Structure and Dynamics: eJournal of Anthropological and Related Sciences

Volume 1, *Issue* 1 2005 *Article* 4

A Primer on Statistical Analysis of Dynamical Systems in Historical Social Sciences (with a Particular Emphasis on Secular Cycles)

Peter Turchin*

*University of Connecticut, peter.turchin@uconn.edu

Abstract

This primer explicates the conceptual foundations of the statistical approach to detecting dynamical feedbacks. It is assumed that we have time-series data on several aspects of the studied system. The basic idea of the approach is to regress discrete rates of change of measured variables on variables themselves. I discuss several issues involved in the analysis, such as how to select the appropriate time step, or the delay parameter. The goal of the analysis is to determine whether a particular predictor variable, or set of variables, has a statistically detectable effect on the response. This is accomplished by cross-validation.

Keywords: population, state breakdown, historical dynamics, methodology

Suggested Citation:
Peter Turchin (2005) "A Primer on Statistical Analysis of Dynamical Systems in Historical Social Sciences (with a Particular Emphasis on Secular Cycles)", *Structure and Dynamics: eJournal of Anthropological and Related Sciences*: Vol. 1: No. 1, Article 4.
http://repositories.cdlib.org/imbs/socdyn/sdeas/vol1/iss1/art4

The dynamical systems approach

Dynamics is the scientific study of any entities that change with time. The two most important goals of dynamics are (1) quantitative description of the observed behavior (sometimes this is called *kinematics*) and (2) explanation of the observed patterns. In natural sciences dynamics may also be used for forecasting and control, but in typical social science applications the theory has not developed to the point where this is possible.

A highly useful approach to studying dynamical phenomena is to think of them as *systems*. This involves artificially separating a holistic natural (or, rather, social) phenomenon into parts—elements or subsystems. The assumption is that the dynamics of the whole can be explained by studying how its parts interact with each other. The success of the enterprise, obviously, depends very much on how the whole is sliced up. One of the most important issues in dynamical analysis, therefore, is to determine just what is the most effective way of representing the studied phenomenon as a system. Efficiency in this context means finding the simplest possible representation of the system (the fewer variables the better) that allows the most explanatory (and, eventually, predictive) power.

The mathematical approaches to the study of dynamical systems have been with us since the days of Newton and Leibnitz. The most common (and incredibly fruitful) mathematical tool is the differential equation, which looks like this:

$$\dot{X} = f(X) \tag{1}$$

where X is a variable describing some aspect of the system (for example, it could be the population density). On the left hand side we see X with a dot on top, which denotes the derivative, or rate of change of X. To the right of the equals sign, $f(X)$ means some function of X. For example, if $f(X) = rX$, then we have an exponential model: $\dot{X} = rX$, which assumes that the rate of change of the variable X is directly proportional to the value of the variable, X.

There are many variations on the basic theme of Eqn. (1). First, we could think of X as not a scalar but a vector, whose elements represent different variables describing the state of the system (see below). Alternatively, we could write it out in scalar form. For example, if there are two variables, X and Y, the model would look like this:

$$\begin{aligned} \dot{X} &= f(X,Y) \\ \dot{Y} &= g(X,Y) \end{aligned} \tag{2}$$

The second variation is to use not continuous models, such as Eqn. (1), but discrete-time models, e.g., $X_{t+1} = f(X_t)$. Other modifications/complications include explicit handling of time delays, adding space (which leads to partial differential equations), and so on. For our purposes, we only need to know about the continuous and discrete forms of dynamical models.

After we have represented the studied phenomenon as a dynamical system, the next step is to decide what variables we will use in describing the state of the system—these are the *state variables*, or *structural variables* (the latter designation is preferable in cliodynamic applications,

since the term *state* is better reserved for a kind of polity). For example, when studying secular cycles (centuries-long oscillations in demographic, economic, social, and political structures of agrarian societies), we might focus on three main parts of a society: the state, the elites, and the commoners. The set of structural variables, then, might be something like population numbers of elites and commoners, average incomes of these two strata, and the fiscal health of the state. Coming up with a set of structural variables is really an integral part of representing the studied phenomenon as a system. *A priori* we don't know what the most efficient (in the sense defined above) set of variables is going to be; we arrive at such a set by the method of trial and error (a well-developed modeling intuition is invaluable in this process).

If we are interested in generating theoretical insights about how the studied system might function, the next step is to start making assumptions about how structural variables are interrelated, writing down explicit models, and investigating their behaviors. This essentially *deductive* process is described in my 2003 book *Historical Dynamics*. An alternative, or rather complementary, investigation is to use statistical methods to analyze time-series data describing various aspects of the system. Such an approach can aid in generating insights about the functioning of the system, and also be used in testing theoretical models.

Analysis of time-series data: the conceptual foundations

Time-series data are repeated measurements of some aspect of the system, typically (but not always) taken at regular time intervals. It is possible to obtain meaningful results even when a univariate data series is available (see Chapter 7 in Turchin 2003a), but our ability to generate insights is greatly expanded when we have multivariate series describing many different aspects of the empirical system. Here I focus on the analysis of multiple time series.

The goal of the analysis is to determine what kind of dynamic interrelations, if any, characterize different aspects of the studied phenomenon for which we have time-series data. The basic model underlying analysis is a modification of Eqns. (1) and (2). Suppose we have measurements of two structural variables, X and Y. Then the model looks like this:

$$\begin{aligned} \dot{X} &= f(X,Y,Z) \\ \dot{Y} &= g(X,Y,Z) \end{aligned} \tag{3}$$

Here Z is the *exogenous* variable, or variable that affects X and Y, but is not affected by them in return. X and Y are *endogenous* variables because they affect each other through feedback loops, designated by the functions f and g. Z can stand for known, and measured, aspects of the empirical system, in which case we can explicitly bring it in the analysis. Alternatively, Z may represent unknown effects, in which case we typically model it as a stochastic variable. In practice, we should always include a stochastic variable—the noise term—in the model which we use to investigate dynamics of real-life systems (whether in natural or social sciences). This is because no model should be expected to capture all aspects of the studied system. Those factors that we don't have explicit data on are folded into the noise term. Sometimes, even when we have data on some factors that have a minor effect on dynamics, we might chose to treat them as part of a collective noise term for the sake of parsimony. Finally, it should be noted that the division of variables into exogenous and endogenous ones is not clear *a priori*, but is arrived at in the process of modeling and data analysis.

If we can somehow estimate the derivatives of X and Y on the left-hand side, then Eqns. (3) becomes a regression model, with the derivatives serving as dependent (or response)

variables, and the values of X and Y themselves serving as independent (or predictor) variables. Z is, as was explained above, the error term. The nonstandard aspects of using Eqns. (3) as a regression model are, first, that functions f and g are most likely nonlinear. Thus, we need to employ extensions of the standard linear regression, for example, polynomial regression. Second, the values of serially measured variables are not statistically independent; they are autocorrelated. This means that we need to use extra care in interpreting regression statistics, such as F and P values.

How can we estimate the derivative $\dot{X}$ from the time series of X values? The variable itself is measured at discrete intervals, while the derivative assumes that X is continuous. One approach is to smooth the observed time series, using such techniques as splines or kernel regression. If following this course, it is critical to choose well the smoothing parameter, the bandwidth. If the chosen bandwidth is too narrow, the "under-smoothed" trajectory will trace closely the measured values of X. But variables in the real world are never measured precisely; there is always a measurement error. When the smoothed trajectory follows closely the stochastic fluctuations in the data series, we will get a highly variable estimate of the derivative. Most of the fluctuation in the derivative estimate will be a result of measurement noise, and it will be difficult to detect the underlying signal of how the structural variables affect the rates of change.

Using a bandwidth that is too broad also causes problems. The over-smoothed trajectory will miss some important signals in the trajectory by smoothing away such revealing movements.

In practice, time-series analysts rarely use Model (3) directly—partly because of the difficulty of finding the optimal smoothing bandwidth, partly because smoothing introduces extra autocorrelations in the data, and partly because using smoothing techniques is a rather technical field. The major alternative to direct estimation of derivatives is to use the discrete rates of change:

$$\Delta X_t = X_{t+\tau} - X_t \tag{4}$$

where τ is the time delay, which could be equal to one, two, or more time intervals at which data were collected. An estimate of the derivative is $\Delta X/\tau$; the smaller τ, the better this quantity approximates the derivative. We can, thus, modify Model (3) as follows:

$$\begin{aligned} \Delta X_t &= f(X_t, Y_t, Z_t) \\ \Delta Y_t &= g(X_t, Y_t, Z_t) \end{aligned} \tag{5}$$

Because many people do not like the implied circularity of using X_t on both left-hand side (as part of the definition of ΔX) and right-hand side, we can simply use $X_{t+\tau}$ as the response variable:

$$\begin{aligned} X_{t+\tau} &= f(X_t, Y_t, Z_t) \\ Y_{t+\tau} &= g(X_t, Y_t, Z_t) \end{aligned} \tag{6}$$

In other words, we have switched from a continuous to a discrete model as the basis for the analysis of data.

Choosing the delay parameter

The appropriate choice of the time delay parameter, τ, is very important (this parameter is analogous to the smoothing bandwidth, when estimating the derivative). Unfortunately, there are no universal guidelines for making this choice, although we now understand that we need to avoid two opposite extremes—redundance and irrelevance (Casdagli et al. 1991). *Redundance* occurs when time delay is very small, so that the expectation of X_t and $X_{t-\tau}$ are practically the same. The opposite problem of *irrelevance* occurs when τ is so large that X_t and $X_{t-\tau}$ are not functionally related anymore, as a result of cumulative effects of noise (and trajectory divergence, if the system is chaotic).

An optimal choice of τ lies between the two extremes of redundance and irrelevance. It would be nice to be able to estimate the lag parameter τ from the data. A naïve approach is to fit models with different values of time delay, and choose the one that yields the highest coefficient of determination, R^2. A little thought shows why this approach does not work. If our response variable is the discrete derivative, such as in Model (5), then by varying τ we also vary the dependent variable. A comparison of how well models fit different response variables is akin to comparing apples and oranges—it is meaningless. Discrete derivatives based on different time lag will be characterized by different variances (generally, as we increase τ from small values, the variance of ΔX will also increase, and will reach a maximum at τ approximating half the period of oscillations). Thus, if we use R^2 then not only its numerator (the variance explained by the model) will change, but also it denominator (the total variance). Other best-fit statistics run into similar problems.

What about using Model (6)? Now we are comparing apples with apples, so the problem explained in the previous paragraph is avoided. However, we run into a different problem—typically the best model is simply $X_t = X_{t-\tau}$ for the smallest possible time lag. This model will explain the enormous proportion of variance—the only variance left unexplained would be that due to measurement noise. Due to this problem of redundance, no other model would be able to do better (except in a spurious sense, by fitting measurement noise), so this approach is also self-defeating. This consideration reinforces the importance of choosing the lag parameter large enough to avoid redundance. Ellner and Turchin (1995) recommend using a time delay long enough to get the correlation coefficient between X_t and $X_{t-\tau}$ down to the vicinity of 0.5 (which would correspond to R^2 of roughly 0.25, ensuring that redundance would account for a relatively minor proportion of variance explained).

The take-home message is that we cannot rely on the data themselves for selecting the optimal time delay (at least, so far nobody figured out a way to get around the problems discussed above). Therefore, we must choose the lag parameter using *a priori* considerations (the data should be used to check for redundance, e.g., using Ellner and Turchin's "rule of thumb"). In fact, when we think about it, it is only appropriate that τ should be chosen based on the goals of the analysis, rather than on the data. To understand this point better, let us take a short excurse into the question of time scales (see also discussion on p. 150ff in Turchin 2003b).

Time scales and the choice of the delay parameter

In general, different social processes operate at a variety of temporal scales. The shorter scales include daily, weekly, monthly, and annual cycles. Beyond that we have human generations, processes occurring on the time scale of centuries, such as secular cycles, and longer-term phenomena such as social and biological evolution. As an example, consider the stock market, as measured by the Dow-Jones Industrial Average (DJIA). DJIA fluctuates on a

variety of scales: daily (because the stock exchange shuts down at night), weekly (no activity on weekends), annual (fiscal year accounting affects trader behavior), multi-annual (business cycles), and multi-decadal (the Kondratieff cycle, although not everybody accepts the reality of such long cycles). The DJIA trajectory looks "fractal," because the amount of fluctuation depends on the time-scale at which the trajectory is viewed.

If we are interested in understanding the effect of business cycle on the stock prices we really don't care about short-term fluctuations. We certainly should ignore price movements with a single day, and probably even within a week. Thus, the time-series with which we would want to investigate multi-annual oscillations would probably use the values of Dow-Jones averaged for each week. Averaging is the simplest kind of smoothing, so what we have done is essentially smoothed away all "uninteresting" short-term fluctuations—uninteresting, that is, from the point of view of the main question of analysis. On the other hand, if we want to know how holiday periods affect stock price movements, we would certainly want to retain within-week fluctuations, and perhaps go down to hourly movements (to see how trading patterns behave during the short pre-holiday days). Now the variation due to the business cycle becomes a nuisance, and it might be a good idea to remove the effect of multiannual and longer-term fluctuations by detrending. The point is that different questions require approaching analysis at different time scales.

Taking population dynamics, they also occur on a variety of scales: monthly (female menstruation periods), yearly (subsistence and epidemic cycles), generational (somewhere between two and three decades), and secular (two or three centuries, if the theory of secular cycles is to be believed). If we are interested in the dynamics of childhood diseases, then the appropriate time scale would be weeks or months, to capture the within-year course of each epidemic (the incidence of measles, for example, begins to grow after children are brought together at the beginning of the school year, and gradually builds up towards a peak in winter).

If we want to understand how secular cycles unfold, on the other hand, we certainly don't care how mortality fluctuates on a weekly or monthly time scale. Or that there may be a deficit of births nine months after the Lent, as a result of devout Christians avoiding sexual intercourse. All such within-year, or even year-to-year fluctuations are irrelevant to the purposes of our investigation. The appropriate time step is one human generation, and we need to average over smaller-scale fluctuations. We also need to do something about very long trends driven by social evolution. This requires some kind of removal of millennial trends, for example as was done for the English population in the accompanying paper (Turchin 2005). By smoothing within-decade fluctuations and removing millennial trends we retain two temporal scales of interest. The longer one is the average period of the secular cycle—this is what needs to be explained. The shorter one is the human generation time—this is the time step of the dynamical process that is postulated to be the explanatory mechanism of secular cycles. Thus, the delay parameter, τ, is set by consideration external to the data themselves. It is a good idea, however, to try a couple of different values of τ, e.g., 20 and 30 years. If results of the analysis are qualitatively the same, everything is fine. If not, we have a problem—high sensitivity to the specific choice of the delay parameter throws in doubt any results we have obtained, and the reason for such sensitivity must be understood.

Thus, the general procedure for selecting the lag parameter is not to estimate it from the data, but to choose it on the *a priori* grounds. Ideally this should be done prior to seeing the data, in order to remove any potential doubts that this parameter was actually selected *a posteriori* to make the results appear better. In the case of secular cycles, this has been done (see p. 154 in

Turchin 2003b, a general recommendation of setting τ equal to general time in the context of animal population dynamics is on p. 187 in Turchin 2003a).

Independent measures of the derivative sidestep the problem of choosing the delay parameter

There is one situation when the whole question of choosing the time lag becomes moot—when we have independent data on the rate of change of a variable. For example, the per capita rate of population change, estimated by the difference between births and deaths (assuming that we can neglect immigration and emigration), is closely related to the rate of change of population numbers (it's the derivative of the log-transformed population size). We are fortunate to have a time-series of per capita rates of population change for England between 1540 and 1870, estimated by Wrigley et al. (1997), who call it the compound annual growth rate. In the regression model where this quantity is used as a response variable we use *non-delayed* values of potential predictor variables—the approach is essentially based on Model (3). The results obtained with the analysis of per capita rates of population change, thus, provide an important testing case of the general approach. Unfortunately, situations where we have an independent estimate of the rate of change are relatively rare. For example, there is no comparable variable corresponding to the rate of change of the instability index.

Steps in the analysis

Once we selected the value (or values) of the time delay, the actual analysis of data is fairly straightforward, and can be accomplished by using any statistical software. The first model that should be tried is the linear regression:

$$X_t = a_0 + a_1 X_{t-\tau} + a_2 Y_{t-\tau} + \varepsilon_t$$

where a_0, a_1, etc are regression coefficients, and ε_t is the error term—a stochastic variable assumed to be normally distributed with mean zero and variance σ^2. Note that the subscripts indicating time have been shifted. This is a convention in time-series analysis—the idea is that we are trying to explain the present given the past (rather than predict the future given the present, which is appropriate in a theoretical model). A viable alternative is to use ΔX_t as the response variable. It also needs to be redefined by shifting subscripts: $\Delta X_t = X_t - X_{t-\tau}$. The two models are very similar and differ in the value of the estimated coefficient a_1.

Many of the issues discussed in standard textbooks on regression apply to fitting statistical models in the time-series data (one important exception, discussed below, is that standard statistical tests cannot be used due to autocorrelations between values of the response variable). The first concern is whether the assumptions of normally distributed residuals hold at least approximately. Often non-Gaussian residuals can be handled by a judicious transformation of the response variable (for example, population numbers should be routinely log-transformed). The second concern is whether the relationship between the predictor and response variables is nonlinear. Mild forms of nonlinearities can be fitted by appropriately transforming the predictor variables (not the response, because that affects the structure of residuals). Strong nonlinearities (e.g., humped functions) are fitted by employing polynomials. For further discussion of these issues, see Turchin (2003a).

The usual statistics printed out by the canned software, such as *F*-ratios and *P*-values are suspect because they are calculated by assuming statistical independence of response variable

values. In the time-series setting, however, sequential values tend to be positively autocorrelated, which tends to inflate the significance of the results (if sequential predictor values are negatively autocorrelated, then the standard statistics, on the contrary, understate the significance of the results). This leads to the confidence intervals on the estimates of the model parameters that are too narrow. Thus, obtaining an unbiased measure of confidence intervals takes some work. The most general approach is the bootstrap (Efron and Tibshirani 1993).

Typically, however, we are not interested in the regression coefficients *per se*, but in determining whether a particular predictor variable, or set of variables, has a statistically detectable effect on the response. In other words, in the model such as

$$\Delta X_t = a_0 + a_1 X_{t-\tau} + a_2 Y_{t-\tau} + a_3 Z_{t-\tau} + ... + \varepsilon_t$$

where X, Y, Z, etc are potential predictor variables, which of them we need, and which we can dispense with? This is the issue of *model selection*. In the last decade or two, statisticians and time-series analysts advanced very substantially the methodology of model selection, so we have a pretty good understanding of what to do, and what not to do. The basic premise of the approach is to select only those variables that increase our ability to predict the values of the response variable (that's why they are called predictors).

Simply using a statistic such as R^2 does not work, because the more variables we include on the right hand side of the regression equation, the better R^2 we get. This procedure results in an overfitted model that, when presented with novel data, does very poorly at predicting the values of the response variables. In other words, an overfitted model does very well in "explaining" in-sample data, on which it was fitted, but poorly on predicting out-sample, novel data. One approach to resolving this problem is to use some information measure, such as the Akaike Information Criterion, AIC (see, for example, Burnham and Anderson 1998). AIC balances how well a model fits the data against how many parameters were needed to achieve this degree of fit. Many standard packages now print out AIC or some other related measure; if not, AIC can be calculated from standard kinds of output. Using AIC is reasonably straightforward, and many people now employ it in model selection.

Although the use of information criteria is perfectly valid, personally I prefer cross-validation. The disadvantage of cross-validation is that it is very computer intensive, but given the power of present-day computers this is not much of a problem. The advantage, on the other hand is that cross-validation gives us a direct method of finding out how well different models predict out-sample data.

The basic approach is simple. We split the data into two halves, fit all different models on the first half, and use the fitted models to predict the response variable in the second half. Then we reverse the procedure: fit the models on the second half, predict the first. This procedure is known as the "double-cross". One problem with it is that although we utilize all the data for testing purposes, at any given step we use only half the data points for fitting purposes. When data are scarce, half of them may not be enough to fit more complex models, even though the use of such complex models may be warranted. The solution is known as the "*k*-fold cross-validation". Instead of dividing the dataset into two halves, we divide it in k parts. We reserve one of the parts for testing, fit the model on the other k–1 parts, test it on the reserved one, and then repeat it for all k parts. In the extreme, we can set $k = n$, the sample size. In such a case, at every step we fit the model on all but one data point (the left-out "predictee"—the point to be predicted), while testing it on all data points. As Mosteller and Tukey (1968), who first clearly

formulated this approach, commented, this procedure "squeezes the data almost dry." However, in the time-series setting we cannot use this procedure as is, because of the redundancy problem. We have to exclude several points on each side of the predictee. Otherwise, it would be too easy to predict the omitted point X_t as an average of X_{t-1} and X_{t+1}. How many points we omit on each side depends on how high is the serial autocorrelation.

This brings me to another important issue—oversampling. An oversampled series is one where sequential values of X are highly autocorrelated. As an example, think of a trajectory showing one "cycle", or more precisely, a single excursion up and down in the phase space. Suppose that the trajectory was sampled at 100 points along the way. Is our sample size, $n = 100$? Not at all, because each subsequent data point is so similar to the preceding one that it adds very little information overall to the data set. We could easily reduce the sampling frequency to just 10 points and not lose any information at all.

As a result, in time-series setting we should quantify our sample sizes as the number of oscillations (up-and-down excursions) in the series. This should not be taken too literally, because each oscillation is a pretty "fat" data point. In a larger sense, however, our confidence in the results can grow only when we analyze more and more oscillations, which can come from the same place (for example, the English data set has a couple of oscillations) or from other places (the Chinese data add two more oscillations for the Han period and two for the Tang period).

Interpretation of the results

The analysis approach that I have been describing above is obviously more involved than a regular regression exercise, but it has one advantage over regular regression. In usual regression situations all we can say is that two variables are closely correlated, we can never tell which one is the cause, and which one the effect. In dynamical analysis a strong correlation between a rate of change and structural variables usually means that the variables themselves are the cause, and the rate of change is the effect. Most graphically this is seen in the discrete setting: since X_t, Y_t and so on precede X_{t+1}, the direction of the causation arrow is pretty obvious. This consideration is a great help in interpreting analytical results, but there are, nevertheless, limitations. For example, if variable Y has a strong effect on the rate of change of X, it is not necessarily true that there is a causal mechanism underlying this relationship. For example, Y could be closely correlated with Z, which is the actual mechanism of change in X, so that the relationship detected by the analysis is, in this way, a spurious one. In general, no statistical analysis can unambiguously identify causal mechanisms. Interpretation of the results of dynamical analysis is aided by the temporal structure of the data set, but gaining understanding still requires erecting and rejecting hypotheses.

Interpretation of regression results is greatly aided by first understanding how different variables in the data set fluctuate with respect to each other. The simplest pattern is two variables moving up and down in synchrony, without a phase shift or, alternatively, in perfect anti-phase. Both types of behavior have the same dynamical implications (Turchin 2003c). One possible explanation of such dynamics is that one of the variables operates on a fast time scale and fluctuates in response to changes in the second variable. This is the most likely explanation of the almost perfect anti-phase movements between population pressure and real wage in the English data set (see Figure 2b in Turchin 2005). Economic processes setting the real wage operate on a much faster time scale than the population movements. When population grows or declines, real wages quickly equilibrate in response to new economic conditions. As a result, real wage curve follows population trajectory without any perceptible lag.

Synchronous oscillations of two variables can also result from a different mechanism: entraining or phase-locking (Turchin and Hall 2003). Mathematical models show that if two (or more) systems separated in space are driven by largely endogenous dynamics, and if their endogenous dynamics are broadly similar (e.g., have approximately the same period), then their cycles may be synchronized by a variety of shared exogenous perturbations. This insight, in particular, is suggestive for possible explanations for the synchrony of population and city-size changes in east and west Asia. Entraining usually results in less rigid synchrony than when a slow variable has a direct effect on a fast one. Therefore, we expect to see the two entrained variables come into and out of phase. However, because data are typically limited in quantity (we rarely can follow the system for more than two or three oscillations) and quality (measurement errors are often high) in practice it is hard to distinguish between the two potential explanations of synchrony using just the time-series data.

When two variables fluctuate synchronously or in perfect anti-phase, dynamical systems theory tells us that the interaction between the two variables is *not* what drives the oscillations (Turchin 2003c). A phase shift of roughly one-quarter between two variables, on the other hand, is consistent with the hypothesis that it is the dynamical interaction between the two variables that drives the cycle (it is also possible that one or both of the variables are closely correlated with the actual drivers of oscillations). When dynamics are relatively simple low-dimensional cycles, graphical analysis in the phase plot could be very revealing (e.g. Figure 4b in the accompanying paper). However, complex trajectory movements in a high-dimensional phase plot are difficult to visualize, and regression analysis of multiple predictor variables comes to the fore.

Based on the phase relations and the results of regressions we can classify variables in the data set into four classes. Let X be the primary variable whose dynamics we are mainly interested in understanding. For example, in the accompanying paper (Turchin 2005) the primary variable is population numbers and what needs to be explained is population oscillations. In what follows, by the "rate of change of X" I mean either an estimate of its derivative, or ΔX_t, depending on whether we are using the continuous or discrete analysis framework.

The first class includes those variables that move together with X synchronously or in anti-phase. An example of such a *first-order endogenous variable* is the real wage in the English data.

The second class includes those variables that are closely correlated with the rate of change of X, and whose own rate of change is closely correlated with X. Typically, their fluctuations will be shifted by a quarter phase with respect to X, but if dynamics of the system are very complex and high-dimensional, then such phase-relations may be hard to see. An example of such *second-order endogenous variables* is the sociopolitical instability.

The third class includes variables that affect the rate of change of X, but who themselves fluctuate independently either of X or of other endogenous variables. A possible example of such an *exogenous variable* might be climatic fluctuations, if it turns out that they have an effect on the carrying capacity of the environment. The final class of *irrelevant variables* includes those that have no effect on either X or its rate of change.

In real life it will not always be easy to sort out all variables neatly into one or another class, but let us think through the implications of different kinds of variables for the explanation of the dynamics of X. Second-order endogenous factors are of most interest because it is their action that drives oscillations up and down. First-order factors set the limits of fluctuations, and are inherently stabilizing. The exogenous factors are responsible for unpredictable stochastic

fluctuations (unless they themselves have a strong periodic component). The irrelevant factors are, well, irrelevant. A rough estimation of the relative importance of various factors is given by the proportion of variance that each of them explains in the regression with the rate of change of *X* as the response variable.

In summary, the statistical approach that I reviewed above can yield valuable insights into the feedback structure characterizing the interactions between different aspects of the studied dynamical system—when data are reasonably plentiful (for example, at least two or three complete oscillations), cover different aspects of the system, and the measurement errors are not too large. Once we have determined which variables have the strongest effect on each other (or, rather, on each other's rates of change), we may desire to obtain quantitative estimates of parameters that govern the strength and functional form of various feedback loops. However, the generic models discussed in this primer are not suitable for this purpose. A better approach is to write models based on explicit mechanisms and fit them to data using such approaches as trajectory matching or nonlinear forecasting. Fitting mechanistic models to time-series data is a large topic, and its discussion is beyond the scope of the present primer. The interested reader may consult Chapter 7 of Turchin (2003a), where I discuss such approaches in the context of population ecology.

Literature cited

Burnham, K. P., and D. R. Anderson. 1998. *Model Selection and Inference: A Practical Information-Theoretic Approach.* Springer, New York.

Casdagli, M., S. Eubank, J. D. Farmer, and J. Gibson. 1991. State space reconstruction in the presence of noise. *Physica D* **51**:52-98.

Efron, B., and R. J. Tibshirani. 1993. *An introduction to the bootstrap.* Chapman and Hall, New York.

Ellner, S., and P. Turchin. 1995. Chaos in a noisy world: new methods and evidence from time series analysis. *American Naturalist* **145**:343-375.

Mosteller, F., and J. F. Tukey. 1968. Data analysis, including statistics. *in* G. Lindzey and E. Aronson, editors. *Handbook of Social Psychology*, Vol. II. Addison-Wesley, Reading, MA.

Turchin, P. 2003a. *Complex Population Dynamics: A Theoretical/Empirical Synthesis.* Princeton University Press, Princeton, NJ.

Turchin, P. 2003b. *Historical dynamics: why states rise and fall.* Princeton University Press, Princeton, NJ.

Turchin, P. 2003c. Evolution in population dynamics. *Nature* **424**:257-258.

Turchin, P. 2005. Dynamical feedbacks between population growth and sociopolitical instability in agrarian states. *Structure and Dynamics* **1**(1): in press.

Turchin, P., and T. D. Hall. 2003. Spatial synchrony among and within world-systems: insights from theoretical ecology. *Journal of World Systems Research* **9**:37-64.

Wrigley, E. A., R. S. Davis, J. E. Oeppen, and R. S. Schofield. 1997. *English population history from family reconstruction: 1580-1837.* Cambridge University Press, Cambridge, UK.