Studying dynamic social processes with ARIMA modelling

Citation for published version (APA):

Vasileiadou, E., & Vliegthart, R. (2014). Studying dynamic social processes with ARIMA modelling. *International Journal of Social Research Methodology*, 17(6), 693-708.
<https://doi.org/10.1080/13645579.2013.816257>

DOI:

[10.1080/13645579.2013.816257](https://doi.org/10.1080/13645579.2013.816257)

Document status and date:

Published: 01/01/2014

Document Version:

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy

If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

This article was downloaded by: [Eindhoven Technical University]

On: 13 October 2014, At: 03:57

Publisher: Routledge

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK

International Journal of Social Research Methodology

Publication details, including instructions for authors and subscription information:

<http://www.tandfonline.com/loi/tsrm20>

Studying dynamic social processes with ARIMA modeling

Eleftheria Vasileiadou^a & Rens Vliegthart^b

^a Institute for Environmental Studies, VU University Amsterdam, Amsterdam, The Netherlands

^b Amsterdam School of Communication Research (ASCoR), University of Amsterdam, Amsterdam, The Netherlands

Published online: 09 Jul 2013.

To cite this article: Eleftheria Vasileiadou & Rens Vliegthart (2013): Studying dynamic social processes with ARIMA modeling, International Journal of Social Research Methodology, DOI: [10.1080/13645579.2013.816257](http://dx.doi.org/10.1080/13645579.2013.816257)

To link to this article: <http://dx.doi.org/10.1080/13645579.2013.816257>

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the "Content") contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms & Conditions of access and use can be found at <http://www.tandfonline.com/page/terms-and-conditions>

Studying dynamic social processes with ARIMA modeling

Eleftheria Vasileiadou^{a*} and Rens Vliegenthart^b

^a*Institute for Environmental Studies, VU University Amsterdam, Amsterdam, The Netherlands;* ^b*Amsterdam School of Communication Research (ASCoR), University of Amsterdam, Amsterdam, The Netherlands*

(Received 16 November 2012; final version received 13 June 2013)

With the increasing use of information and communication technologies, there is a wealth of longitudinal data available, which open up new research directions. This availability necessitates special analytical tools, namely time series analysis methods. The paper focuses on Auto Regressive Integrated Moving Average (ARIMA) modeling and provides an outline of how it can be used in social sciences to study dynamic social processes. It provides a typology of dynamics of social processes, using the distinctions between stability vs. fluctuation of a communication process and exogenous vs. endogenous changes. Five distinct types of dynamics of social processes are outlined: stability; linear trend; different attractors; permanent effect; and not permanent effect. Further, the paper examines how these types can be analyzed with the use of ARIMA modeling, and what this means for understanding of the underlying social process. Conclusions are drawn for the use of ARIMA in social sciences, and for understanding of dynamics of social processes.

Keywords: ARIMA; dynamics; time series; communication science

Introduction

With the increasing use of computers and information and communication technologies, large amounts of longitudinal data have become easily accessible to social scientists: databases such as Lexis-Nexis, with archives of print media; large-scale longitudinal surveys of children (Hansen & Joshi, 2007); and panel survey data (such as the British Household Panel Survey or the Panel Study of Income Dynamics). The availability of such longitudinal data enables the study of the evolution of a variable. The paper focuses on how social scientific longitudinal data can be studied, especially on communication processes, with the use of Auto Regressive Integrated Moving Average (ARIMA) modeling.

In social scientific research, longitudinal data are used less frequently than data derived from measurements at a single point in time, because collecting longitudinal data is often time-, money-, and energy-consuming, even though interest in temporal patterns of change in social behavior is increasing (Fitzmaurice, Laird, & Ware, 2011). Yet, if temporal patterns of social processes are studied, the results might be quite different (Kelly & McGrath, 1988). For instance, reviewing studies on the use of communication technologies by groups, Walther suggests that there

*Corresponding author. Email: e.vasileiadou@vu.nl

are fundamental differences between the results obtained from experimental, short-term research and the results from longitudinal research (Walther, 2002), as differences between face-to-face and electronic groups are reduced over time as members interact. Moreover, longitudinal data-sets are required in order to reveal the gender dimension of poverty (Ruspini, 2001). In short, taking time into account offers new perspectives on social processes.

When longitudinal data are available, the methods of analysis used do not always reflect the time series character of the data. This means that crucial information may be lost: for instance, trends in the data-set that are caused by the internal dynamics of the variable under study (Hollanders & Vliegenthart, 2008). This limitation is generally addressed with research designs with two or three discrete data selection points in time, or with panel data-sets. The problem with such designs is that social behavior variables are expected to have random variability, so selecting data every three or four years, which is common practice in e.g. child cohort studies (Wilson, Huttly, & Fenn, 2006), may result in wrong identification of a trend (Kelly & McGrath, 1988).

There is little theorizing about dynamic process theories (Kooiman, 2003), and more precisely about the dynamics of social processes. Variables change over time in distinct, patterned ways; what do these patterns indicate for the social process that the variables measure?

The current paper aims at addressing this gap by studying the dynamics of different communication processes, with a view to building a typology of those dynamics, and moving on to discussing how different types of dynamics can be analyzed with the use of computer-based ARIMA modeling. ARIMA modeling is a statistical modeling method developed to deal with time series data and considered to be the most general type of time series analysis (Boef & Keele, 2008).

The current paper is exploratory and addresses the following two research questions:

- (1) How can different types of dynamics of social processes be classified?
- (2) How can ARIMA be used to understand different communication dynamics?

The aim in this paper is not to present a complete overview of available techniques for time series data (Romer, 2006; Yanovitzky & VanLear, 2007), but to understand dynamics of social processes over time, going beyond a mere application of ARIMA on specific cases.

Communication dynamics: an overview and typology

Interest in dynamic processes has come from many different social sciences: economics, political science, and communication science. Complexity science is also fundamentally interested in issues of continuity and change (Medd, 2002). Thus, even though we use examples from communication sciences to illuminate the typology, the insights are not tied to communication dynamics, but apply to other social processes as well.

Stability vs. change

Some processes may remain fundamentally unchanged over the period of observation; for instance, newspaper coverage of uncontested political issues. Such

processes would be *stable*, even if the indicator used, e.g. number of newspaper articles on scientific topics may vary around a mean. This stability can be more pronounced, especially if a sufficiently long time period is taken into account, or a high aggregation level (e.g. articles in a month vs. articles in a week). Frisch defined a system in stationary equilibrium or stationary state if no tendency to change exists; in other words, if given a point in time t , the values of variables x_t y_t z_t only change randomly (Frisch, 1933). When in such a system in stationary equilibrium a small disturbance is introduced, the system afterwards tends back to the original equilibrium situation. Media attention for a range of issue can be regarded as being in a state of long-term equilibrium. Such pattern can be also seen in communication patterns of small groups, as, after a relatively short period of adjustment and change at the beginning of a group, the group settles in a 'stable state, a fixed value or narrow range of values', which is indicated as a single *attractor* (Arrow, McGrath, & Berdahl, 2000).

Other variables would be *not stable*, for instance they may be increasing over time, such as the amount of scientific publications on climate change. In his work in political science, Kooiman suggests that social systems are usually unstable or disequilibrated; they 'fluctuate' constantly (Kooiman, 2003). This instability is also intuitively understood in processes such as, for instance, media selection in the context of virtual teams, being in constant fluctuation (Shachaf & Hara, 2007). There may be substantial degree of fluctuation between different equilibriums. In the previous example of group communication, initially the amount of communication of a new group fluctuates, until it settles in equilibrium. With an external event, such as a new member joining the group, fluctuation may arise again, until the amount of internal communication settles again in a new equilibrium, perhaps higher than before.

But, there is also the possibility that a variable exhibits cyclic behavior. A system may visit two or more attractors sequentially, a process which may correspond to different phases in a group process: for instance, a group's work-related communication may have two states, between which it alternates, during periods of high workload and low workload (Arrow et al., 2000). Such an attractor, instead of one fixed value, can be a cycle of values that the variable moves in a repetitive sequence (periodic attractor) (Arrow et al., 2000).

Endogenous vs. exogenous change

A further distinction of change is whether the change is internal in the process, or is caused or triggered by an external event or process. In the former case, the change is *endogenous*, as in, for instance, dyadic communication where the interaction between two people feeds from their interaction at previous time lags (Buder, 1991). An endogenous change has also been discussed in media communication, as news remains news because it was news yesterday (Hollanders & Vliegthart, 2008).

Sometimes the change is *exogenous*, as an event independent from the communication process under study influences the process. For instance, an event such as 9/11 will influence media attention on Afghanistan. RSS and Blogs also exhibit exogenous changes (bursts triggered by events such as US elections) and endogenous dynamics, e.g. use of specific words in bloggers' discussions (Lambiotte, Ausloos, & Thelwall, 2007). Such an exogenous change may have a temporary effect on the variable, or a more permanent effect. Focusing on external triggers of

events is often a more fruitful approach to understand social processes than conventional causality (Morrison, 2012).

Thus, a number of different processes or combinations of those processes can be expected:

- (1) Stability: some communication would be stable through time, having reached equilibrium (attractor), indicated with a steady mean of the variable.
- (2) Linear trend: some communication processes would increase or decrease as time goes by, and are thus a function of time, as an independent variable. It can be described as deterministic because the passage of time determines the trajectory of the variable.
- (3) Different attractors: some communication would fluctuate around more than one attractor, for instance, conflict-resolution in an organization, with different periodic attractors in periods of relative peace and periods of internal conflicts.
- (4) Permanent effect: sometimes, a one-off exogenous event may have a permanent effect on the communication dynamics.
- (5) Not permanent effect: sometimes, a one-off exogenous event may have some impact on a communication process, which is not maintained through time. After some time interval, the process returns to its original state.

This is a limited set of processes, as this typology aims to simplify and identify patterns in such inherent complexity. However, these processes occur often in communication studies, as well as other social processes. Moreover, these are processes that can be analyzed with the ARIMA modeling technique.

Method

ARIMA was initially developed by Box and Jenkins (1970) and is, therefore, also called Box-Jenkins transfer modeling. Its main assumption is that a time series variable own past can help to explain its current value and, therefore, before exogenous explanatory variables can even be considered, it is first necessary to model the series' own past and thus capture its endogenous dynamics. ARIMA models are especially suitable for time series with more than 50 observations (Yanovitzky & VanLear, 2007).

A series can be described as a (p,I,q) process; with p autoregressive (AR) terms, I orders of integration (number of times a series needs to be differenced before it is stationary), and q lagged moving-average terms.

Testing for stationarity

The first step in the ARIMA model-building process is to establish whether the time series variable is stationary. The assumption of stationarity is one of the basic requirements for properly modeling time series. A series is stationary if: (1) its mean is unaffected by a change of time origin and thus the expected value is the same for all time points and (2) the variance of the variable is stable through time and neither mean nor variance depend on the time of measurement (Gottman, 1981, p. 61). Stationarity is present when the variable has a stable attractor (mean), and it fluctuates randomly (stable variance) around a fixed attractor (stable mean).

Most often, the augmented Dickey–Fuller (ADF) test is used to establish whether a series is stationary. If the ADF test indicates that the series is non-stationary, the series is said to contain a ‘unit root’ and to be integrated of order one – I(1). It is then necessary to take differences and continue the model-building process with the differenced series. For the differenced series, it is again required to test whether it is stationary. If not, the series is integrated of order two – I(2); and we need to difference the differenced series. Several versions of the ADF test exist (Dickey & Fuller, 1981). This paper uses three different versions: one that includes a trend, one that includes a drift, and one that does not include a drift (Table A1 appendix). Further, a variable may require log transformation, when it is not stationary in its variance.

Determining the AR and moving average terms

Once a stationary series is obtained, one can continue with the second step in the model-building process, determining the AR and moving average (MA) terms. AR terms indicate the lagged endogenous variables that are added to the model and resemble the effects of previous values of the series on the current value. In an ARIMA (1,0,0) model, the value at time t is correlated with the value at time $t - 1$, which in turn was correlated with value $t - 2$, etc. Thus, each shock or disturbance to the system has a diminishing effect on all subsequent time periods (McCleary, Hay, Meidinger, & McDowall, 1980). The formula of the model is:

$$\text{Value}(t) = \text{disturbance}(t) + \varphi \times \text{Value}(t - 1)$$

MAs represent the influence of residuals from previous values on the current value. They resemble shocks (resulting in larger unexplained residuals or error terms) in the series that affect later values of the same series at specific points in time (Yanovitzky & VanLear, 2007). So, in an ARIMA (0,0,1) model, a disturbance affects the process for the current week and the week after that, and then it abruptly ceases to affect it. The formula of the model is:

$$\text{Value}(t) = \text{disturbance}(t) + \theta \times \text{disturbance}(t - 1)$$

The integrated models – e.g. ARIMA (0,1,0) – were already discussed above. Substantially, they reflect the cumulative sum of the shocks in each previous time period. Each value equals the previous value (which is the cumulative sum of changes/differences in the previous stages) and some random fluctuation (disturbance). This type of process is the most sensitive to disturbances, since any shock (event) has a ‘permanent effect’ (Vasileiadou, 2009), and not a diminishing effect (as in the AR-model). In practice, this means that the variable is not stationary. The formula of the model is:

$$\text{Value}(t) = \text{value}(t - 1) + \text{disturbance}$$

The ultimate goal is to properly model the series’ own past, without leaving any information from this past unused. Unused information shows up as autocorrelation in the residuals. In that case, the residuals are not mimicking *white noise*, a term

indicating a signal or process that does not correlate in time, with independent random values with a normal distribution. Therefore, the residual of the model needs to resemble white noise (no autocorrelation). This was tested with the Ljung-Box Q^1 statistic.

In the model-building process, one should strive for parsimony. Different ARIMA specifications of a series may result in residuals that resemble white noise. Model building is usually done by comparing the pattern of the autocorrelation function (ACF; giving the correlation of the series at different lag lengths) and the partial autocorrelations (PACF; which measures the correlation between y_t and y_{t-s} controlling for the effect of intervening values of y between t and $t-s$) of the original series and comparing them to patterns typical for different AR and MA models (see Appendix). In practice, it is often not easy to detect differences between the ACF and PACF figures, so there are a number of additional criteria for the best-fit model: the lowest AIC (Akaike Information Criterion) and Schwarz Bayesian Criterion indicators indicate the most parsimonious model.

Adding exogenous explanatory variables

One can expand the model and add exogenous, explanatory variables. Again, the residuals of the model should mimic white noise. Coefficients and standard errors of the independent variables show whether the effect of the independent variable on the dependent variable is significant, while the AIC and BIC indicate whether the model has improved compared to the univariate model.

Brief comparison with alternative modeling techniques

Other common modeling techniques for time series data are Markov chains and Vector Autoregression (VAR). Markov chains are widely used, because their properties are well understood and they are easy to estimate. However, they are adequate only when a series is stationary. VAR is using only the autoregression component of the more general ARIMA model and assumes two variables that are both independent and dependent. As such, it does not have the flexibility allowed by ARIMA.

Analysis

In this section, we use five different data-sets, from our previous work, shortly mentioning where the data are coming from. Additional details on data collection can be found in the cited references in each of the examples. We use these data-sets as examples of how different ARIMA models relate to different types of dynamic processes.

ARIMA (0,0,1): stable process

The first example comes from the study of a distributed research team (Vasileiadou, 2009). Data were collected from the team-wide emailing list of the research team by downloading all emails from the distribution server, for a period of more than three years. Figure 1 shows the distribution of the variable ‘amount of emails to the emailing list per week.’

Figure 1. ARIMA (0,0,1), a stable process.

The use of the emailing list is divided in three periods of steady increased use between two shorter periods of low use: week 1–10 (approximately); week 28–93; and week 107–171. The distribution of the frequency of the general list is more or less even through time.

The dynamic of the variable is more or less stable through time, having a single attractor (mean). The graph at first glance indicates a stationary process. The series is indeed stationary: the Dickey–Fuller tests indicate that the null-hypothesis of non-stationarity needs to be refuted (Table 1, Appendix, Figure 2). Indeed, there is an ARIMA (0,0,1) with ($\theta = -.40$, $p < .001$) and a constant (15.889, $p < .001$). The MA model indicates that each value of the variable is determined by the current disturbance and the previous disturbance. That means that a shock affects values for one subsequent week. Moreover, the coefficient (θ) is negative, which means that the disturbance has a reverse effect on its use the following week. So, a shock in week 20 may increase the use of the list that week, but in the following week the variable would move back to its initial values (equilibrium) and the shock would be corrected for (Hollanders & Vliegenthart, 2008).

In conclusion, a variable described by an ARIMA (0,0,1), with negative coefficient, suggests an underlying process that is stable over time, with one attractor and evenly distributed variance. Further, the process is not especially sensitive to external shocks, whose impact is only short-term, and is offset when the coefficient is negative. The negative coefficient of the shocks on a communication system suggests a system of low degree of complexity, which tends to be stable and predictable (Vasileiadou, 2012). This is a type of process indicated in the theoretical section as a stable process.

ARIMA (0,0,1) with linear trend

The second example comes from the study of the same distributed research team. Data were collected from the management emailing list of the research team by downloading all emails from the distribution server, for a period of more than three years. The management list was used for communication of all administrative messages documents. Figure 2 shows the distribution of the variable ‘amount of emails to the management list per week.’

The use of the management list shows a linear negative trend, as emails decrease over time. The research team had more frequent communication around

Figure 2. ARIMA (0,0,1) with linear trend.

administration issues in the beginning of the project than in the end. Since the administration issues were the same throughout the duration of the research project, it is assumed that the decrease in the amount of administration-related emails indicates a learning process in the team, as the team members understood the procedures.

The graph at first glance indicates a negative linear trend. The series is, however, stationary: the Dickey–Fuller tests indicate that the null-hypothesis of non-stationarity needs to be rejected (Table 1, Appendix, Figure 2). However, it is still useful to model this linear trend, which can provide some indication about the behavior of the variable. The linear trend can be modeled with the use of a dummy variable $t = \text{time}$ with values $t_1 = 1, t_2 = 2 \dots t_x = x$. This dummy variable is then used as independent variable in the ARIMA modeling. The model that was fit was an ARIMA (0,0,1) with ($\theta = -.398, p < .001$) a constant (7.031, $p < .001$), and the independent variable time ($-.059, p = .002$). The MA model indicates that each value of the variable is determined by the current disturbance and the previous disturbance. That means that a shock affects values for one subsequent week. Moreover, the coefficient (θ) is negative, which means that the disturbance has a reverse effect on its use the following week. So, a shock in week 20 may increase the use of the list that week, but in the following week the variable would move back to its initial values (equilibrium) and the shock would be corrected for. The negative linear trend indicates the learning process in the team.

In conclusion, a variable described by an ARIMA (0,0,1), with negative coefficient and linear trend, suggests an underlying process that is decreasing (or increasing) in a linear way over time. Further, it is not especially sensitive to external shocks, whose impact is only short-term, and is offset when the coefficient is negative.

ARIMA (1,0,1): permanent effect

A variable that lends itself very well for ARIMA modeling is media attention for actors or issues. This example is about the weekly attention that is being devoted to

the issue of immigration and integration in the Dutch newspaper *Telegraaf* in the period 1999–2004. We collected all articles on the topic in this newspaper per week. Figure 3 presents the series.

A visual inspection of this figure indicates that there are roughly two periods: one of low attention on the topic, and another, subsequent period, of constant and higher attention on the topic. Analysis elsewhere has indicated that the two periods can be distinguished before and after 9/11 (Roggeband & Vliegthart, 2007). This indicates that the process might be non-stationary and resembles a process with multiple periodic attractors.

Going through the model specification process, the series turns out to be stationary: the Dickey–Fuller tests indicate that the null-hypothesis of non-stationary needs to be refuted (Table 1, Appendix, Figure 3). Apparently, the difference in means between the two periods is not large enough to make the process non-stationary. The final model is an ARIMA (1,0,1), which includes a positive AR component ($\phi = .825$, $p < .001$) – indicating that current week’s attention depends on the attention the issue received in the previous week – and a negative MA component also with a lag of one ($\theta = -.526$, $p < .001$) – indicating that the previous error term affects the current value of the series.

The dynamics specification makes a lot of sense intuitively and has a straightforward interpretation. The AR(1) term reflects the stable part of the news agenda: news is (partly) being news, because it was news last week. The negative MA term often shows up when modeling media attention for specific issues, where an increase in attention caused by a newsworthy event results in a high value and a relatively high unexplained error term. This increase in attention is often followed by a quick return to initial values, resulting in a situation where higher error terms in the past result in sharper decreases afterwards, which can be modeled with negative MAs.

Adding an exogenous variable

Continuing with the same example (Figure 3), the level of attention on immigration issues fluctuates considerably over time, but is clearly higher in September 2001

Figure 3. ARIMA (1,0,1).

and also after that. 9/11 seems to have structurally increased the attention that is being devoted to immigration and integration issues.

Adding an exogenous dummy variable to the model that takes the value ‘0’ before 9/11 and ‘1’ after that, the model improves. The effect of the variable is significant (coefficient=2.73, $p < .001$). After 9/11, there are on average almost three more articles published on the issue per week.

Substantially, the results indicate that events that take place outside the media realm can substantially and even structurally – at least for the period considered here – change the level of attention journalists devote to certain issues. Thus, the type of dynamic discussed as permanent effect in the theoretical section is modeled here with the use of an exogenous dummy variable in the ARIMA (1,0,1).

ARIMA (0,1,1): different attractors

The following example comes from the study of the climategate incident, when emails by climate scientists were published on the internet without their authors’ consent. Data were collected from LexisNexis, selecting all international English-speaking newspapers, with the search term ‘climategate’ occurring anywhere in the text, from 17 November 2009 to 28 February 2010. Figure 4 shows the amount of newspaper articles mentioning ‘climategate’ per day.

Figure 4 shows a clear distinction of the newspaper articles in three periods. First, the debate hypes in the period 1 December–23 December, followed by a relatively low-attention period between 23 December and 16 January 2010, probably because of Christmas. From 17 January onwards, the issue gains more attention, but far less than in the original hype phase. These three periods identified here have completely different intensity of communication, with three temporary attractors (means), indicating a non-stationary variable.

The variable is indeed non-stationary: the Dickey–Fuller tests indicate that the null-hypothesis of non-stationarity holds (Table 1, Appendix, Figure 4). The model

Figure 4. ARIMA (0,1,1).

that was fit was an ARIMA (0,1,1) with positive MA coefficient ($\theta = .361$, $p < .001$) without a constant.

The integrated process indicates that a random shock, such as the hacking of the scientists' emails, had an accumulated effect on the number of articles, influencing it for all the period. This suggests permanent effect of random shocks (Vasileiadou, 2009). Further, the coefficient in the MA model is positive, indicating that the influence of the disturbance of the previous time lag is positive. So, in the (0,1,1) models, a larger than expected change in day t results in greater than expected change in day $t+1$, and the influence of the shock is not corrected for.

In conclusion, a variable described by an ARIMA (0,1,1), with positive coefficient, suggests an underlying communication process that is not stable over time: it has more than one temporary attractors. Further, it is very sensitive to random shocks, modeled as disturbances by ARIMA, indicating a system with high degree of complexity (Vasileiadou, 2012).

The dynamic of periodic attractors indicates a non-stationary process that may need differencing and can thus be modeled with an ARIMA (0,1,1).

ARIMA (0,1,1) with log transformation: different attractors

The following example comes from another distributed research team (Vasileiadou, 2009). Figure 5 indicates the number of documents downloaded from the common server of the team per week. The data were collected from the server (log files). Almost all activities are concentrated in the period week 76–105. After this period, the downloading activity decreased substantially, with only sparse activity. Here, again the distribution of the downloading activity is not even in time: there is no stable mean. This visual inspection already suggests that differencing will be needed. In addition, Figure 5 indicates that the higher the values of the variable, the higher its variance. This suggests the need for log transformation in the ARIMA modeling.

Figure 5. ARIMA (0,1,1) with log transformation.

Table 1. Analyzing dynamics with ARIMA.

Type of dynamic	Modeling element
Stable process	ARIMA (0,0,1) with negative MA co-efficient
Linear fluctuation (trend)	Exogenous dummy variable, with one in $t=1$, 2 in $t=2$, etc.
Different attractors	Integrated model (need for differencing) Positive coefficients for error terms in the MA component Need for log transformation when variance is unstable
Permanent effect	Exogenous dichotomous dummy variable, with different values in the 'before' and 'after' periods
Not permanent effect	A range of non-permanent effects would indicate stable process (negative MA coefficient) ^a

^aOne big non-permanent effect would need to be modeled with exogenous dichotomous dummy variable.

The variable is not stationary: the Dickey–Fuller tests indicate that the null-hypothesis of non-stationary holds (Table 1, Appendix, Figure 5). The ARIMA model that was fit was an ARIMA (0,1,1) with natural log transformation with positive MA coefficient ($\theta = .624$, $p < .001$).

The variable follows an integrated process, which suggests permanent effect of random shocks, and lack of stabilization of the communication activity. Further, the coefficient of the MA component is positive, indicating that the influence of the disturbance of the previous time lag is not corrected for at subsequent time lags.

In conclusion, a variable described by an ARIMA requiring log transformation suggests an underlying communication process that is not stable in time. The process under study here fluctuates more in higher values, than in lower values, which suggests an activity which has its 'normal' values at low level. A lot of communication activities, e.g. time spent on communication media per day, have a natural 'floor.'

Therefore, a process with periodic attractors can be modeled with an ARIMA (0,1,1), adding a log transformation when the variance of the variable is unstable.

Table 1 summarizes the results, matching the different types of dynamic processes discussed in the theoretical section, with the ARIMA modeling elements from this current section.

Conclusions and discussion

Often the social process under study is more or less stable, fluctuating around a mean (attractor), especially when the time lags are long enough so that daily or weekly fluctuations are averaged out. For instance, the amount of time an individual spends on face-to-face communication over time would probably remain the same, if one studies it month by month. Such a process would be modeled with an AR(1) or a MA(1) model with a negative coefficient, as the disturbance at each time lag would be corrected for in subsequent time lags. Such stationary models are suited when a variable has a stable mean and variance.

The case of linear upward or downward trend can be found often when studying communication processes. An example identified here was exchange of emails on administration issues in a distributed research team, but it would apply also to number of new homepages per month, Facebook visitors, newspaper circulation, etc. It can be described as deterministic because time determines the fact that the variable increases, or decreases. This was modeled with an exogenous dummy variable that added explanatory power to the (endogenous) ARIMA process.

Other social processes are not stable in time, especially when the time horizon of the study is long enough to show fluctuations. A process may fluctuate between two or even more different attractors (means). The example used here was the amount of newspaper articles on climategate. Such a process would be modeled using an integrated model (differencing the values), in addition to possible AR or MA parts. The coefficient of the AR or MA components would be positive, which suggests that the influence of the past (value or disturbance) is not corrected for in subsequent time lags.

Furthermore, there are social processes whose indicators fluctuate more around a specific attractor, than around another attractor: the variance of the variable is not stable. Modeling such variables would probably include a log transformation, which would even out the fluctuation of the variance over time, in addition to other ARIMA components.

Finally, there are processes that are clearly influenced by exogenous variables, apart from their own dynamics, the so-called bursts. The example here is the newspaper attention on immigration issues before and after 9/11. Such exogenous variables can be modeled with the addition of an exogenous, dichotomous dummy variable, in addition to the other ARIMA components.

These insights provide a novel contribution to theory and methodology of dynamic social processes. Unfortunately, there is little conceptual work on types of communication dynamics. There is even less work which combines conceptual insights on dynamics with specific methods. However, when modeling a variable with a specific kind of ARIMA model, there are a number of conceptual implications for the underlying communication process, as shown here. The paper is not unique in its plea for the application of time series techniques. In political science, where suitable data (e.g. on president approval and socio-economic indicators) are more widely available, all kinds of econometric time series techniques have been introduced (Box-Steffensmeier & Smith, 1998; Freeman, Williams, & Lin, 1989). However, there is little other work that attempts to use the ARIMA model and its different components to theorize specifically about different types of dynamics.

One limitation of the current paper is the lack of multivariate ARIMA modeling. The types of models presented, and theorized in this paper, relate (mostly) to univariate modeling, whereby a complex social phenomenon, such as a media hype, is simplified, and thus reduced, because we only use one indicator to measure it (amount of newspaper articles). Similarly, team communication can be reduced to one or two indicators (i.e. amount of emails exchanged, amount of documents exchanged). One of the strengths of ARIMA is that it offers the possibility to conduct multivariate modeling, to establish inter-relations among variables at different levels of analysis, and different time lags. This is particularly pertinent in complex dynamic systems, where there are individual-, system-, and context-level variables influencing each other (Vasileiadou & Safarzyńska, 2010). A similar pertinent

method for complex systems is dynamic modeling of longitudinal network data (Snijders, 2001).

One final note relates to the distinction between prediction of endogenous change and causation, including exogenous factors. Even though the main aim of most social scientific research is to discover causation between variables, this may not always be possible, or desirable. For instance, in complex dynamic systems, where many elements mutually influence each other (in a negative or positive feedback loop), causality may be more difficult (if at all possible) to discern (Arrow et al., 2000). For social scientific variables, it is too simple to suggest that a series is determined by its own past, plus some stochastic shocks (Hollanders & Vliegenthart, 2008). In fact, it is obvious that social behavior is influenced in one or another way by countless other things. Even when the interest is in those exogenous factors, identifying endogenous change is important, because it helps us, at a second step, identify and establish in a robust way how those factors have an impact. Thus, prediction of endogenous change is the first step towards establishing causation with exogenous factors.

Note

1. This was preferred over the widely used Durbin-Watson statistic, which only tests for first-order autocorrelation and is not valid when the model includes AR terms.

Notes on contributors

Eleftheria Vasileiadou is an assistant professor in science and public policy at the Institute for Environmental Studies, VU University Amsterdam, and lecturer at the department of Communication Science University of Amsterdam (UvA). She works on issues on science production and communication, and complexity theory and methodology.

Rens Vliegenthart is an associate professor in political communication at the department of Communication Science and at the Amsterdam School of Communication Research (ASCoR), University of Amsterdam (UvA).

References

- Arrow, H., McGrath, J. E., & Berdahl, J. (2000). *Small groups as complex systems. Formation, coordination, development and adaptation*. Thousand Oaks: Sage.
- Boef, S. D., & Keele, L. (2008). Taking time seriously. *American Journal of Political Science*, 52, 184–200.
- Box, G., & Jenkins, G. (1970). *Time series analysis, forecasting and control*. San Francisco: Holden Day.
- Box-Steffensmeier, J. M., & Smith, R. M. (1998). Investigating political dynamics using fractional integration method. *American Journal of Political Science*, 42, 661–689.
- Buder, E. H. (1991). A nonlinear dynamic model of social interaction. *Communication Research*, 18, 174–198.
- Dickey, D. A., & Fuller, W. A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica*, 49, 1057–1072.
- Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied longitudinal Analysis* (2nd ed.). Hoboken, NJ: Wiley.
- Freeman, J. R., Williams, J. T., & Lin, T. (1989). Vector autoregression and the study of politics. *American Journal of Political Science*, 33, 842–877.
- Frisch, R. (1933). On the notion of equilibrium and disequilibrium. *Review of Economic Studies*, 3, 100–106.

- Gottman, J. M. (1981). *Time-series analysis: A comprehensive introduction for social scientists*. Cambridge: Cambridge University Press.
- Hansen, K., & Joshi, H. (2007). Editorial: Special issue on child cohort studies. *International Journal of Social Research Methodology*, 10, 319–323.
- Hollanders, D., & Vliegthart, R. (2008). Telling what yesterday's news might be tomorrow: Modeling media dynamics. *Communications*, 33, 47–68.
- Kelly, J. R., & McGrath, J. E. (1988). *On time and method* (Vol. 13). Newbury Park: Sage.
- Koolman, J. (2003). *Governing as governance*. London: Sage.
- Lambiotte, R., Ausloos, M., & Thelwall, M. (2007). Word statistics in blogs and RSS feeds: Towards empirical universal evidence. *Journal of Informetrics*, 1, 277–286.
- McCleary, R., Hay, J. R. A., Meidinger, E. E., & McDowall, D. (1980). *Applied time series analysis for the social sciences*. Beverly Hills: Sage.
- Medd, W. (2002). Complexity and the social world. *International Journal of Social Research Methodology*, 5, 71–81.
- Morrison, K. (2012). Searching for causality in the wrong places. *International Journal of Social Research Methodology*, 15, 15–30.
- Roggeband, C. M., & Vliegthart, R. (2007). Divergent framing: The public debate on migration in the Dutch Parliament and media, 1995–2004. *West European Politics*, 30, 524–548.
- Romer, D. (2006). Time series models. In D. Romer, K. Kenski, K. Winneg, C. Adasiewicz, & K. H. Jamieson (Eds.), *Capturing campaign dynamics, 2000 and 2004; The National Annenberg Election Survey* (pp. 165–243). Philadelphia, PA: University of Pennsylvania Press.
- Ruspini, E. (2001). The study of women's deprivation: How to reveal the gender dimension of poverty. *International Journal of Social Research Methodology*, 4, 101–118.
- Shachaf, P., & Hara, N. (2007). Behavioral complexity theory of media selection: A proposed theory for global virtual teams. *Journal of Information Science*, 33, 63–75.
- Snijders, T. A. B. (2001). The statistical evaluation of social network dynamics. *Sociological Methodology*, 31, 361–395.
- Vasileiadou, E. (2009). Stabilisation operationalised: Using time series analysis to understand the dynamics of research collaboration. *Journal of Informetrics*, 3, 36–48.
- Vasileiadou, E. (2012). Research teams as complex systems: Implications for knowledge management. *Knowledge Management Research & Practice*, 10, 118–127.
- Vasileiadou, E., & Safarzynska, K. (2010). Transitions: Taking complexity seriously. *Futures*, 42, 1176–1186. doi: 10.1016/j.futures.2010.07.001
- Walther, J. B. (2002). Time effects in computer-mediated groups: Past, present and future. In P. Hinds & S. Kiesler (Eds.), *Distributed work* (pp. 235–257). Cambridge: The MIT Press.
- Wilson, I., Huttly, S. R. A., & Fenn, B. (2006). A case study of sample design for longitudinal research: Young lives. *International Journal of Social Research Methodology*, 9, 351–365.
- Yanovitzky, I., & VanLear, A. (2007). Time series analysis. Traditional and contemporary approaches. In A. F. Hayes, M. Slater, & L. B. Snyder (Eds.), *The SAGE sourcebook of advanced data analysis methods for communication research* (pp. 89–124). Thousand Oaks: Sage.

Appendix 1. Using ACF and PACF to identify ARIMA models

The ACF of an AR(1) process exhibits a geometrically declining pattern, falling down to zero as the lag length increases. For a MA(1) process, the autocorrelation drops to zero for all lag lengths larger than one, and, more generally, drops to zero for lag length larger than q for a MA(q) process. The partial ACF mirrors the pattern of the ACF. It is declining for a MA(1) process and drops to zero at lag length larger than one for AR(1) processes.

The following figures show ACF and PACF for exemplary ARIMA models.

Table A1. ADF tests.

	Without drift and trend (random walk)	With drift and trend	With drift, without trend
Figure 1	−2.341	−4.924	−4.943
Figure 2	−2.724	−5.071	−3.730
Figure 3	−2.703	−4.520	−5.648
Figure 4	−1.359*	−2.521*	−2.000*
Figure 5	−1.589*	−1.857*	−2.406*

*indicates non-stationarity at a 5%-level.