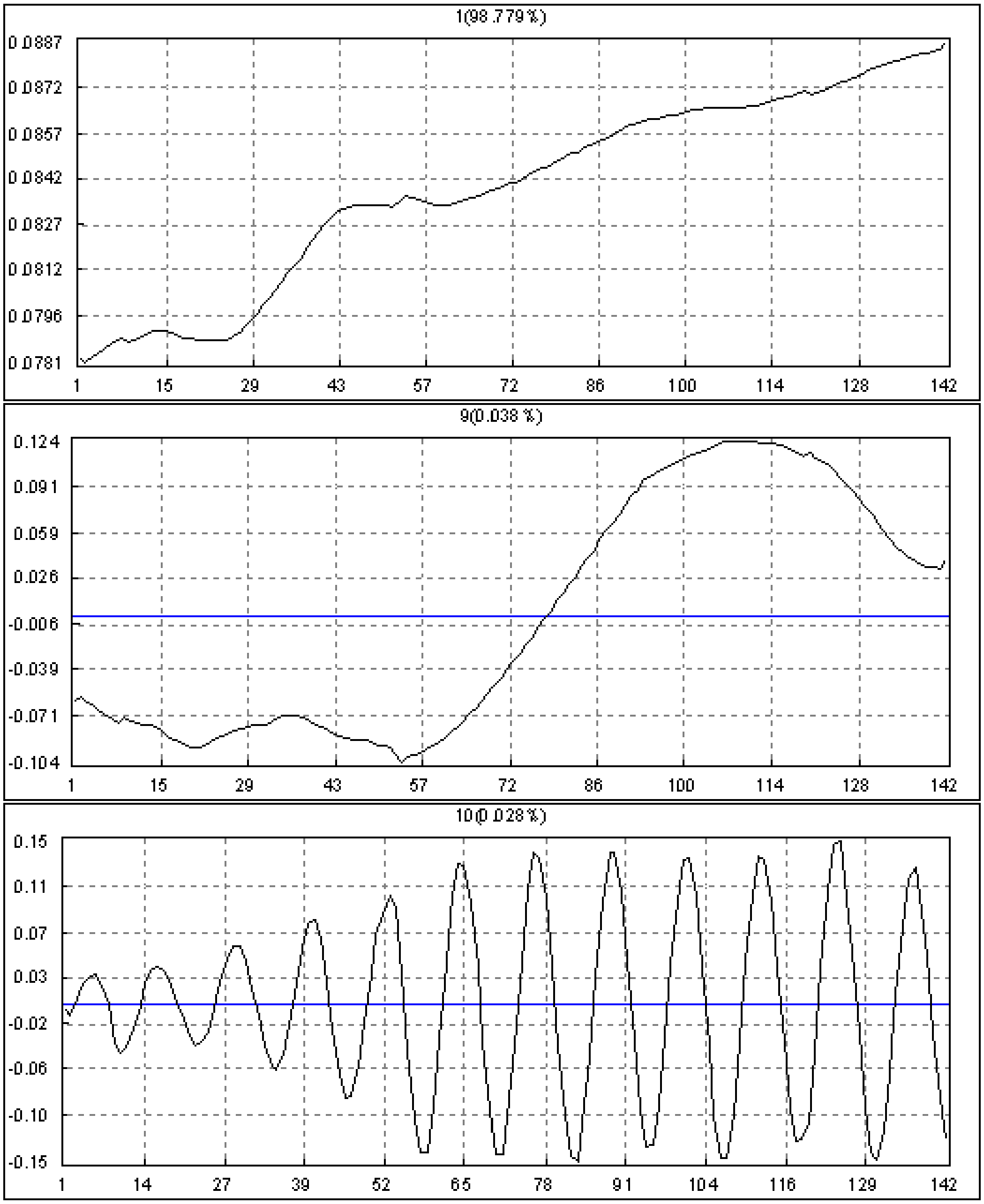
	Estimator				
Parameters	OLS	M-Reg	M-NL	SSA	SSAb
alpha	0.397	0.380	0.243		
beta	0.00036	0.00049	0.322		
delta			0.791		
gamma			25.260		
R2	0.304	0.608	0.646	0.867	0.397
F	122.650	436.543	513.368	1662.659	186.589
Prob	0.000	0.000	0.000	0.000	0.000

Results from most methods are remarkably close. With the exception of the SSA trend line, to which we will refer later, they all point to a significant upward trend in the visibility share of Mrs. Royal. The robust M-reg estimator suggests that the OLS slope parameter is slightly biased towards zero by the outliers, as we should have expected by mere visual inspection of the data. More interestingly, perhaps, the robust non-linear estimator suggests that the upward trend is closing on to its new plateau of .565, such that the last predicted value is the lowest of all parametric estimates. Even though differences are not material and keeping in mind the fact that visibility in the digital sphere and voting behavior are not the same, it may be appropriate to remember that Mrs. Royal ended-up with a .472 share of the popular vote.

[3] The data lines are in color, meant to be read on screen or on a color printed document. An annotated grey-scale version can be found in the appendix.

SSA trends have been extracted automatically (SSA) and manually overridden (SSAb) using the longest possible (half-length) window as is suggested for relatively short time series. The automated procedure kept nine underlying series (1, 2, 3, 4, 5, 6, 7, 9 and 13) in a reconstruction process based on the low frequency criterion. Because a closer inspection showed that the first series captured the overwhelming fraction of the variance (99%) we also report a distinct SSA reconstruction based on the first eigen values.

Figure 5: Correlations across some eigen functions

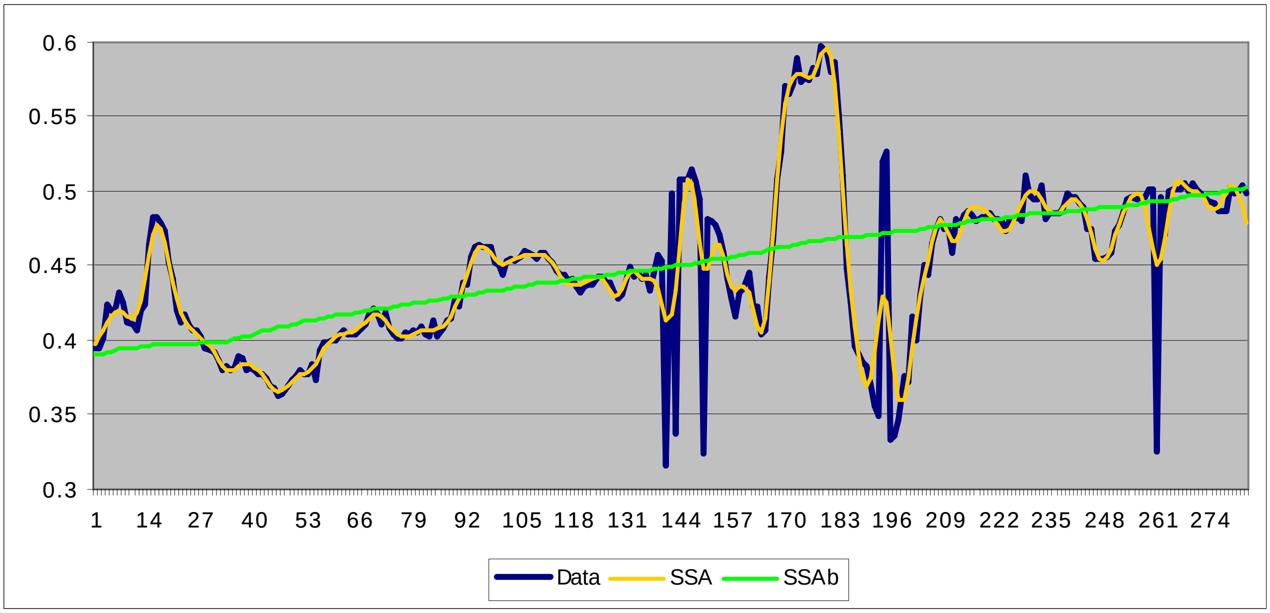


To illustrate what the procedure does, consider figure 5. We see that the first series accounts for 98.8% of the variance captured with the computation window. The series exhibits a slow varying correlation with the original series, suggesting that it captures trend data rather than cycling or seasonal patterns. We see a similar pattern with series 9 with the autocorrelation function varying slowly over time, although in this case, the eigen vector captures a mere 0.38% of the variance. In sharp contrast, function 10 shows a definite oscillating pattern that appears to capture weekly variations in the initial time series; many functions, not shown here because of space limitation, show definite daily patterns while others show unstructured patterns.

This is one of the paradoxes of SSA trend extraction – quite like qualitative methods, SSA yields rich results in comparison to the reductionism of parametric models, so rich in fact that they defy simple interpretation. Consider figure 6 (similar to five but without OLS and robust trend estimates) – the SSA reconstructed series tracks the actual series very closely. It allows the identification of severe outliers (around periods 144, 196 and 216) yet threads a large number of up- and downswings that occurred throughout the electoral campaign. A tantalizing result is visible near the end of the series, where the reconstructed series drops sharply. Whereas the raw data indicated that Mrs. Royal visibility share was at .498, the SSA trend estimate put the share at .478, much closer to the electoral outcome.

This being said, including all low frequency underlying series in the reconstruction phase obscures the fact that except for the first, extremely robust, underlying series, additional signal components are weak and may well be misleading. Another way to put it – while it is true that severe outliers are still readily identifiable, it is nonetheless obvious that the spectral analysis has been influenced by these points, as each outlier is accompanied by a corresponding turn in the trend line. In other words, SSA will integrate strong signal in its low frequency category, even if the signal is of very short duration.

Figure 6: Results – presidential series, SSA and SSAb trend estimates



The limitations of automated SSA trend extraction will become more readily apparent when we turn our attention to the industry time series, which presented themselves with several types of large scale events. Figure 7 below presents the results in a graphical way. Table 2 shows essential statistics.

If the results for the presidential time series were fairly consistent across methods, the story is quite different with the industry data. Industry series were affected by severe outliers, major events of short duration or major regime changes all of which are considerable challenges for estimation procedures.

A glance at table 2 shows that the robust estimator for the non linear model fared the best of all, which is to be expected since it has both more flexibility than the robust linear model and can handle outliers better than the OLS model and differences can be considerable under the "right" circumstances. Consider time series 2 for instance, where there is an apparent regime change. The OLS estimator slices through plateaus, yielding what appears to be an inappropriate verdict on the series (i.e. predicted values will be overly optimistic for later periods as it considers the higher plateau to be an indication of a strong positive, and ongoing, trend. The robust linear estimator discounts the first, lower, plateau which yields the unusual result of a lower R-square compared to the OLS estimator, and while the robust slope is less than that the biased OLS estimate, it is still generating overly optimistic predictions for later periods. The robust non linear estimator in contrast captures remarkably well the regime shift, even though it appears to be tracking on the low side of the upper regime.

Figure 7 – Results: industry time series

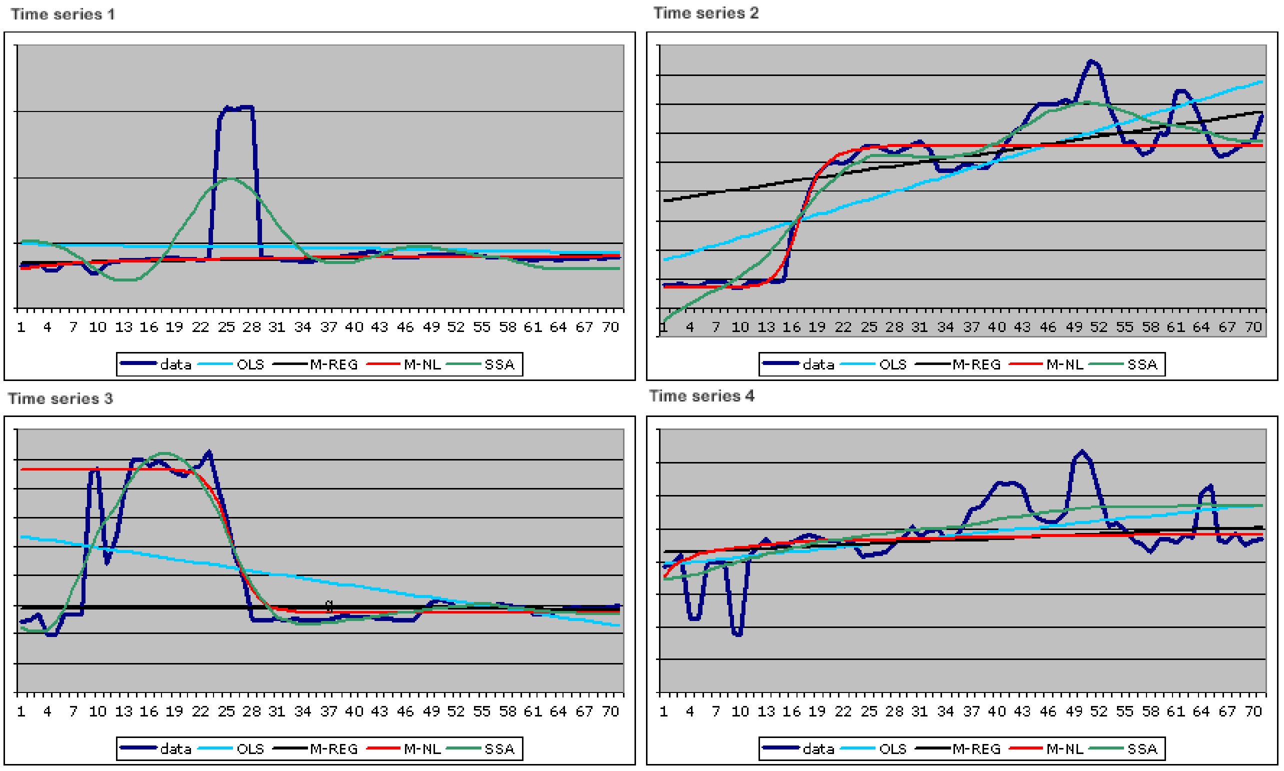


Table 2 – Results: industry time series

	Time series 1				Time series 2			
Parameters	OLS	M-reg	M-NL	SSA	OLS	M-reg	M-NL	SSA
alpha	1.95E-03	1.35E-03	3.50E-04		0.003	0.007	0.001	
beta	-3.92E-06	3.55E-06	3.55E-02		0.00018	0.00009	0.010	
delta			0.096				6.730	
gamma			4.16E-05				1.68E+02	
R2	0.005	0.900	0.919	0.487	0.648	0.510	0.847	0.731
F	0.325	618.564	780.214	67.384	126.999	71.674	382.036	263.665
Prob	5.71E-01	3.60E-36	2.44E-39	7.54E-12	2.69E-17	2.81E-12	7.69E-30	1.95E-25

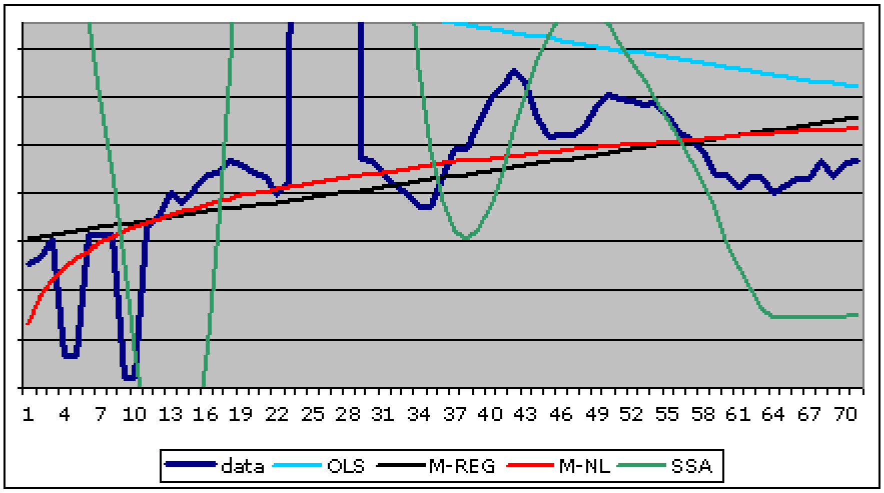
	Time series 3				Time series 4			
Parameters	OLS	M-reg	M-NL		OLS	M-reg	M-NL	SSA
alpha	0.011	0.006	0.015		4.93E-03	5.12E-03	3.76E-03	
beta	-8.70E-05	-1.71E-06	-9.86E-03		1.30E-05	5.65E-06	1.77E-03	
delta			18.466				0.557	
gamma			1.00E+20				7.78E-07	
R2	0.199	0.605	0.834	0.907	0.268	0.255	0.315	0.397
F	17.134	105.699	345.905	690.372	25.316	23.669	31.705	46.812
Prob	9.69E-05	1.47E-15	1.38E-28	5.29E-37	3.70E-06	6.95E-06	3.58E-07	2.42E-09

The robust non linear estimator is also the clear winner of the three parametric models for time series 3, although the story has an interesting twist. The results reported here show that the OLS estimator has been pushed into a negative slope by the event occurring near the beginning of the series. This appears to be inappropriate as the underlying series otherwise tracks upwards. The robust non-linear estimator also reports a negative trend, albeit much smaller. The robust non-linear estimator shows a regime change, starting at a high plateau and plunging to a lower one and then tracking on a flat line. The non-linear model yields a better fit overall and predicted values for later periods are a much closer to the actual data.

Interestingly, in the course of analyzing these data, we found out that the robust non-linear estimator was sensitive to the initial values fed into the maximum likelihood algorithm. If starting values for alpha and beta were given on the basis of the prior estimation of the robust linear parameters, the iterative process stopped to progress almost immediately and the algorithm yielded a function practically indistinguishable from the robust linear (i.e. a negligible negative slope). Clearly, the loss function is not a well behaved convex hull, but instead appears to have two distinct minima – one close to the linear estimates and the one we reported. What may appear to be annoying at first is in fact a useful reminder that at some point, “events” such as those occurring in time series 3 are in a grey zone between obvious outliers on the one hand, and bona fide regimes on the other. Consider a series similar to time series 3, but starting 10 periods later. Intuitively then, one would infer that there has been a downward regime shift. Contrast this with another series where the event occurring between periods 10 to 28 is ten periods shorter. Intuitively we would consider that the event is an aberration similar to what we saw in time series 1.

Results for time series 1 show how robust estimators are important. The OLS slope estimate is negative, because the trend line is pulled upwards by the event occurring between periods 23 and 28. In contrast, the robust linear estimator cleans the series and yields a more appealing positive slope estimate. Similarly, the robust non-linear estimator yields a small, positive, "regime" shift. A closer look at the results may be useful, because the magnitude of the event is such that we misconstrue what happens at the series baseline. Figure 8 shows that the series baseline is trending upwards. The OLS estimator (pale blue line) misses completely the baseline. In contrast, both robust estimators track the baseline correctly, with the non-linear model (red) providing a better adjustment. The SSA trend line, which at first might have appeared to provide an overall better summary of the series because it reacts to the event, turns out to gyrate excessively around the baseline.

Figure 8 – A close-up view of industry time series 1



Finally, all estimators provide a similar picture for uneventful time series 4, i.e. a weak uptrend. Robust estimators do not provide dramatic improvements over the OLS benchmark, although both provide a better adjustment over later periods.

Our review of these results would be incomplete if we did not draw attention to the SSA trend estimates. All estimates were obtained by a fully automated procedure. These trend lines track series fairly well and when these lines diverge from the robust non-linear estimates in a meaningful way (such as time series 4), that remains certainly a plausible interpretation of the data. This is perhaps most obvious in the case of the third series, where the robust non-linear model would have us think that there was a regime shift, while SSA tracks as if it were an isolated event.

Pseudo R^2 values for the SSA trend lines are usually close to those of the robust non-linear estimator except for the first series where SSA was unable to fully capture the event occurring near the beginning of the series.

A more detailed description of the SSA analysis is beyond the scope of this paper as it would entail producing vast amounts of charts and data. An illustration of the procedure can be found in Ghil et al. (2000).

5 CONCLUSION

The purpose of our exercise was to explore the feasibility of modelling digital sphere time series in order to track public opinion. We had two fundamental concerns. First, how much noise vs signal is present in the series. This question has an unambiguous answer: our data clearly shows that such series have a strong signal component. All models reported healthy fit statistics and the SSA procedure yielded first eigen values that were both large and of low frequency (i.e. belonging to the trend component).

Second, we wanted to report on two types of analysis procedures that appeared to be reasonable candidates to capture the underlying trend – robust parametric models (both linear and non-linear) and the very appealing non-parametric SSA model. Here the answer is more nuanced. On the one hand, our analysis shows that robust estimation is essential. Casual inspection of the data showed that series exhibit extreme values in the form of spurious outliers as well as what we called "events", i.e. a short run of outliers. We also showed that a robust estimator for a non-linear model appears to be the safest approach since several series exhibit what appears to be regime shifts.

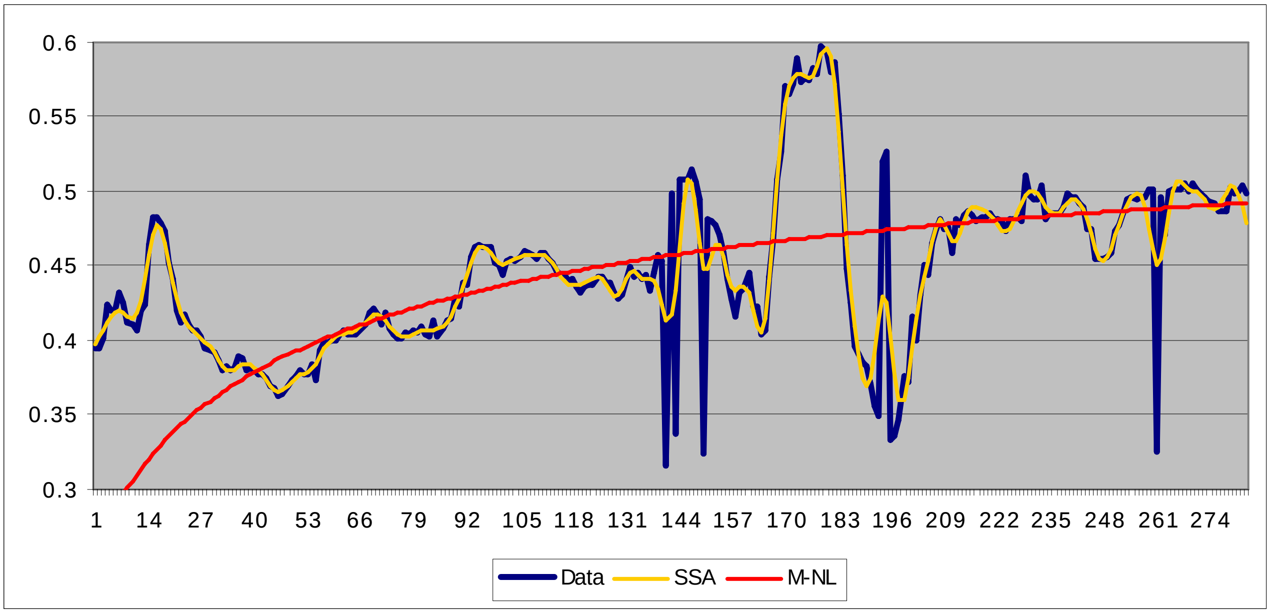
On the other hand, the SSA procedure delivers richer trend extractions that would probably make a useful complement to the parsimonious parametric models that we used. SSA is fairly robust to outliers as it treats them as a part of the signal that is carried by lower rank underlying series. However, SSA trend estimates appear to be vulnerable to events of significant duration, as was the case in industry time series 1.

The upshot of all this seems to be that there is a promising future for the systematic analysis of trend data gathered from the digital sphere. Although we are in the earliest phase of our research, we are considering a procedure in which both robust non-linear and SSA analyses are conducted on an ongoing basis. The analyst would rely on the robust non-linear model to obtain key facts about the series of interest (overall fit, regime towards which the series is headed or has attained, identification of outliers) and on SSA to generate a rich visual of the series' underlying trend.

Consider for instance figure 9. (1) The simple depiction of the robust non-linear trend analysis clearly conveys the general idea that Mrs. Royal visibility index was catching up with Mr. Sarkozy's and that even though it suggested a deficit at election time, it appeared to be on track to surpass it. (2) The juxtaposition of the SSA trend over the M-NL trend draws our attention to the event occurring around period 180. (3) Outliers occurring around period 133, 196 and 261 also stand out. (4) The SSA trend extraction generates vital data to try to ascertain the impact of strategic campaign decisions and/or the influence of external events.

In the case of industry data, the ease by which robust non-linear methods can synthesize trends by means of direct measures of eventual plateaus, the current distance from such equilibrium and the presence of outliers, allows business analysts to watch vast numbers of indicators by drawing directly on whatever indicator they choose (i.e. flag current extreme values, track plateaus for shifts over/under critical values).

Figure 9: Presidential time series trend data extracted with the two leading methods



This research is at its very beginning. Now that we have established that there is potential for modelling, our first goal will be to develop our models more carefully. Since we are ultimately concerned with anticipating the turn of future events, we must first develop appropriate forecasting models with reasonable hold-out samples, and then turn our attention to the external validity of such predictions. The first remark may not be obvious to the uninitiated but it is customary and important to make a distinction between data models that are trying to fit a series (which is what we did) and forecasting models where the decision criterion is not the internal fit, but rather how a model can forecast accurately over a given horizon.

We also want to work towards the integration of a watch system (i.e. partial automation of the analysis in conjunction with manual investigation processes).

As final words, it is useful to keep in mind that even though results are extremely encouraging as it appears that digital sphere data accurately track real-world opinions and are amenable to powerful trend extraction procedures, there is a need for replication across a variety of conditions before we can advise business managers to devote resources to such activity.

6 REFERENCES

Alexandrov, Theodore (2006), "Batch extraction of additive components of time series by means of the Caterpillar-SSA method," Vestnik St. Petersburg Univ. Math. 39(2): 71-74

Antoniou, I., Vi.V. Ivanov, Va.V. Ivanov and P.V. Zrelov (2003), "Wavelet Filtering of Network Traffic Measurements," *Physica A* 324: 733-753

Bagchi, Kallol and Somnath Mukhopadhyay (2006), "Predicting Global Internet Growth Using Augmented Diffusion, Fuzzy Regression and Neural Network Models," *International Journal of Information Technology & Decision Making*, 5(1): 155-171

Balazs, George H. and Milani Chaloupka (2004), "Thirty-year Recovery Trend in the Once Depleted Hawaiian Green Sea Turtle Stock," *Biological Conservation* 117: 491-498

Baumgartner, Robert, Oliver Frölich, Georg Gottlob, Patrick Harz, Marcus Herzog and Peter Lehmann (2005) "Web Data Extraction for Business Intelligence: the Lixto approach," Working Paper, DBAI Technology University of Wien

Chau, Michael and Ivy Chan (2005), "A Refinery of an Internet-Based Search Tool – exploring perceptions from information systems practitioners," Working paper, University of Hong Kong

Dekimpe, Marnik G, Philip Hans Franses, Dominique M. Hanssens and Prasad A. Naik (2006) "Time-Series Models in Marketing," Working paper ERS-2006-049-MKT, Erasmus Research Institute of Management

Dekimpe, Marnik G. and Dominique Hanssens (2000), "Time-series Models in Marketing: past, present and future," *International Journal of Research in Marketing,* 17: 183-193

Drezner, Daniel W. and Henry Farrell (2004), "The Power and Politics of Blogs," *Proceedings of the American Political Science Organization.*

Edwards, Samuel Z. (2006), *Forecasting Highly-Aggregate Internet Time Series Using Wavelet Techniques*, unpublished dissertation, Virginia Tech University

Fok, Dennis, Richard Paap, Csilla Horvaqth and Philip Hans Franses (2005), "A Hierarchical Bayes Error Correction Model to Explain Dynamic Effects of Price Changes," Working paper ERS-2005-047-MKT, Erasmus Research Institute of Management

Frances, Philip Hans, Teuns Kloek and André Lucas (1996), "Outlier Robust Analysis of Market Share and Distribution Relations for Weekly Scanning Data," *Working paper*

Franses, Philip Hans (2005), "On the Use of Econometric Models for Policy Simulation in Marketing," *Journal of Marketing Research*, 52(February): 4-14

Fried, Roland (2004), "Robust Filtering of Time Series with Trends," *Nonparametric Statistics* 16(3-4): 313-328

Ghil, M., R. Allen, M.D. Dettinger, K. Ide, D. Kondrashov, M.E. Mann, A.W. Robertson, A. Saunders, Y. Tian, F. Varadi and P. Yiou (2002), "Advanced Spectral Methods for Climactic Time Series," *Reviews of Geophysics* 40(1): 3.1-3.41

Golyandina, Nina, Vladimir Nekrutkin and Anatoly Zhigljavsky (2000), *Analysis of Time Series Structure: SSA and related techniques*, Chapman & Hall/CRC

Howard, Philip N. (2005), "Deep Democracy, Thin Citizenship: the impact of digital media in political campaign strategy," *The Annals of the American Academy of Political and Social Sciences*, 597: 153-170

Huber, Peter J. (1981), *Robust Statistics*, New York: Wiley

Hyung, Namwon and Philip Hans Franses (2005), "Forecasting Time Series with Long Memory and Level Shifts," *Journal of Forecasting*, 24: 1-16

Koelle, Katia and Mercedes Pascual (2004), "Disentangling Extrinsic from Intrinsic Factors in Disease Dynamics: a nonlinear time series approach with an application to cholera," *The American Naturalist* 163(6): 901-913

Little, John D.C. (1970), "Models and Managers: The concept of a Decision Calculus," *Management Science*, 16(8): B466-B485

Marjanovic, Oliviera (2007), "The Next Stage of Operational Business Intelligence: creating new challenges for business process management," *Proceedings of the 40th HICSS*

Maynard, Diana, Milena Yankova, Alexandros Kourakis and Antonio Kokossis (2005) "Ontology-based Information Extraction for Market Monitoring and Technology Watch," Working paper, University of Sheffield

McCann, Michael and William Haltom (2004), "Framing the Food Fights: how mass media construct and constrict public interest litigation," Unpublished manuscript, University of Washington

Nielsen, Heino Bohn (2003), "Cointegration Analysis in the Presence of Outliers," Working Paper, Institute of Economics, University of Copenhagen

Papagiannaki, Konstantina, Nina Taft, Zhi-Li Zhang and Christophe Diot (2005), "Long-Term Forecasting of Internet Backbone Traffic," *IEEE Transactions on Neural Neetworks*, 16(5): 1110-1124

Pauwels, Koen, Imram Currim, Markin G. Dekimpe, Eric Ghysels, Dominique M. Hanssens, Nathalie Mizik and Prasad Naik (2004), "Modeling Marketing Dynamics by Time Series Econometrics," *Marketing Letters*, 15(4): 167-183

Soper, Daniel S. (2005), "A Framework for Automated Web Business Intelligence Systems, *Proceedings of the 38th HICSS*

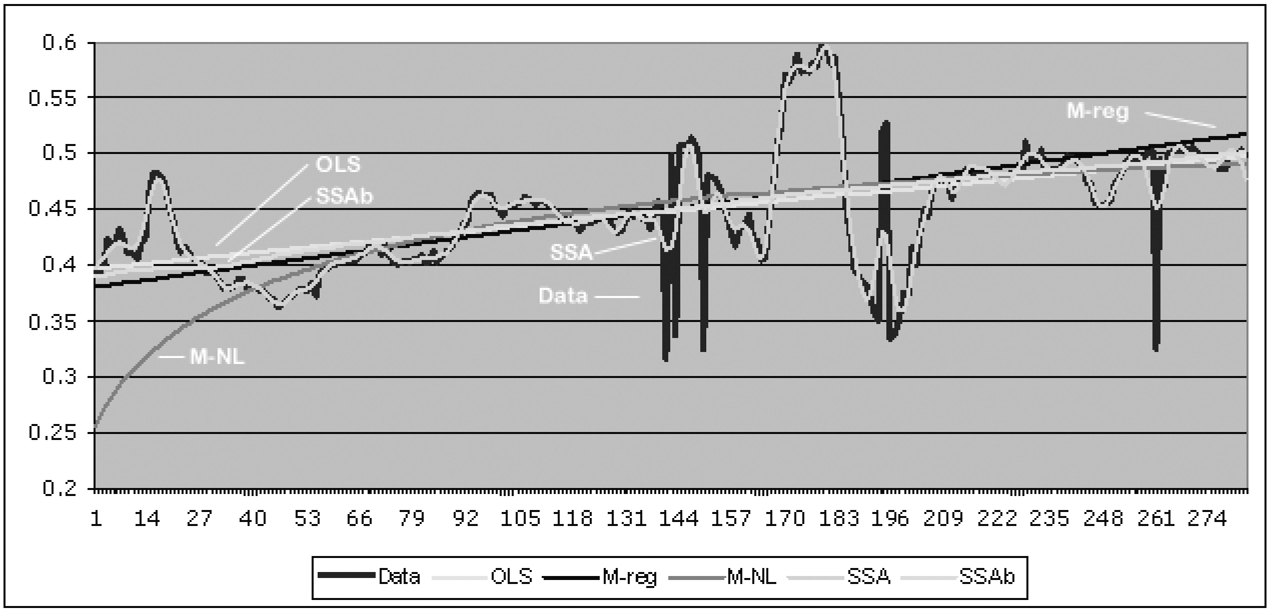
Strnad, Jeff (2005), "Conceptualizing the 'Fat Tax': the role of food taxes in developed economies," *Southern California Law Review*, 78(122): 1221-1326

Surman, Mark and Katherine Reilly (2003), *Appropriating the Internet for Social Change: Towards the strategic use of networked technologies by transnational civil societies organizations,* Final report to the Social Sciences Research Council – Information Technology and International Cooperation Program

Western, Bruce (1995), "Concepts and Suggestions for Robust Regression Analysis," *American Journal of Political Science*, 39(3): 786-817

7 APPENDIX

Figure 1 – Results of the presidential time series (greyscale, annotated)


OLS
SSAb
M-reg
SSA
Data
M-NL
Data
OLS
M-reg
M-NL
SSA
SSAb