Cliodynamics: The Journal of Quantitative History and Cultural Evolution

Fitting Dynamic Regression Models to Seshat Data

Peter Turchin
University of Connecticut

Abstract

This article presents a general statistical approach suitable for the analysis of time-resolved (time-series) cross-cultural data. The goal is to test theories about the evolutionary processes that generate cultural change. This approach allows us to investigate the effects of predictor variables (proxying for theory-suggested mechanisms), while controlling for spatial diffusion and autocorrelations due to shared cultural history (known as Galton's Problem). It also fits autoregressive terms to account for serial correlations in the data and tests for nonlinear effects. I illustrate these ideas and methods with an analysis of processes that may influence the evolution of one component of social complexity, information systems, using the *Seshat: Global History Databank*.

General Introduction

Ten thousand years ago all humans lived in egalitarian small-scale societies of some hundreds, or perhaps a few thousands of individuals. Today we live in huge societies with millions of members (the average population of a member of the United Nations is 30 million), with extensive inequalities, intricate division of labor, and elaborate government structures. Understanding how and why this "major evolutionary transition" (Maynard Smith and Szathmáry 1995) occurred is of huge intellectual and practical importance (Turchin 2016).

Recent analysis of social complexity[1] data in the Seshat Databank revealed that characteristics, such as social scale, economy, features of governance, and information systems, show strong evolutionary relationships with each other and that complexity of a society across different world regions can be meaningfully measured using a single principal component of variation (Turchin et al. 2018). Although different world regions began the transition to complex societies at different times, and the pace of the evolution of social complexity was highly variable, our results suggest that key aspects of social organization are

[1] I explain "social complexity" later in the article.

Corresponding author's e-mail: peter.turchin@uconn.edu

Citation: Turchin, Peter. 2018. Fitting Dynamic Regression Models to Seshat Data. *Cliodynamics* 9: 25–58.

functionally related and tend to coevolve in predictable ways. The next step, thus, is to understand general principles that guided these evolutionary trajectories.

We don't lack explanatory theories (e.g., Fukuyama 2011; Johnson and Earle 2000; Trigger 2003; Turchin 2016); the main issue is how we test them empirically. Because it is not feasible (or even ethical) to subject large-scale societies to experiments, the main avenue of empirical testing must rely on historical and cross-cultural analyses. Such approaches have already yielded many insights into the evolution of human societies. As an example, one of the most fruitful data compilations in anthropology has been the Standard Cross-Cultural Sample (SCCS), which codes the characteristics of 186 well-described ethnographic societies (Murdock and White 1969). Every year, between 70 and 80 articles are published that use data in the SCCS to test a variety of anthropological theories (Turchin et al. 2012). Overall, more than 1200 analyses of the SCCS data were published since its introduction in 1969.

There are, however, serious limitations of the SCCS and similar repositories of cultural information that restrict its application in testing theories about the evolution of complex societies. One problem is that the SCCS focuses on small-scale societies. As a result, large-scale societies are seriously undersampled. More important, the SCCS is a synchronic or static database: it codes the characteristics of any particular society at a single point in time. Because the SCCS does not tell us how societies change with time, its data are less suitable to testing evolutionary theories. After all, sociocultural evolution is all about change. A similar problem afflicts another popular, and related, anthropological database, the Ethnographic Atlas (Murdock 1967).

As I discuss in the next section, the temporal dimension greatly enhances the ability of statistical analysis to make inferences about causal mechanisms in cultural evolution. Investigators have circumvented the limitation of classical ethnographic databases, which lack such a temporal dimension, in a variety of ways. For example, one fruitful avenue has been to use the methods of phylogenetic analysis developed in evolutionary biology (Currie and Mace 2012; Mace and Holden 2005; Watts et al. 2015). In the "deep roots" literature (Spolaore and Wacziarg 2013), which attempts to discern long-term processes affecting economic development, researchers supplement easily obtainable data on modern societies (e.g., the Gross Domestic Product per capita, GDPpc) with historical "snapshots" taken at irregular time intervals. For example, Comin et al. (2010) assembled a dataset on technology adoption in 1000 BCE, 0 CE, and 1500 CE for the predecessors of today's nation states and regressed GDPpc in 2002 on these measures of technological sophistication in the past. Peregrine's (2003) *Atlas of Cultural Evolution* and the related *Encyclopedia of Prehistory* (Peregrine

and Ember 2001) more systematically captures the time dimension, but at a relatively coarse scale with a time step of one thousand years. Most recently, the Seshat project has started publishing historical and archaeological data on a global sample of past societies spanning the time period between the Neolithic and Industrial Revolutions (Turchin et al. 2018). These data are resolved at 100-year intervals, enabling us to fit dynamic regression models and, thus, greatly increasing our statistical power to empirically test predictions from rival theories.

An Approach to the Analysis of Time-Resolved Cross-Cultural Data

A huge advantage of dynamic (time-series) data is that it enhances our ability to investigate the evolutionary processes that generate cultural *change*. When we observe a high degree of correlation between two variables in a synchronic data set, it is not possible to determine which of the factors is the cause, and which is the effect. This problem is sometimes referred to as the problem of "endogeneity" or simultaneous causality. Recent developments in statistical methods of causal inference (Pearl 2009) hold the promise of circumventing this problem by analyzing simultaneously a large number of variables representing potential causes (for an example in cross-cultural analysis, see Eff and Dow 2009).

Analyzing time-series data, on the other hand, enables us to capitalize on the general observation that causes tend to precede effects in time (Suppes 1970). In econometric literature this principle is known as Granger causality (Granger 1969). There are certain limitations to this idea (a crowing rooster does not cause the sun to rise), and I deal with such caveats in the next section. However, it is clear that time-resolved data, at the very least, provide us with a more informative data set to test dynamical theories.

Consider the following dynamic model:

$$X_{t+1} = f(X_t, Y_t, U_t)$$

$$Y_{t+1} = g(X_t, Y_t, U_t)$$

Here X_t and Y_t are two variables that could be dynamically inter-related. For example, X_t could be population density at time t, and Y_t is warfare intensity, measured by death rates resulting from war (Turchin and Korotayev 2006). U_t summarizes the influences of *exogenous* variables (also known as dynamical noise), which affect the change of X_t and Y_t, but are not themselves affected by the main variables. The generative functions f and g specify how future values of X and Y (at time t + 1) are determined by joint effects of all variables at present (time t). If we know the generative functions, initial conditions, and how U_t

changes with time (for example, we can model it as a stochastic process), then we can iterate this model forward in time to study what dynamics it predicts.

In data analysis we have an inverse problem: we know the trajectories of X_t and Y_t, but wish to make some inferences about the generative functions. In particular, we wish to know whether X_t and Y_t affect the time evolution of each other. For simplicity, I will assume that we will use the following linear regression model to answer this question:

$$X_{t+1} = a + bX_t + cY_t + U_t$$

$$Y_{t+1} = d + eX_t + hY_t + U_t$$

where a, b, … h are regression coefficients and U_t is the error term. However, it is not necessary to make the linearity assumption (see Turchin 2003) and later in this article I will be fitting nonlinear dynamic models to data.

There are four possible outcomes that such analysis could yield.

- If neither c nor e are significantly different from zero, we conclude that X_t and Y_t evolve independently of each other.
- If c is significantly different from zero, but e is not, we have the situation of one-directional causation: Y affects the evolution of X, but is not itself affected by X.
- If e is significantly different from zero, but c is not, we again have the situation of one-directional causation, but now it is X that is a cause for Y.
- Finally, if both regression coefficients are significantly different from zero, then X and Y co-evolve as a dynamic complex (this is sometimes known as "circular causation," which is different from the problem of simultaneous causality).

A famous example of dynamics characterized by circular causation is the interaction between predators and prey, which is known to have the capacity to generate population cycles (Lotka 1925; Volterra 1926). Another example, involving the interaction between population dynamics and warfare, is discussed by Turchin and Korotayev (2006). In that article we also demonstrate that in a dynamical system with circular causation it is possible to observe no correlation between X_t and Y_t, when these variables are measured at the same time. This paradoxical result (X affects Y's rate of change and Y affects X's rate of change, but X and Y are not correlated) is an important reason to fit dynamic models to time-series data (or, if such data are unavailable, build dynamic databases).

Complicating Factors

Time-series data, thus, offer us a possibility of resolving causal relationships between variables. However, in real-life applications there are many factors that could defeat our ability to detect cause-effect arrows. One fundamental difficulty is that although our analysis might implicate *X* as a causal factor for *Y*, in reality the true cause could be *Z*, with which *X* is closely correlated. This is known as the "hidden variable" problem or omitted variable bias. Failure to address this problem can result in spurious results. As an example, in his analysis of SCCS data Nolan (2003) found a positive relationship between high population densities and incidence of warfare. Yet a reanalysis of the same data by Eff and Routon (Eff and Routon 2012) indicated a reverse relationship: high population densities lead to *less* war. Eff and Routon show that the previous result was due to omitted variable bias (omitting variables from the analysis that are highly correlated with high population densities, such as metal-working and writing systems, biasing their results).

A somewhat less obvious problem is an "uninformative dataset," which could result when the observed trajectories sample a small subset of the phase space in which dynamics could potentially evolve. For example, if the relationship between *Y* and *X*'s rate of change is ∩-shaped, and our dataset contains observations only for the middle range of *Y*s, where the relationship is nearly flat, our analysis will incorrectly conclude that *Y* does not influence change in *X*. Such potential complications need to be considered during the design stage of database building. Generally speaking, one should aim to include as much variability in as many different variables as possible. An additional reason for uninformative datasets are noise, especially measurement errors (on which more below).

Another well-known complication in the analysis of cross-cultural data is Galton's problem (Eff and Dow 2009; White et al. 2011). A fundamental assumption of regression analysis is that analytical units are statistically independent of each other. However, historical societies are not independent: they interact with each other by trade, conquest, and the exchange of cultural information. As a result, two different societies may possess the same trait not because it evolved independently in each, but because of borrowing, conquest, or inheritance from a common ancestor. Our statistical methods need to take such processes into account.

Finally, our knowledge about past societies is usually imprecise and often incomplete. Furthermore, even professional academic historians often disagree about certain aspects of the societies they study. This means that our statistical methods must be able to deal with missing data, uncertainty, and disagreement

between the experts. Currently the preferred method for dealing with these issues uses the technique of Multiple Imputation (Eff and Dow 2009).

The Evolution of Information Systems

This article will illustrate these ideas and methods of analysis by focusing on processes that may influence the evolution of one component of social complexity, information systems, using the *Seshat: Global History Databank* (Turchin et al. 2015). There is a long tradition in anthropology that identifies writing, literacy, and specialized knowledge as some of the key features that define civilization (Childe 1950; Goody 1986; Trigger 2003; van der Leeuw 1981). Trigger (2003) identified three different functions that early information systems served: state propaganda (commemorating kings and their deeds), administration, and religion (including divination and sacred texts). Earlier Goody (1986) explored a list of four functions of writing: religion, economy, state administration, and law.

Because the main purpose of this article is not a thorough test of evolutionary hypotheses explaining the rise and development of information systems, but presenting the dynamic regression methodology (with information systems serving as an illustration), in the following I will focus on a limited set of hypotheses. The main theory that I will investigate is that the evolution of information systems was driven by administrative needs, and I will use the Seshat measure of specialized governance as a proxy for this explanation. A subsidiary hypothesis suggests that writing would be particularly useful in territorially extensive polities, which required an efficient and accurate delivery of information between distant regional capitals and the central government. This hypothesis will be proxied by the extent of polity territory as a covariate in the regressions. Finally, I will investigate the importance of economic functions of writing by using the Seshat variable Money, which reflects the sophistication of monetary instruments.

I will also include in the analysis all other Seshat variables that reflect various characteristics of social complexity. Including as many covariates as are available helps to reduce the omitted variable bias, as was explained in the previous section. Thus, the plan is to fit dynamic regression models with *Info* (the sophistication of information systems) as the response variable and other components of social complexity as predictor variables. I will use the dataset that was recently published by Turchin et al. (2018).

The rest of the article is organized as follows. I begin with a brief introduction to the goals, structure, and methodology employed by the Seshat project. Next, I describe the methodology, based on Multiple Imputation, for dealing with missing

data, uncertainty, and expert disagreement. My goal is to gather all information needed to understand the Seshat data and analyses in this methodology article; to do this I borrow some text from the Supplementary Online Information of (Turchin et al. 2018). Finally, I fit a series of dynamic regression models, including nonlinear versions, to the data with the goal of detecting the causal factors influencing the evolution of information systems.

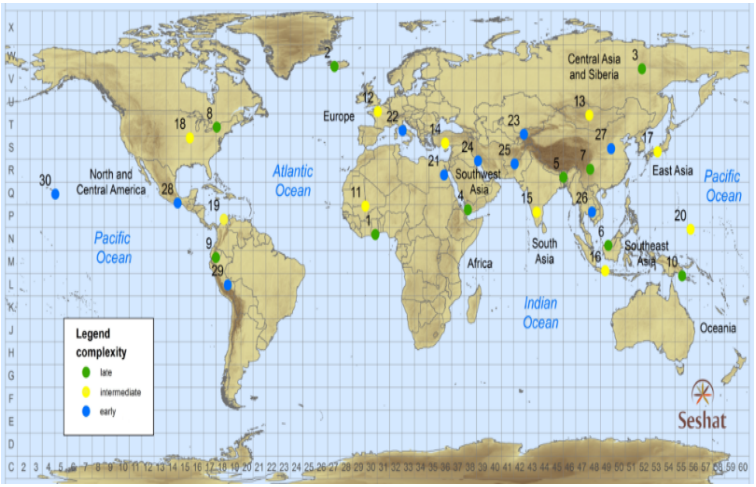
A Quick Introduction to Seshat: Global History Databank

Founded in 2011, *Seshat: Global History Databank* systematically collects what is currently known about the social and political organization of human societies and how they have evolved over time (François et al. 2016; Turchin et al. 2015). The overall goal of Seshat is to enable researchers to conduct comparative analyses of human societies and rigorously test different hypotheses about the social and cultural evolution of societies across the globe, spanning long periods of human history.

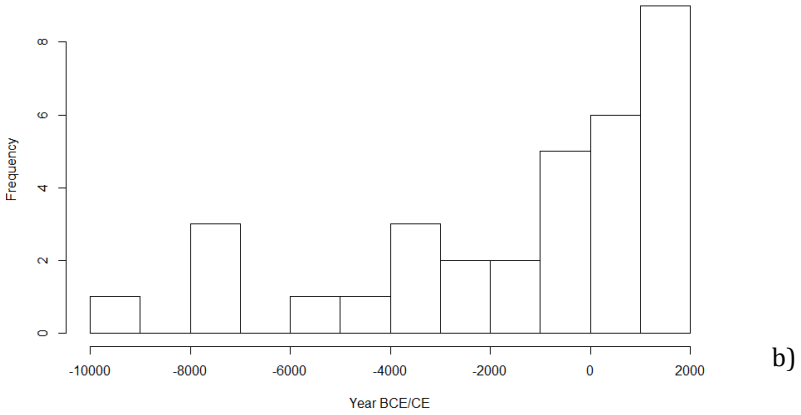
Temporal and Geographic Scope

Currently Seshat focuses on the time period between the Neolithic and Industrial Revolutions. The spatial reach is global, and eventually we plan to include in the Databank information on all past societies, up to the present, for which historical or archaeological data are available. However, reaching this goal will take time. As a first step, we collected data on a sample of 30 locations across the globe, stratified by the world region and the antiquity of complex societies (Turchin et al. 2018). We are now in the process of expanding the global coverage, and the results below are based on 32 locations. For each of the 32 global points we start at a period just before the Industrial Revolution (typically, 1800 or 1900 CE depending on the location) and go back in time to the Neolithic (subject to the limitation of data).

Our unit of analysis is a *polity*, an independent political unit that ranges in scale from villages (independent local communities) through simple and complex chiefdoms to states and empires. For each polity we code variables on social complexity, warfare, religion and rituals, agriculture and resources, institutions, well-being, and the production of public goods. Overall, the current codebook includes over 1500 variables. These variables are coded for any past polity that occupied one of our 30 world locations between the Neolithic and Industrial Revolutions. Currently there are over 400 such polities in Seshat. As of December 2017, the Databank contains >200,000 coded values ("Seshat records", see below). In this article, however, I will focus only on Social Complexity variables (Turchin et al. 2017).



a)



b)

Figure 1. (a) Locations of Natural Geographic Areas (NGAs) that sample global variation in cultural evolution. For the current list of NGAs see our Methods page. (b) Frequency distribution of the starting dates for data sequences in Seshat Databank. For "late complexity" NGAs data series are short, often starting only when European explorers reached the area in the eighteenth or nineteenth century. For "early complexity" locations data sequences extend back in time between 4,000 and 10,000 years ago. "Intermediate complexity" cases are usually located between these two extremes.

In order to assess whether different societies show commonalities in the way they have evolved we developed a geo-temporal, stratified sampling scheme that aimed (1) to include as much variation among the sampled societies as possible in terms of social organization and (2) to ensure representation of different parts of the world. This issue is challenging as societies can expand or contract in geographical space, appear or disappear in the historical and archaeological records, and show varying degrees of continuity with earlier or later societies.

To overcome these issues and ensure that we collected data in a systematic manner we divided the world into ten major regions (Figure 1a). Within each region we selected three natural geographic areas (NGAs), our basic geographical sampling units. Each NGA is defined spatially by a boundary drawn on the world map that encloses an area delimited by naturally occurring geographical features (for example, river basins, coastal plains, valleys, and islands). The extent of the NGAs does not change over time, and NGAs thus act as our fixed points which determine which societies we collected data for. The data themselves, however, are collected not for an NGA, but for the entire society, or polity, that happened to occupy the NGA at a given time. Each NGA, then, serves as geographic "anchor" from which we generate a list of all the polities that occupied it over the course of history. Such a sampling approach allows us to be consistent and methodical about designating societies for which we gather data. It also allows us to construct spatially anchored time-series, as long as it is understood that the spatial extent of sampled societies fluctuates with time (as polities rise, expand, go into decline, and collapse).

Within each world region we looked for NGAs that would allow us to cover as wide a range of forms of social organization as possible. In effect we wanted to ensure that we captured information about the kinds of societies that researchers have previously discussed in relation to social complexity ("states", "chiefdoms", "stratified societies", "empires", etc.) without using typological definitions of such societies or employing a strong, limiting definition about what features such societies should have. We also wanted to make sure that we captured information about societies that are not traditionally thought of as complex ("small scale societies", "egalitarian tribes", "acephalous societies").

Accordingly, within each world region one NGA was selected that saw the earliest developments of a centralized, stratified society. We also chose a second NGA that was the opposite; ideally, it was free of centralized societies until the colonial period. Finally, the third NGA was intermediate in terms of the time when political centralization emerged within the world region. Because different world regions acquired centralized societies at different times there can be substantial variation across "early complexity" NGAs both in the time at which

our measures of social complexity start increasing and the degree of social complexity that is eventually reached at the end of our sampling period. For example, Susiana, the early complexity NGA in Southwest Asia has a much longer history of large societies than Hawaii, the early complexity NGA in the Pacific region. The distribution of starting dates for all NGAs currently in Seshat is shown in Figure 1b.

More recently (in 2017) we began expanding data coverage beyond the original sample of 30 NGAs. Currently, we have good data on two additional NGAs, and I will include these data in the analyses reported below.

Data Collection

To populate the Databank, for each NGA we consult the literature and chronologically list all polities that were located in the NGA, or encompassed it. We chose a temporal sampling rate of one hundred years, and we only included polities that span a century mark (for example, 300 CE, 400 CE, 500 CE, and so on) while omitting any polities of short duration that only inhabited an NGA between these points. One century is short enough to capture meaningful changes in the social complexity of historical societies, but not too short to lead to *oversampled* data ("oversampling" results when the succeeding point in time contains the same information as the preceding one, thus not adding to the overall information content of the data set).

For those periods when the NGA is divided up among a multitude of small-scale polities (e.g., independent villages, or small chiefdoms) it is not feasible to code each individual polity. In such instances we use the concept of "quasi-polity," which is defined as a geographic area with some degree of cultural homogeneity that is distinct from surrounding areas and approximately corresponds to an ethnological "culture" (Murdock 1967; Murdock and White 1969) or an archaeological sub-tradition (Peregrine 2003). We then collect data for each quasi-polity as a whole. This way we can integrate over (often patchy) data from different sites and different polities within the NGA to estimate what a "generic" polity was like. Such an approach is especially useful for societies known only archaeologically, for which we usually don't know polity boundaries.

It is important to point out that our use of polities and quasi-polities is best understood as a means of sampling the vast literature on past human societies rather than trying to impose a rigid framework on the human past. Our data coding procedures enable us to capture changes in a particular variable within the lifetime of a polity and also allow us to capture variation within a polity or quasi-polity where there is such evidence. We also allow a gradual emergence or disappearance of a polity, as when an empire slowly disintegrates and its

constituent pieces gain an increasing degree of independence from the old imperial master. Finally, we are able to flexibly incorporate multiple lines of evidence and uncertainty as we outline below.

When gathering data into Seshat, our approach is to avoid forcing information about a past society into an arbitrary scale (e.g., "rate the social complexity of this society on a scale from 0 to 10"). Instead, and prior to collecting the data, we run a workshop that develops a conceptual scheme for the particular aspect that we aim to capture in Seshat. Generally speaking, we aim to use either a quantitative variable (e.g., an estimate of the population of the coded polity) or break up complex variables into multiple simple variables that can be coded in a binary fashion (absent/present). The initial coding scheme is then tested by Seshat research assistants (RAs) applying it to several test cases, in consultation with experts (archaeologists or historians who study the coded polities). The coding scheme is then refined based on the suggestions from both experts and RAs and is applied to the whole sample.

Once a coding scheme is defined, data collection occurs in several phases. First, RAs search published articles and books on a particular polity (with advice from a regional or polity expert on what sources are likely to be most useful) in order to find information about each variable and enter it into the databank. Second, RAs compile lists of questions on values that cannot be coded unambiguously, or on which information in the published sources is lacking, and seek help from the experts on the polity. In the final phase we ask experts to go over the data to check coding decisions made by RAs and help us fill any remaining gaps. Experts also indicate when the value should be coded as "unknown" (RAs may use the code "suspected unknown," but only experts can definitively state that something is indeed "unknown").

When two or more experts disagree about the value or there is ongoing debate in the literature, all choices are entered as alternatives. For quantitative variables whose values are known only approximately, coders are instructed to enter a likely range [min, max] that roughly corresponds to a 90 percent confidence interval (i.e., omitting possible, but unlikely or unrepresentative values).

We refer to a coded value of a particular variable for a particular polity as a "Seshat record." Seshat records have complex internal structure. First, there is the value of the coded variable. For a numerical variable the value can be either a point estimate, or a range approximating the 90-percent confidence interval. Binary variables can take the following values: present, absent, inferred present, inferred absent, and unknown (a numerical variable can also be coded as unknown). "Inferred" presence or absence indicates some degree of uncertainty:

when direct evidence of presence (for example) is lacking, but the expert can confidently infer it. For example, if iron smelting has been attested both for the period preceding the one that is coded, and for the subsequent period, we code it as "inferred present" even though there is no direct evidence for it (assuming there are no indications that this technology was lost and then regained).

Binary variables can also have temporal uncertainty associated with them. For example, if we know that iron smelting appeared in a particular polity at some point between 300 and 600 CE, we code period previous to 300 CE as absent, the period following 600 CE as present, and the period between 300 and 600 CE as effectively "either absent, or present" (this is different from "unknown").

The second important part of a Seshat record is a narrative paragraph explaining why this particular variable was coded in this particular way. Typically, this narrative is first written by an RA, who may quote the relevant text from a reference (a book or an article) of from a personal communication by an expert. The narrative is then checked and edited by experts. Subsequent experts can add to it and disagree with previously recorded estimates.

The third part of a Seshat record is the references to publications or other databases. Reference can also be a "personal communication" from an expert or from several experts participating in a Seshat workshop.

Aggregation of Social Complexity Data into "Complexity Components"

As the preceding discussion shows, during the data collection stage complex variables are broken down into simpler components and data are gathered for each component. Before analysis we assemble simpler (often, binary) variables into more quantitative measures that are more suitable for statistical analysis. I illustrate the process with Information Complexity (*Info* for short; variables are italicized for readability).

Info is based on 15 binary Seshat variables. The first four provide the basis for measuring the sophistication of the writing system (following the description of the variables in the Seshat codebook):

> **Mnemonic devices** such as tallies
>
> **Non-written records**, which are more extensive than mnemonics, e.g., quipu
>
> **Script** as indicated at least by fragmentary inscriptions (note that if written records are present, then so is script)
>
> **Written records**: these are more than short and fragmentary inscriptions (such as those found on tombs or runic stones). There must be at least several sentences strung together. For example, royal

proclamations from Mesopotamia and Egypt, which can be quite lengthy, qualify as written records

These binary measures are combined to produce a *writing* scale from 0 to 4:

0 = no evidence of writing for the coded polity
1 = only evidence of mnemonic devices is present
2 = non-written records are present, but no script
3 = evidence for script in fragmentary inscriptions, but no lengthy texts
4 = written records are present

Presence or absence of a "less sophisticated" writing variable doesn't affect this scale (so if "script" is present, it does not matter wither non-written records are present or absent).

The next two Seshat variables code for whether the writing system is phonetic (e.g., alphabetic) or non-phonetic (e.g., ideographic). They are not used in the scale for *Info*, as they only code for the type of script.

The next nine binary variables code for presence or absence of various kinds of *texts*:

- Lists, tables, and classifications, as used in trade
- Calendar
- **Sacred Texts**, which originate from supernatural agents (deities), or are directly inspired by them
- **Religious literature**, which differs from the sacred texts in that it provides commentary on the sacred texts, or advice on how to live a virtuous life
- **Practical literature**: for example, manuals on agriculture, military, cooking, etc.
- History
- Philosophy
- **Scientific literature** including mathematics, natural sciences, social sciences
- **Fiction** including poetry (but must be written down)

The variable *texts*, which scales from 0 to 9, sums the number of securely attested types of texts (that is, coded as "present"). The idea here is that the more sophisticated a society is informationally, the more different types of texts it will have in circulation. For this reason, we only count those types of texts that were definitely in circulation and left clear evidence of their use (in other words, "absent", "unknown", or even "inferred present" do not constitute such strong evidence of presence). Finally, to construct *Info* we simply sum *writing* and *texts* scores. Thus, *Info* can vary between 0 and 13.

It is important to note that the above scheme is only one of the possible ways to come up with a quantitative measure of information sophistication. Other analysts are free to combine and recombine Seshat variables in different ways. One of our goals, when designing the conceptual approach used in Seshat, was to separate data coding and data analysis steps as much as possible, providing analysts with freedom to define entities of interest that are most suitable to their analysis goals.

Furthermore, the current set of variables is optimized to capture the evolution of information complexity in early agricultural societies, and *Info* tends to max out by the time we get to the Middle Ages, or even earlier for some world regions (for example, Roman Empire during the Principate scores the maximum on the *Info* scale). Additional variables are needed to adequately score the informational complexity of modern societies. These could be an elaboration on the current scheme (e.g., splitting "history" into "world history", "military history", "biography", "economic history", "intellectual history", etc.), or adding an entirely different set (for example, capturing the sophistication of information technology due to computers).

Whereas *Info* is the response ("dependent") variable in the analysis, the predictor ("independent") variables represent other aspects of social complexity, or "Complexity Characteristics" as defined in Turchin et al. (2018: see Figure 2a). The first set of predictor variables relates to the size of polities: polity population (*PolPop*), extent of polity territory (*PolTerr*), and "capital" population (the size of the largest urban center, *CapPop*). These three variables serve as proxies for the social scale, and are log-transformed (base 10) prior to analysis. Log-transformation is an appropriate way to treat these quantitative variables because our main focus is on the order of magnitude. In other words, a transition in polity population from 1000 to 10,000 (1 on the logarithmic scale) is similar to a transition from 10 million to 100 million (also 1 on the logarithmic scale). If polity population is not log-transformed, nearly all variation in this variable would be dominated by the upper end of the scale (tens of millions).

Another set of variables measures hierarchical complexity focusing on the number of control/decision levels in the administrative, religious, and military hierarchies, and on the hierarchy of settlement types (village, town, provincial capital, etc.). These four Seshat variables were combined into a single measure of hierarchical levels (*Hier*) by averaging (over non-missing values).

Government (*Gov*) variables code for the presence or absence of official specialized positions that perform various functions in the polity: professional soldiers, officers, priests, bureaucrats, and judges. This class also includes characteristics of the bureaucracy (e.g. presence of an examination system), the

judicial system, and specialized buildings (e.g. courts). Infrastructure (*Infra*) captures the variety of observable structures and facilities that are involved in the functioning of the polity. Both *Gov* and *Infra* measures add together a number of binary variables (11 for *Gov* and 12 for *Infra*).

Finally, economic development is reflected in Monetary System (*Money*). The *Money* scale reflects the "most sophisticated" monetary instrument present in the coded society (0: none, 1: Articles, 2: Tokens, 3: Precious metals, 4: Foreign coins, 5: Indigenous coins, 6: Paper currency). We refer to *PolTerr*, *PolPop*, *CapPop*, and *Hier* as "social scale" Complexity Characteristics, and *Info*, *Gov*, *Infra*, and *Money* as "non-scale" Complexity Characteristics.

Sample Size and Structure

Once all Complexity Characteristics (CCs) are aggregated they are put together in a data file whose columns are polity name, NGA name, time (in centuries), and the values of eight CCs (see the SOM for a description of the data file published as part of R-scripts). The data file analyzed in this article has 456 rows. This number is larger than the 332 unique polities, because many polities experience changes in one or more of CCs during their duration. When data change, an additional row for the polity is created and added to the dataset. The two rows will have the same polity name, but different time periods.

The proportion of missing values varies by CCs. The least well sampled CCs is *PolPop* with 310 observations (see SOM for these statistics). By definition, the response variable (*Info*) has no missing values, because it is not subjected to multiple imputation. Overall, 222 rows out of 456 in the data table have no missing values. Frequency distributions for all variables are plotted in Figure S1 in the SOM.

Time-series analysis requires data sampled at regular time intervals (set to one century: see *Dynamic Regression: Methods* below). This requires interpolation, which is best explained with a concrete example. Thus, the Seshat polity Latium – Iron Age starts at 1000 BCE end ends at 717 BCE. Because none of the CCs changed during this period, the data file has a single row for this polity. The interpolated data file, on the other hand, devotes three rows to the polity (for 1000, 900, and 800 BCE). Each row is identical except for time.

Interpolated data, used in time-series analysis, has 902 observations (of which about half simply repeat the data values for the previous century). Note, however, that if we need to estimate autoregressive terms, we end up with fewer observations. For example, if for an NGA we have a sequence of centuries from 500 to 1500 CE, or 11 observations, estimation of a model that includes AR(2) terms will leave us with only 9 observations (because the dependent variable,

sampled at 700, 800, … 1500 CE, requires Lag1 sampled at 600, 700, … 1400 and Lag2 sampled at 500, 600, … 1300). The effective sample sizes and degrees of freedom for each specific analysis can be found in the R output included in the SOM.

Finally, the length of time-series varies by NGA (see Figure 1a). The longest one is 115 time steps (for Konya Plain), and the shortest are several "late complexity" NGAs with just two or three observations.

Multiple Imputation

Dealing with Missing Data, Uncertainty, and Expert Disagreement

Due to the fragmentary nature of the information that is available about past societies it is not possible to reliably code all variables for all polities. There is therefore a non-trivial amount of data points which we have to code as "unknown". The presence of such missing data is an important feature of Seshat in that it accurately reflects our current understanding (or lack of it) about any particular feature in any particular past society. Missing data, however, present a challenge for statistical analyses.

One way of dealing with incomplete data sets is to simply omit the rows in the data matrix that contain missing values. There are two problems with this approach. First, it can be very wasteful in that omitted rows may contain much useful information relating to the variables that were coded. Had we used this approach with our social complexity data, for example, we would have to throw away approximately half the rows. Second, row deletion may lead to biased estimates because there are often systematic differences between the complete and incomplete cases. In our case, in many NGAs small-scale societies were present far back in time and, as a result, they are much harder to code. Additionally, some regions of the world have been subject to greater levels of research effort than others. Omitting many of the lesser known cases, due to their larger proportion of missing values, would give too much weight to later, better known societies from only some parts of the world. As an example, had we used the row deletion approach for our current dataset, we would end up with very few observations for Australia-Oceania. Such unequal deletion of observations would very likely bias the results, since the analysis would be dominated by such regions as Europe and Southwest Asia.

In order to deal with missing values, as well as incorporating uncertainty and expert disagreement into our analyses, we employ a technique known as *multiple imputation* (Rubin 1987), which utilizes modern computing power to extract as much information from the data as possible. Imputation involves replacing

missing entries with plausible values, and this allows us to retain all cases for the analysis. A simple form of imputation, "single imputation", might replace any unknown cases for a binary "present/absent" variable with simply "absent", or to replace unknown cases of continuous variables with the mean for that variable. These approaches have similar drawbacks to row deletion in that they tend to introduce a bias. Multiple Imputation avoids this problem: Analysis is done on many data sets, each created with different imputed values that are sampled in probabilistic manner. This approach results in valid statistical inferences that properly reflect the uncertainty due to missing values (Yuan 2010). Multiple imputation procedures can vary depending on the type of variable and the type of data coding issue faced.

Expert disagreement. In cases where experts disagree we select each alternative coding with equal probability. Thus, if there are two conflicting values coded by different experts and we create 20 imputed sets, each alternative will be used roughly 10 times.

Uncertainty. Scalar values that are coded with a confidence interval are sampled from a Gaussian distribution whose mean and variance are estimated assuming that the interval covers 90 percent of the probability. For example, if a value of [1000–2000] was entered for polity population, we draw values from a normal distribution centered on 1500 with a standard deviation of 304. It is worth noting that this procedure means that in 10 percent of cases the value entered into the imputed set will be outside the data interval coded in Seshat. For categorical or binary variables we sample coded values in proportion to the number of categories that are presented as plausible. For example, if our degree of knowledge doesn't allow us to tell whether a certain feature was present or absent at a particular time then the imputed data sets will contain "present" for roughly half the imputed sets and "absent" for roughly half the sets.

Missing data. For missing data we impute values as follows. Suppose for some polity we have a missing value for variable A and coded values for variables B through H. We select a subset of cases from the full dataset in which all variables A through H have values and build a regression model for A based on predictors B-H. Not all predictors B–H may be relevant to predicting A, and thus the first step is selecting which of the predictors should enter the model (see below on model selection). Once the optimal model is identified, we estimate its parameters. Then we go back to the polity where variable A is missing and use the known values of predictor variables for this polity to calculate the expected value of A using the estimated regression coefficients. However, we do not simply substitute the missing value with the expected one (because, as explained above, this will result in biased estimates). Instead, we sample from the posterior

distribution characterizing the prediction of the regression model (in practice, we randomly sample the regression residuals and add it to the expected value). We apply the same approach to each missing value in the data set, yielding an imputed data set without gaps.

Multiple imputation was applied only to the predictor variables (*PolPop*, *PolTerr*, *CapPop*, *Hier*, *Gov*, *Infra*, and *Money*). The response (dependent) variable (*Info*) was not imputed, and *Info* was not used for the imputation of predictor variables. Thus, the data rows with missing values of the predictor variable are not used in regressions. The overall imputation procedure was repeated 20 times, yielding 20 imputed sets that were used in the analyses below.

Cross-validation

One interesting issue in helping us interpret multiple imputation results is how accurately the stochastic regression approach can predict missing values. Most importantly, does this approach actually yield better estimates than, for example, simply using the mean of the variable? In order to answer this question I employ a statistical technique known as *k*-fold cross-validation (Kohavi 1995).

Cross-validation estimates the true predictability characterizing a statistical model by splitting data into two sets. The parameters of statistical model are estimated on the *fitting set*. Next, this fitted model is used to predict the data in the *testing set*. Because the prediction is evaluated on the "out of sample" data (data that were not used for fitting the model), the results of the prediction exercise give us a much better idea of how generalizable the model is, compared to, for example, such regression statistics as the coefficient of determination, R^2.

The accuracy of prediction is often quantified with the coefficient of prediction:

$$\rho^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i^* - Y_i)^2}{\sum_{i=1}^{n}(\bar{Y} - Y_i)^2}$$

where Y_i are the observations from the testing set (the omitted values), Y_i^* is the predicted value, $\bar{Y}$ is the mean of Y_i, and n is the number of values to be predicted. The coefficient of prediction ρ^2 equals 1 if all data are perfectly predicted and 0 if the regression model predicts as well as the data average (in other words, if the model is simply $Y_i^* = \bar{Y}$). Unlike the regression R^2, which varies between 0 and 1, prediction ρ^2 can be negative—when the regression model predicts data worse than the data mean. Prediction ρ^2 becomes negative when the sum of squares of deviations between predicted and observed is greater than the sum of squares of deviations from the mean.

In *k*-fold cross-validation, rather than having simply a single fitting set and a single testing set, we divide the data into *k* sets. Because the Seshat World Sample was designed to code polities in 10 major regions spanning the globe, it makes sense to use these regions as data sets (Turchin et al. 2018). Such a procedure is superior to dividing data into 10 sets randomly, because it automatically eliminates temporal and within-region correlations between the values in the fitting set and values in the testing set. Thus, "out of sample" prediction becomes "out of region" prediction, which is obviously more challenging than when the prediction method utilizes data points within the same region both for building the predictive model and for testing it.

Here I illustrate this approach by testing how well the method predicts values of the Info variable. In the previous article we tested the approach on the part of the dataset for which all rows lacked missing values, $n = 203$ (Turchin et al. 2018: Supplementary Results). I will use a modification that allows me to more fully utilize the Seshat data ($n = 456$).

The procedure works as follows. We start by selecting a row of data that has a coded value for the *Info* variable and check how many other variables are not missing. For example, for Copper Age Italy (3600–1800 BCE) we have estimates for *Hier*, *Gov*, *Infra*, and *Money* variables (in addition to *Info*), but we lack estimates for *PolPop*, *PolTerr*, and *CapPop*. Next, we construct a fitting set by selecting all rows which have data for Info, *Hier*, *Gov*, *Infra*, and *Money*. We omit, however, any rows associated with the World Region in which Italy is located, Europe (because our goal is to calculate the accuracy of out-of-region prediction). The remainder is our fitting data set.

We don't know whether all predictor variables (*Hier*, *Gov*, *Infra*, and *Money*) are needed to predict *Info*. One of the most commonly used methods of model selection is to rely on the AIC, or Akaike Information Criterion (Burnham and Anderson 1998), which estimates the relative information loss between models based on different predictors. I performed an exhaustive search (fitting regressions with all possible combinations of predictor variables) and selected the combination that minimizes the AIC. As an aside, such a "brute force" approach is not the most efficient way of doing model selection, but modern computers are so powerful that it doesn't make sense to optimize the search strategy.

Once the best model (as indicated by the AIC) is found, we can use it to predict the value of *Info* for Copper Age Italy, using whichever of the potential predictor variables (*Hier*, *Gov*, *Infra*, and *Money*) that the best model needs. We now store both the observed value of *Info* and the predicted one, and repeat the procedure for all 456 rows in the data matrix. Finally, we calculate the coefficient

of prediction using the formula above. All analyses were conducted using scripts[2] written in the R statistical programming language.

The *k*-fold cross-validation is computer-intensive but it utilizes the data in a very efficient way: all data eventually end up in the testing set, while each fitting set uses, on average, 90 percent of data (data from all regions except the one in which the datum to be predicted is located).

Results: *k*-Fold Cross-Validation of *Info*

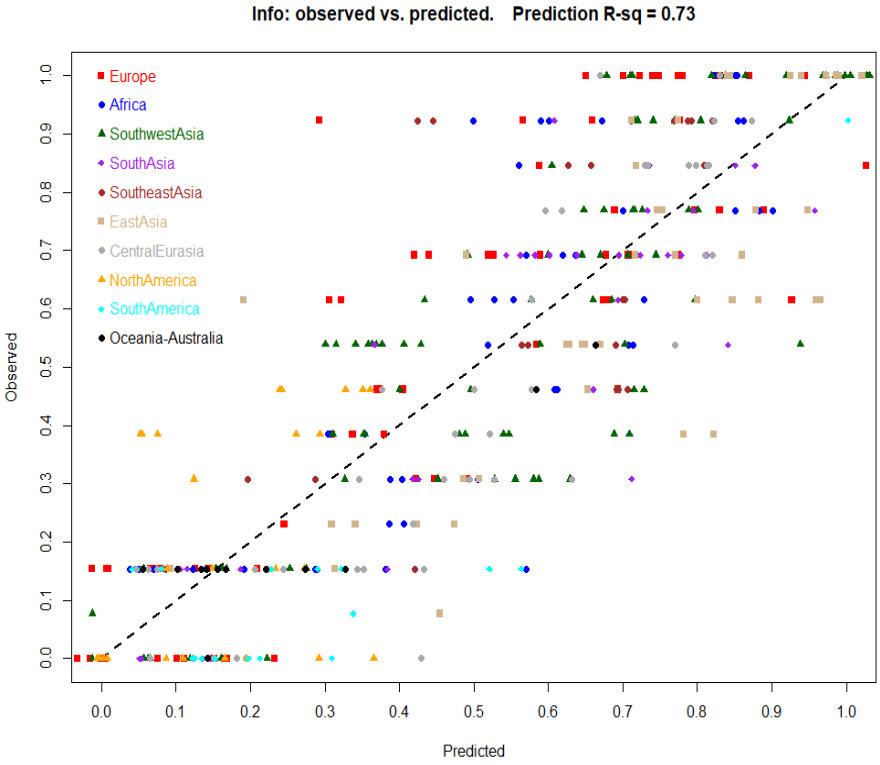


Figure 2. Comparing observed to predicted *Info* values. Data are color-coded by world region. The broken straight line is one-to-one correspondence (or perfect prediction).

[2] These are available at: https://doi.org/10.17916/P6159W.

Plotting observed versus predicted values of *Info* for the Seshat data indicates a strong linear relationship between the two. Prediction ρ^2 is a very credible 0.73, showing that over 70 percent of variance in *Info* can be predicted from the knowledge of other variables for the particular society. This result supports one of the main conclusions in Turchin et al. (2018) that key aspects of human social organization tend to co-evolve in predictable ways. Note that the cross-validation analysis in this article is based on a substantially larger sample size (n = 456) compared to n = 203 in Turchin et al. (2018).

Ability to accurately predict values of Complexity Components also supports the approach of using stochastic regression with multiple imputation. Using existing values to impute the missing ones yields smaller variation in imputed values, compared to, for example, sampling from the overall distribution, and this approach enables us to more reliably detect the patterns present in the data.

As a final note, in the previous article (Turchin et al. 2018) we also assessed whether the multiple imputation (MI) method used in this analysis could have introduced bias into our results. We created 100 artificial data sets that randomly introduced missing values into our "complete data set" (which contained only complete rows), reproducing the pattern of missing values in the "overall data set". We then applied the MI procedure to each of the artificial data sets in exactly the same way as the overall data were analyzed. By comparing the PCA results based on artificial datasets with results from the complete dataset, we saw that the Multiple Imputation procedure accurately captures the overall patterns in the data both in terms of the number and pattern of Principal Components (PCs) produced (Turchin et al. 2018: Figure SI7), and the loadings of the different variables on to PC1 (Figure SI8). Our overall conclusion was that the MI procedure does not introduce a systematic bias into analysis results.

Dynamic Regression

Methods

The general regression model that I used above to investigate factors affecting the evolution of the Seshat measure of information complexity (*Info*) takes the following form:

$$Y_{i,t} = a + \sum_{\tau} b_\tau Y_{i,t-\tau} + c \sum_{i \neq j} \exp\left[-\frac{\delta_{i,j}}{d}\right] Y_{j,t-1} + h \sum_{i \neq j} w_{i,j}\, Y_{j,t-1} + \sum_{k} g_k\, X_{k,i,t-1} + \epsilon_{i,t}$$

Here $Y_{i,t}$ is the response variable; in our case it is the value of *Info* coded for location *i* at time *t*. Recollect that we construct a spatio-temporal series for Seshat variables by following polities (or quasipolities, such as archaeologically attested cultures) that occupied a specific NGA (Natural Geographic Area) at each century mark during the sampled period. Thus, the time step Δt = 100 years.

On the right-hand side, *a* is the regression constant (intercept). The next term captures the influences of past history of *Info* ("autoregressive terms"), with τ = 1, 2, … indexing time-lagged values of *Y* (as time is measured in centuries, $Y_{i,t-1}$ refers to the value of *Info* 100 years before *t*). The third term represents potential effects resulting from geographic diffusion of *Info* (Eff and Dow 2009; Eff and Routon 2012). I use a negative-exponential form to relate the distance between society *i* and society *j*, $\delta_{i,j}$, to the influence of *j* on *i* because, unlike with a linear kernel, the negative-exponential one does not become negative at very long $\delta_{i,j}$, instead approaching 0 smoothly. Note that we avoid the problem of endogeneity (simultaneous causality) by using time lagged $Y_{j,t-1}$ (see *General Introduction*). The third term, thus, is a weighted average of *Info* values in the vicinity of society *i* at the previous time step, with weights falling off to 0 as distance from *i* increases. Parameter *d* measures how steeply the influence falls with distance, and parameter *c* is a regression coefficient measuring the importance of geographic diffusion.

The fourth term detects autocorrelations due to any shared cultural history of *Info* at location *i* with other regions *j*. Here *w* represents the weight due to *phylogenetic* (linguistic) distance between locations (set to 1 if locations *i* and *j* share the same language, 0.5 if they are in the same linguistic genus, and 0.25 if they are in the same linguistic family). Linguistic genera and families were taken from *The World Atlas of Language Structures* (Dryer and Haspelmath 2013) and Glottolog (Hammarström et al. 2017). The rest of the right-hand side represents the effects of the predictor variables $X_{k,i,t-1}$ (time-lagged); g_k are regression coefficients; and $\varepsilon_{i,t}$ is the error term.

This approach allows us to investigate the effects of the predictor variables (Complexity Components other than *Info*), while controlling for serial autocorrelations, spatial diffusion, and autocorrelations due to shared cultural history (as discussed in *General Introduction*, the latter two are known as Galton's

Problem). It is important to note that effects of spatial diffusion and shared origin are not simply "nuisance parameters" that we want to eliminate. These are interesting mechanisms of cultural evolution in their own right. In particular, differentiating the effects of spatial diffusion from influences by the predictors allows us to estimate the relative importance of cultural evolution *in situ* (under the influence of the predictor variables) versus cultural borrowing from nearby areas.

All parameters in the equation above can be estimated with linear regression, except for d (which scales the distance effect of the spatial diffusion term). I set an initial value of d equal to 1000 km, because this was a characteristic distance between neighbor NGAs (over two-thirds of nearest neighbor distances were between 500 km and 1500 km). Next, I searched for an optimal d value by running the best regression (see below) for a range of d-values on a grid of (100, 200, 300, …, 1900, 2000). The estimated value of d (the one with the smallest AIC) was 1100 km (however, for none of these values the spatial term was significant at the $P = 0.05$ level, see below).

Given a value of d, I fitted model (3) using the R function glm (Generalized Linear Models). Model selection was accomplished by fitting all possible linear models and selecting the one that yielded the smallest AIC.

All analyses reported in this article are based on data scraped from the Seshat Databank on March 11, 2018. These data will be available for download at http://seshatdatabank.info/datasets/ upon publication of the article.

The summary statistics and distributions of the response and predictor variables are given in the Supplementary Online Material.

Results: Processes Influencing the Evolution of *Info*

The first step in the analysis aims to understand how much variability in the results is introduced by missing data, uncertainty, and expert disagreement. As was explained above (*Dealing with Missing Data, Uncertainty, and Expert Disagreement*), I generated 20 imputed datasets. I then submitted each dataset to an R-script, which finds the combinations of predictors that yield the best AIC (*Dynamic Regression: Methods*). As we shall see later, variation due to imputation was slight. For this reason, I first report the regression results for a dataset with averaged predictors (in which each value was an average of 20 imputed values; as a reminder, the response variable was not imputed and thus we only have one set of values for it). The spatial diffusion term did not have a statistically significant effect in any of the models, and thus it is not included in the following results. Furthermore, additional checks indicated that the time evolution of Info is

appropriately described by an autoregressive process of order 2 (in other words, only the first two autoregressive terms have a significant effect).

Fitting regression models with all possible combinations of linear terms suggests strong effects (*t*-values greater than 3) of the autoregressive terms (Lag1 and Lag2), Phylogeny, and Money (Table 1). Two other predictors, PolPop and Gov have weaker effects (with *t*-values around 2). Infra appears to have a negative effect on the trajectory of Info; however, this effect is not significant at the conventional level of $P < 0.05$. The frequency with which various terms are included in the ten best models roughly corresponds to this variation in *t*-values.

Table 1. Results of the best 10 linear models (with the smallest AIC).

Lag1	Lag2	PolPop	PolTerr	CapPop	Hier	Gov	Infra	Money	Phyl	R^2	ΔAIC
18.73	3.86	2.15				1.83	−1.48	3.40	3.67	0.91	0.00
18.89	3.80	1.78				1.49		3.08	3.72	0.91	0.22
20.14	4.04	2.65						3.53	3.71	0.91	0.45
18.75	3.86	2.06		−1.07		1.67		3.19	3.56	0.91	1.06
18.73	3.85	1.93	−0.83			1.78	−1.67	3.33	3.72	0.91	1.30
20.14	4.10	2.84					−1.04	3.65	3.67	0.91	1.37
19.30	3.81					2.46		3.39	3.67	0.91	1.40
18.64	3.89	2.20		−0.75		1.91	−1.27	3.43	3.55	0.91	1.43
18.53	3.82	1.90			0.43	1.59	−1.52	3.42	3.67	0.91	1.81
19.51	3.93	1.92			0.79			3.47	3.70	0.91	1.83

Note: This table reports *t*-values associated with various terms in the best-fitting models. Lag1 and Lag2 refer to autoregressive (lagged) terms ($Y_{i,t-1}$ and $Y_{i,t-2}$); PolPop through Money are the predictors, Phyl is the effect of phylogeny, R^2 is the regression coefficient of determination, and ΔAIC is the difference in the Akaike Information Criterion with respect to the best model (with the lowest AIC).

Before proceeding further, we need to run some diagnostic checks that assess the validity of various assumptions of the regression model. Figure 3 shows four standard diagnostic tests. The two top panels look good (there are no trends). The patterning, which we see in the Residuals vs. Fitted pane (and in Scale-Location pane), is due to the fact that the dependent variable is not truly continuous, but takes one of 14 discrete values between 0 and 13 (see *Aggregation into "Complexity Components"*). This is also why the upper right portion of the data cloud the panel looks "sliced off": there is a hard upper limit

above which Info cannot increase. The plot of Residuals vs Leverage also looks good, as all cases are well inside the Cook's distance lines.

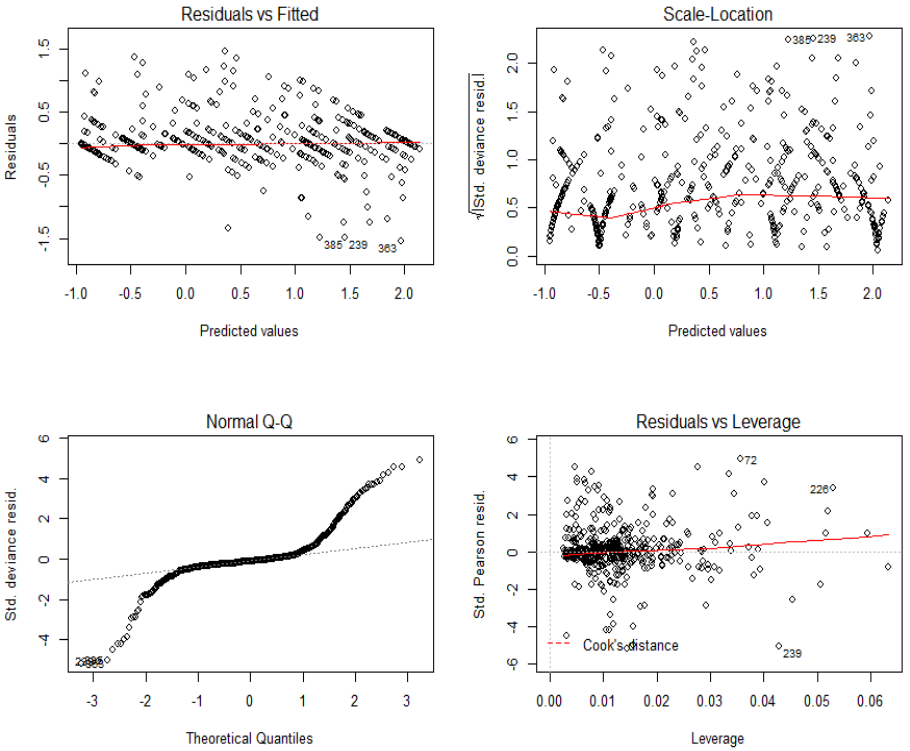


Figure 3. Diagnostic plots for the best-fitting model.

The Normal Q-Q plot, on the other hand, shows that residuals are clearly non-Gaussian. Instead of points following an approximately straight line, there is a "plateau" in the middle, indicating an overabundance of small deviations (residuals near 0). This pattern arises from a tendency of many traits under cultural evolution to be faithfully transmitted to future generations. As a result, temporal trajectories of such traits are characterized by periods of stasis, punctuated by (sometimes) rapid change. Fortunately, unlike the problem of "heavy tails", which results in inflating the significance levels calculated under the

normality assumption, "heavy middle", if anything, results in overly conservative *P*-values and can be ignored (Faraway 2002).

Another diagnostic check that we need to do is whether there are any lingering temporal autocorrelations not captured by the regression model (recollect, that all best models include both Lag1 and Lag2 terms). A check of higher order effects indicated that there are no significant correlations between residuals and lags 3, 4, 5 … at $P = 0.05$ level. In other words, an AR2 process (autoregressive model with two lags) appears to adequately capture all time dependencies in Info.

Seshat data is a space-time series: a time trajectory anchored at each NGA (Natural Geographic Area). One way to check for the influence of unobserved variables is to hold constant all time invariant NGA characteristics that could be driving correlations between the response variable and regressors. Accordingly, I fitted a regression model to the data with NGAs as fixed effects. The results are very interesting (see the SOM: *Supplementary Results*): two NGAs, Iceland and Ghanaian Coast, are highlighted by this analysis as being different from the rest. At the same time, the puzzling negative effect of Infra disappears in the fixed-effects model. Something is different about these two NGAs, but it is difficult to determine what is going on because they together have contributed only three observations to the data set. Dropping these three observations results in a simplified model, in which there are no differences between remaining NGAs and no effect of Infra (SOM: *Supplementary Results*).

Apart from clearing up the puzzling negative effect of Infra, the output of the fixed-effect regression (on both the complete data, or the one in which the three suspect observations are dropped) is reassuring in another, and more important way. It suggests that unobserved heterogeneity between NGAs is not what drives the main findings about factors affecting the evolution of Info. The fixed-effect regression identifies the same predictors: Lag1, Lag2, Phylogeny, Money, PolPop, and Gov. This observation helps buttress the case that the partial correlations, revealed by these regression analyses, represent causal relationships.

Another approach of testing for the effect of unobserved variables is to fit a regression model with time as a covariate. Because the term associated with time results in an improvement of the fit (SOM: *Supplementary Results*), my conclusion is that, indeed there are some, as yet unidentified, factors affecting the evolution of Info (this will be further addressed in *Discussion*).

Finally, and as was discussed in *General Introduction*, it is always a good idea to test for nonlinear effects. Accordingly, I fitted a series of models by adding quadratic terms associated with those predictors that were selected in the best-fitting linear model. This investigation showed that adding squared terms for

Lag1 and Money significantly improved the degree of fit (SOM: *Supplementary Results*). It also indicated that dropping the Gov term further betters the AIC.

Table 2 presents the results of the best model (smallest AIC) that considers both linear and quadratic effects.

Table 2. Regression results: the best nonlinear model. Regression coefficients have been standardized by scaling all variables in the model to mean = 0 and variance = 1.

Standardized Coefficients:

	Estimate	*Std. Error*	*t value*	*Pr(>|t|)*
(Intercept)	0.000	0.010	0.000	1.00000
Lag1	0.750	0.040	18.877	0.00000
Lag1.sq	–0.078	0.016	–4.874	0.00000
Lag2	0.124	0.031	4.001	0.00007
Gov	0.058	0.026	2.221	0.02665
Money	0.052	0.019	2.724	0.00658
Money.sq	0.031	0.013	2.354	0.01883
Phylogeny	0.033	0.013	2.550	0.01095
Time	0.025	0.013	1.894	0.05861

Discussion

This article presents a general methodology for fitting dynamic regression models to time-resolved cross-cultural data. My primary focus is methodological, and I illustrate the general approach by analyzing *Info*, the Seshat variable that attempts to capture the sophistication of information system of past societies. After a variety of analyses and diagnostic checks, documented in the previous section and in *Supplementary Online Materials: Results*, I converged on a particular model, the results of which are presented in Table 2. What do these results tell us about the evolution of *Info*? Most importantly, what are the caveats?

Before dealing with the three hypotheses described in the introduction, let's discuss the autoregressive terms in the model. It is not surprising that *Lag1* (the value of a location's *Info* lagged by one century) has an enormous impact on *Info*: this effect captures the memory in the system. Note that the estimated standardized coefficient for *Lag1* is 0.75. This means that changing the value of *Info* at the previous time step by one unit will result in a change of Info at the next time step by 0.75 units. The quadratic term, *Lag1.sq*, is also highly influential. The negative sign associated with the coefficient tells us that it is a "regulatory" term: when *Info* gets too high, this term "steps on the brakes" to impede its further increase. One possible explanation of this effect is that it is simply a result of the

hard ceiling on *Info* values that I discussed above. But it is also plausible that when other complexity characteristics, such as *Gov* and *Money*, decline, so does the equilibrium level of *Info*. A possible example of such dynamics is the collapse of Roman Empire, which brought in its wake a substantial decline in the sophistication of government and economy within Italy, followed by a "Dark Age" characterized by a drastic decline in literacy and the number of text types in circulation.

The coefficient associated with *Lag2* is positive, suggesting that evolving social systems have a longer-term memory than preceding time step. Such longer memory effect could help *Info* to recover after a negative shock. For example, suppose that a systemic collapse results in a loss of knowledge, depressing *Info*. If *Info* was high before the collapse, then it would recover faster after it, because the social system could "remember" (or recover) recently lost information. More broadly, however, a highly influential second-order autoregressive term is a strong indication that we are missing information on an important dynamic driver (Turchin and Ellner 2000).

The coefficient associated with *Phylogeny* is also positive, while the direct spatial effect is not statistically significant. This result suggests that *Info* tends to be transmitted between linguistically similar cultures, no matter how distant they are, rather from linguistically dissimilar, but geographically near neighbors. This result is, perhaps, not surprising, as *Info* has an obvious linguistic component. It highlights the possible importance of cultural closeness in the diffusion of cultural traits, over and above the impact of geographic proximity. However, an important caveat here is that in the current Seshat data space is seriously "undersampled," as a typical distance between neighbor NGAs is on the order of 1000 km. It is quite possible that when we analyze denser-sampled data in Seshat, we will be better able to detect the influences due to geographic proximity.

Turning to Complexity Components, we observe that the greatest influence on the evolution of *Info* is exerted by *Money*. In fact, *Money* is the most important predictor of them all, barring the *Lag1* and *Lag2* terms. As Table 2 shows, *Money* enters the best model in a nonlinear (quadratic) form. The positive sign of the coefficient associated with the quadratic term (*Money.sq*) indicates that it is the higher ranges of this variable that have the greatest effect on increasing *Info*.

Gov is also a part of the best-fitting model. It is an interesting and somewhat unexpected result, however, that the economic sophistication, proxied by *Money*, has a stronger influence on *Info* than government specialization. However, at the current stage of the development of the database and analysis methodologies, this result should be treated as a very tentative one. In particular, it is entirely

possible that the inclusion into the analysis of additional variables, currently coded by the Seshat team, will change our understanding of this evolutionary process in both quantitative terms (relative strength of various predictors) and qualitative terms. For example, one of the predictors currently included in the best model may be replaced by a better causal variable, as more variables become available for analysis.

Finally, *PolTerr* does not appear in the model, suggesting that the need to get messages to far flung regional centers was not a strong factor in the evolution of information systems. In fact, none of the social scale characteristics (*PolPop*, *PolTerr*, or *CapPop*) appears to influence *Info*, once other factors (e.g., *Money* and *Gov*) are taken into account. Again, however, this conclusion must be treated as preliminary. First, note that models including *PolPop*, for example, were among the best linear models (Table 1). Second, the difference in AIC between the best-fitting nonlinear model, and the one that also includes *PolPop* is only 1.4 (see *Supplementary Online Material*). As a rule of thumb, models that differ in AIC by less than 2 are considered to have a similar degree of support. Thus, we need additional data, which will either strengthen, or further weaken this result.

I emphasize that these results are tentative and preliminary, and have been included in the paper merely for illustrative purposes. An important caveat is that we need to run similarly careful analyses on other Complexity Characteristics before we can characterize the web of causal pathways connecting different aspects of Social Complexity. Second, we should keep in mind that the effect of *Gov* or *Money* could be due to some other, hidden variables with which these predictors could be closely correlated. In fact, there are indications in the analysis that we do not have data on all relevant predictors, since the best model includes a temporal trend and second-order autoregressive terms.

Thus, one general conclusion from this analysis is as expected: we need more data on different aspects of past societies. What this analysis adds, however, is that we can confidently invest resources in collecting such data, because we have now demonstrated that they will be very useful in helping us distinguish between various hypotheses about cultural evolution of social complexity. Prior to collecting Seshat data and running statistical analyses on it, it was not clear at all that this approach would work. One possible outcome could have been an *uninformative* data set and analytical results saying that we cannot distinguish between various hypothesis. In other words, "history is just one damn thing after another," as Arnold Toynbee famously characterized this view. But our analyses in this article (and in Turchin et al. 2018) show that there are strong patterns in historical data; and that the data already in Seshat allow us to go a long way

towards uncovering them. We can test rival theories against each other and obtain clear results on their validity. The historical record is highly informative.

In addition to more data, we also need more sophisticated methods of analysis. In particular, the geographic diffusion term in the general regression model (*Dynamic Regression: Methods*) is too vague. We need better models to distinguish between multiple ways that cultural traits could be transmitted across space: trade, conquest, and direct cultural borrowing. As I have pointed out above, horizontal transmission is not a "nuisance term" to include in the model and then to forget about. It represents processes of high interest to Cultural Evolution, and we need better methodology to detect such influences. We also need data that samples space better, because that would increase our ability to distinguish between different models of horizontal transmission.

The original motivation behind developing Standard Cross-Cultural Sample was to eliminate Galton's Problem. Subsequent analyses showed that the SCCS failed in this respect. Better statistical methods, which estimated Galton terms (spatial diffusion and related phylogeny), resulted in huge progress. The Seshat project has made great strides in improving the temporal sampling, and this enables us to fit dynamic regression models, as was illustrated in this article, and by doing so, to approach causal processes in cultural evolution. Expanding the Seshat approach to a denser spatial sample will allow us to investigate the influence of horizontal transmission of cultural traits resulting from trade, warfare, and direct information exchange.

We also need to link more tightly theory development with data analysis. Atheoretical "Big Data" approaches are of limited utility for revealing causal mechanisms of cultural evolution. One potential problem is the one-to-many mapping between any empirical factor and a set of potential causal mechanisms. Another is that there could be a mismatch between functional forms used in regression analysis (e.g., linear, or even polynomial relationships) and functional forms arising from mechanism-based models. While there are statistical methods that fit arbitrary functional forms to data, such approaches are extremely data-hungry. An alternative approach is to use mechanism-based models directly in data analysis. A discussion of such methodologies is well beyond the scope of this article. I merely note that my colleagues and I are working in this direction (for example, see Turchin et al. 2013). A combination of theory, built from first principles, with sophisticated data analysis has been used very effectively in such nonlinear dynamics fields as complex population dynamics (for a review, see Turchin 2003). I fully expect that this approach would work equally well in helping us understand the evolution of complex societies.

Acknowledgement

The author is grateful to Pat Savage, Dan Hoyer, Jim Bennett, Romain Wacziarg, Ömer Özak, and two anonymous reviewers for their comments on manuscript. This work was supported by a John Templeton Foundation grant to the Evolution Institute, entitled "Axial-Age Religions and the Z-Curve of Human Egalitarianism," a Tricoastal Foundation grant to the Evolution Institute, entitled "The Deep Roots of the Modern World: The Cultural Evolution of Economic Growth and Political Stability," an ESRC Large Grant to the University of Oxford, entitled "Ritual, Community, and Conflict" (REF RES-060-25-0085), a grant from the European Union Horizon 2020 research and innovation programme (grant agreement No 644055 [ALIGNED, www.aligned-project.eu]), and an European Research Council Advanced Grant to the University of Oxford, entitled "Ritual Modes: Divergent modes of ritual, social cohesion, prosociality, and conflict." I gratefully acknowledge the contributions of our team of research assistants, post-doctoral researchers, consultants, and experts. Additionally, I have received invaluable assistance from our collaborators. Please see the Seshat website (http://seshatdatabank.info/seshat-about-us/acknowledgements/) for a comprehensive list of private donors, partners, experts, and consultants and their respective areas of expertise.

References

Burnham, K. P. and D. R. Anderson. 1998. *Model Selection and Inference: A Practical Information-Theoretic Approach*. New York: Springer. doi: 10.1007/978-1-4757-2917-7.

Childe, V. G. 1950. "The Urban Revolution." *Town Planning Review* 21: 3–17. doi: 10.3828/tpr.21.1.k853061t614q42qh.

Comin, D., W. Easterly, and E. Gong. 2010. "Was the Wealth of Nations Determined in 1000 BC?" *American Economic Journal: Macroeconomics* 2: 65–97. doi: 10. 1257.mac.2.3.65.

Currie, T. E., and R. Mace. 2012. "The Evolution of Ethnolinguistic Diversity." *Advance in Complex Systems* 15:1–20. doi: 10.1142/s0219525911003372.

Dryer, M. S., and M. Haspelmath, editors. 2013. *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Eff, E. A., and M. M. Dow. 2009. "How to deal with missing data and Galton's problem in cross-cultural survey research: a primer for R." *Structure and Dynamics* 3 (2): Article 1.

Eff, E. A., and P. W. Routon. 2012. "Farming and Fighting: An Empirical Analysis of the Ecological-Evolutionary Theory of the Incidence of Warfare." *Structure and Dynamics* 5: 1–33.

Faraway, J. J. 2002. Practical Regression and Anova using R. July 2002.

François, P., J. G. Manning, H. Whitehouse, R. Brennan, T. Currie, K. Feeney, and P. Turchin. 2016. "A Macroscope for Global History: Seshat Global History Databank: a methodological overview." *Digital Humanities Quarterly* 10.

Fukuyama, F. 2011. The Origins of Political Order: From Prehuman Times to the French Revolution. New York: Farrar, Straus and Giroux.

Goody, J. 1986. *The Logic of Writing and the Organization of Society*. Cambridge: Cambridge University Press. doi: 10.1017/cbo9780511621598.

Granger, C. W. J. 1969. "Investigating Causal Relations by Econometric Models and Cross-spectral Methods." *Econometrica* 37: 424–438. doi: 10.2307/1912791.

Hammarström, H., S. Bank, R. Forkel, and M. Haspelmath, editors. 2017. "Glottolog 3.1." *Max Planck Institute for the Science of Human History, Jena*.

Johnson, A. W., and T. Earle. 2000. *The Evolution of Human Societies: From Foraging Group to Agrarian State*, 2nd edition. Stanford, CA: Stanford University Press.

Kohavi, R. 1995. "A study of cross-validation and bootstrap for accuracy estimation and model selection." *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* 2: 1137–1143.

Lotka, A. 1925. *Elements of Physical Biology.* Baltimore: Williams and Wilkins.

Mace, R., and C. J. Holden. 2005. "A phylogenetic approach to cultural evolution." *Trends in Ecology and Evolution* 20: 116–121. doi: 10.1016/j.tree.2004.12.002.

Maynard Smith, J., and E. Szathmáry. 1995. *The Major Transitions in Evolution.* New York: W. H. Freeman.

Murdock, G. P. 1967. *Ethnographic Atlas*. Pittsburgh: University of Pittsburgh Press.

Murdock, G. P., and D. R. White. 1969. "Standard Cross-Cultural Sample." *Ethnology* 8: 329–369. doi: 10.2307/3772907.

Nolan, P. D. 2003. "Toward an Ecological-Evolutionary Theory of the Incidence of Warfare in Preindustrial Societies." *Sociological Theory* 21: 18–30. doi: 10.1111/1476-9558.00172.

Pearl, J. 2009. *Causality: Models, Reasoning and Inference*. 2nd Edition. Cambridge: Cambridge University Press.

Peregrine, P. N. 2003. "Atlas of Cultural Evolution." *World Cultures* 14: 2–88.

Peregrine, P. N., and M. Ember, editors. 2001. *Encyclopedia of Prehistory*. New York: Kluwer Academic/Plenum Publishers.

Rubin, D. B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley. doi: 10.1002/9780470316696.

Spolaore, E., and R. Wacziarg. 2013. "How Deep are the Roots of Economic Development?" *Journal of Economic Literature* 51: 145. doi: 10.1257/jel.51.2.325.

Suppes, P. 1970. *A Probabilistic Theory of Causality*. Amsterdam: North-Holland Publishing.

Trigger, B. G. 2003. *Understanding Early Civilizations*. Cambridge: Cambridge University Press. doi: 10.1017/cbo9780511840630.

Turchin, P. 2003. *Complex Population Dynamics: A Theoretical/Empirical Synthesis*. Princeton, NJ: Princeton University Press. doi: 10.1515/9781400847280.

Turchin, P. 2016. Ultrasociety: How 10,000 Years of War Made Humans the Greatest Cooperators on Earth. Chaplin, CT: Beresta Books.

Turchin, P., R. Brennan, T. Currie, K. Feeney, P. Francois, D. Hoyer, J. Manning, A. Marciniak, D. Mullins, A. Palmisano, P. Peregrine, E. A. L. Turner, and H. Whitehouse. 2015. "Seshat: The Global History Databank." *Cliodynamics* 6: 77–107. doi: 10.21237/c7clio6127917.

Turchin, P., T. E. Currie, E. A. L. Turner, and S. Gavrilets. 2013. "War, Space, and the Evolution of Old World Complex Societies." *PNAS* 110:16384–16389. doi: 10.1073/pnas.1308825110.

Turchin, P., T. E. Currie, H. Whitehouse, P. Francois, K. Feeney, D. Mullins, D. Hoyer, C. Collins, S. Grohmann, P. Savage, G. Mendel-Gleason, E. Turner, A. Dupeyron, E. Cioni, J. Reddish, J. Levine, G. Jordan, E. Brandl, A. Williams, R. Cesaretti, M. Krueger, A. Ceccarelli, J. Figliulo-Rosswurm, P.-J. Tuan, P. Peregrine, A. Marciniak, J. Preiser-Kapeller, N. Kradin, A. Korotayev, A. Palmisano, D. Baker, J. Bidmead, P. Bol, D. Christian, C. Cook, A. Covey, G. Feinman, A. D. Juliusson, A. Kristinsson, J. Miksic, R. Mostern, C. Petrie, P. Rudiak-Gould, B. Ter Haar, V. Wallace, V. Mair, L. Xie, J. Baines, E. Bridges, J. Manning, B. Lockhart, A. Bogaard, and C. Spencer. 2018. "Quantitative historical analysis uncovers a single dimension of complexity that structures global variation in human social organization." *PNAS* 115: e144–e151. doi: 10.1073/pnas.1708800115.

Turchin, P., and S. P. Ellner. 2000. "Living on the Edge of Chaos: Population Dynamics of Fennoscandian Voles." *Ecology* 81: 3099–3116. doi: 10.2307/177404.

Turchin, P., and A. Korotayev. 2006. "Population Dynamics and Internal Warfare: a Reconsideration." *Social Science and History* 5 (2): 121–158.

Turchin, P., H. Whitehouse, P. Francois, E. Slingerland, and M. Collard. 2012. "A Historical Database of Sociocultural Evolution." *Cliodynamics* 3: 271–293.

van der Leeuw, S. E. 1981. "Information flows, flow structures, and the explanation of change in human institutions." In *Archaeological Approaches to the Study of Complexity*, edited by S. E. van der Leeuw, 229–329. Amsterdam: Cingula 6.

Volterra, V. 1926. "Fluctuations in the Abundance of a Species considered Mathematically." *Nature* 118: 558–600. doi: 10.1038/118558a0.

Watts, J., S. J. Greenhill, Q. D. Atkinson, T. E. Currie, J. Bulbulia, and R. D. Gray. 2015. "Broad supernatural punishment but not moralizing high gods precede the evolution of political complexity in Austronesia." *Proceedings of the Royal Society* B 282: 20142556. doi: 10.1098/rspb.2014.2556.

White, D. R., R. Feng, G. Gosti, and T. Oztan. 2011. "Easy R scripts for Two-Stage Least Squares, Instruments, Inferential Statistics and Latent Variables." *Sociological Methodology*.

Yuan, Y. C. 2010. Multiple Imputation for Missing Data: Concepts and New Development. SAS Institute.