Preface and Introduction by Douglas R. White

I. Introductions

- White, Douglas. Introduction to Transformations of Cross Cultural Research
- Ilkay, Altintas, Douglas R. White, Tolga Oztan, Workflows in Cross Cultural Research

II. Development of GeoPhylo Regression and Multiple Imputation Solutions

Eff on the prevalence of autocorrelation in cross-cultural data:


Dow and Eff on incorporating multiple imputation and network lags to estimate cross-cultural models:

- Eff, E. Anthon and Malcolm M. Dow. 2009. How to Deal with Missing Data and Galton’s Problem in Cross-Cultural Survey Research: A Primer for R.

III. Models and Scales

Substantive papers employing the Dow-Eff functions and Creation of Topical Scales for the Standard Cross-Cultural Sample:


IV. Dow on missing data imputation and DEf autocorrelation:
• Dow, Malcolm M. Missing Data Imputation (MI), With and Without Deletion (MID) of Data where the Dependent Variable is Missing.
• Dow, Malcolm M. Dow-Eff Two-Stage Least Squares for Wy and Tests of Autocorrelation Control.

V. Eff on Cross-Cultural datasets and 2SLS:
• Eff, E. Anthon. DEf01d Two-Stage Least Squares (2SLS) for Regression, Logistic, and Multiple Logistic Analysis.
• Eff, E. Anthon. DEf01d Computing Scales.

VI. White and Feng on on-line research and teaching:
• White, Douglas R. Bayesian Network Learning, Programming and Results.
• White, Douglas R. Neighbor-nets: Programming Distance and Evolutionary Results.
• Ren Feng, Online Classroom Teaching of Cross Cultural Research

VII. Fischer and others on access to ethnographic texts:
• Fischer, Michael. Automated Coding of Ethnographic Texts
• Fischer, Michael, Carol Ember, Douglas R. White. Combining Random Samples with Standard Cross-Cultural data and eHRAF texts.

VIII. van de Leuuw, Turchin and Kohler on larger perspectives:
• van de Leuuw. Sander. Big theory meets Big data.
• Turchin, Peter. Economic Growth Models meet Cross-Cultural data
• Kohler, Timothy A. Turchin’s Population and Warfare Oscillations: Archaeological Evidence in the Southwest.

IX. Oztan and White on Galton’s Problem and kin avoidances
• Oztan, B. Tolga and Douglas R. Are there Solutions to Tylor’s and Driver’s Evolutionary Puzzles of Avoidances?

X. Wallace on Mitochondrial genetics and cross-cultural data
• Wallace, Douglas. Mitochondrial Adaptations and Climate.

XI. Johnson on DEF01 and the Binford-Johnson datasets
• Johnson, Amber. Ecological Constraints among Hunter-Gatherer Societies: Does Cross-Cultural Modeling of the Binford Data improve with DEF?

XII. Roberts, Wynn, Acevedo, White, Oztan and Snarey on religion
• Acevedo, Gabriel. SCCS and Gendered Religions Institutions.
• Wynn, Eleanor and Douglas R. White. Computing Reincarnation Beliefs Across Cultures.

XIII. Other
• Chick, Gary. Sports and War
• de Munck, Victor. Romantic love

XIV. Conclusion
• TBA

Permission sought Dow and Eff as authors of three articles in Cross-Cultural Research. The various authors of articles in World Cultures and Structure and Dynamics each hold their own copyrights: as per the Wiley Contract statement below dated April 27, 2012 p2-3:

Wiley contract language, p2-3
MULTI EDITOR BOOK 2009 White Dow and Eff, Companion to Cross Cultural Research
If the Work contains extracts (including illustrations) from, or is based in whole or in
part on, other copyright works p3 (including, for the avoidance of doubt, material from on-line or intranet sources), the Editors, at the Editors' expense, and in the form specified by the Publisher, will obtain from the owners of the respective copyrights written permission to reproduce those extracts in the Work in all territories and editions and in all media of expression and languages or will procure such permissions from the Contributor.

All permission forms must be submitted to the Publisher on delivery of the Work. The Editors shall provide full credit to such works. Should the Editors be unable to obtain any permission in the required form the Publisher may, in its discretion, decide not to include the relevant material in the Work, or may clear such third-party rights and charge the costs thereof against royalties owed hereunder.

Chapters reprinted: down to 8 copyright permissions held by Companion editors Eff and Dow, no longer requested: 3 from the Journal of Cross-Cultural Research (Sage), 1 from the Journal of Social, Evolutionary, and Cultural Psychology.

Dow and Eff on the prevalence of autocorrelation in cross-cultural data:


Dow and Eff on incorporating multiple imputation and network lags to estimate cross-cultural models:


Substantive papers employing the Dow-Eff functions:


Papers related to data creation for the Standard Cross-Cultural Sample:

Introduction: Transforming Cross-Cultural Research

Humanity has evolved as a network but has not succeeded in knowing its own past. It is only proper that a transformation of cross-cultural research should play back its evolution, history, causalities and processes. In 1889 the configuration of issues later to be called Galton’s problem after Sir Francis Galton was first recognized. It raised the problem of how to identify adequate statistical methods that deal with different kinds of causality and explanation in cross-cultural research. As Eff stated:

In 1889, Edward Tylor presented what was to become the seminal paper in statistical cross-cultural analysis, before a panel at the Royal Anthropological Institute. Sitting on the panel was Sir Francis Galton, the statistician and eugenicist. Tylor compiled information on institutions of marriage and descent for 350 cultures and examined the correlations among these institutions.
The results showed that certain institutions were associated with each other far more often than chance would imply; Tylor interpreted these results as indications of a general evolutionary sequence, in which institutions changed focus from the maternal line to the paternal line. Galton disagreed, pointing out that *similarity between cultures could be due to borrowing, could be due to common descent, or could be due to evolutionary development; he maintained that without controlling for borrowing and common descent one cannot make valid inferences regarding evolutionary development.* In the literature, Galton’s critique has become the eponymous “Galton’s Problem.”

Galton’s problem appears in a peculiar form in regression analysis. The statistical technique requires that the disturbance [i.e., error] term in the estimated model have certain properties, one of which is that the disturbances not be correlated with each other. Violation of this property causes the estimated standard errors of the coefficients to be biased, so that one cannot trust the t-statistics, and one cannot make hypothesis tests regarding the estimated coefficients. (Eff 2004:153-143).

The problems that Galton diagnosed involving regression or causality (X “due to” Y or Z) as associated with problems of statistical analysis were much more difficult than they seemed: Malcolm Dow (Dow 2007), following Brandsma and Ketellepper (1979), only solved these problems a century later, as implemented by Anthon Eff.

One might think that the core of Galton’s problem is due to the effects of “clusters of separate homogeneities” (clusters due to borrowing, common descent, evolutionary development, etc.) on data analysis. Clustering, however, is the wrong concept: one that is not well defined (Kleinberg 2002) in terms of three essential properties (achievable partitions, consistency, and scale-invariance) that cannot all be achieved simultaneously. Clustering is thus poorly defined as a characteristic of groups of nodes in a network. Dow’s solution to Galton’s problem is not about clustering but a group-level property of the set of variables used in a regression equation, defined instead by four equations listed in section 2. The solution to Galton’s problem also depends on the extent to which the total regression model produced by those equations has an independent error term and passes statistical criteria as to whether the model is correctly specified.

Single variables also exhibit autocorrelation: in spatial statistics: Moran’s (1950) I coefficient is a measure of spatial autocorrelation characterized by “a correlation in a signal among nearby locations in space.” White (1993) and Eff (2004) have used Ord’s (1975) version of Moran’s I to show the very large extent of spatial autocorrelation in the variables of the Standard Cross-Cultural Sample. Eff (2004) calculated that roughly 44% of the variables in the Standard Cross-Cultural Sample dataset (Murdock and White 1969) were significantly autocorrelated; White’s (1993) results for an earlier year were similar. Such similarities may occur between pairs of cases or pairs of variables or in larger configurations, including clusters. Spatial clustering of homogeneities is only one of the forms of autocorrelation.

If cultures were completely independent, each a totally autochthonous invention, with no mutual knowledge, contact, or interaction with one another, there would be no examples of Galton’s problem: spatial or linguistic autocorrelations of single
variables would not be significant, and ordinary least squares (ols) for dependent variables predicted by other sets of variables in a dataset (given appropriate statistical tests) would have error terms uncorrelated with the independent variables. This scenario is nearly impossible to find in samples in which observed variables are recorded.

Patterns of autocorrelation include vertical inheritance and/or horizontal influences. Spatial autocorrelation is ubiquitous in observational studies, and for humans, linguistic autocorrelation (whether for sampled or complete observations and/or in a local or worldwide context). Samples of societies form networks of similarities partly due to historical or evolutionary origins from a common root, often found in language trees. Locational proximities facilitate, for each culture, the emulation of neighbors or nearby network centers.

Sociological and network literatures recognize confounded aspects of these problems, if without a time dimension, as between nonindependence due to matching traits (sociologically: homophily), social influence, and the causal effect of a society’s covariates on its characteristics (Shalizi, Thomas 2011). “Society” here is substituted for the authors’ term “individual” but “generically, all of these are confounded with each other”,”Symmetries in regression coefficients cannot identify causal effects and ... very simple models of imitation (a form of social contagion) can produce substantial correlations between” a society’s enduring traits and its preferences for affiliation with other societies. In cross-cultural research vertical inheritance studied through the time dimension of phylogenetic languages trees simplifies the problem. How confounded elements will play out for spatial autocorrelation remains to be seen.

Clustered homogeneities in observational samples were not a problem solved by survey researchers (Hansen et al. 1953, Kish 1965)? As recognized in survey sampling, clustered homogeneities are only those of purposely created subsamples with different sampling frequencies, weighted accordingly. What these weightings resolve for survey sampling is only the “effective sample size” problem that takes into account only the effects of how subsamples were constructed. In each case, some of the subsamples of a survey are purposefully organized as clusters of homogeneity relative to other clusters designed to be convenient for the aims and costs of the survey. Surveys do not attempt to measure empirical clusters of separate homogeneities that can only be evaluated after the survey results are available for analysis. Raoul Naroll (1961, 1965, 1970) took his clue from Hansen, and later from Kish (1965), in trying to define an “effective sample size” for cross-cultural samples by computing the likely effect of similarities between pairs of cultures. Murdock and White (1969) followed suite. These ad hoc measures, however, cannot solve a problem of “controlling for clusters of separate homogeneities in ways that provided corrections for autocorrelation” because (1) following Kleinberg (2002), such clusters are not logically well defined, and (2) the solution to Galton’s problem (Dow 2007), as implemented by Eff, are of a totally different order. Dow’s solutions to Galton’s problem are very general in that they
would also apply to results of survey samples and to what are called “experimental control groups” that act as protocols for epidemiological research and other fields of observational research.

This introduction reviews topics covered in the Wiley companion. Each is necessarily affected by autocorrelation. Yet, for a century, following Galton’s 1889 description of the problem, no “solution” existed for autocorrelation at the level of regression analysis (Moran’s autocorrelation coefficient was only for single variables). Simple in principle, resulting in a single computed variable added to the regression equation if it passes a series of statistical tests, Dow’s (2007) solution affects all cross-cultural and other observational studies.

The inclusion of research on human evolution using language-based phylogenies began in strength in the 1990s. These studies only partly solved the problem of autocorrelation by focusing on studies of vertical transmission. A fuller solution, that of GeoPhylo Regression (DEf, Dow-Eff or GeoPhyloR), as discussed in these chapters, did not emerge until published by Dow (2007) and Eff and Dow (2009), a 120 year time lag from the first recognition of Galton’s problem.

Given the uneven development of research on samples of observed variables, I discuss here how a generic approach to controlling autocorrelation in regression analysis developed by Dow (2007) applies in all aspects of cross-cultural research if generalized to other modeling methods (e.g., logit). Other fields of study can benefit from the DEf methods, which are essential to cross-cultural research because of the extent of the autocorrelation problem. This introduction considers aspects of our topic in succession that entail the need for a general “Transformation” of cross-cultural research. This has become my choice for this Wiley Companion’s title. Because comparisons of ethnographic descriptions are part of the basis of cross-cultural research (along with archaeology and evolutionary studies), this is where the introduction begins.

1. **Fieldwork vs Surveys, Cross-Cultural Research, & Controlled Experiments**

Fieldwork-based ethnographies typically trace out patterns of cultural sharing, in contrast to survey findings about populations where patches of cultural homogeneities are evident but typically ignored. Ethnographic field studies are essential to understanding the ways that cultures and communities differ in specific geographic spaces. An ethnographer’s time is often spent in minute observations and comparisons, as does learning how to compare the meanings behind cultural beliefs and behaviors. Comparisons over time include seasonality, annual, ritual and generational or intergenerational. Ways of understanding complex human behavior and symbolization are part of anthropological or sociopolitical fieldwork. Human societies exhibit both immense diversity and local or far-ranging homogeneities. Common history, evolutionary continuities, and interaction through proximity or communication produce multiple structures of homogeneity. Understanding how homogeneous features cluster, near and far, can help in comparative research even if they cannot in themselves contribute directly to the solution of Galton’s problem.
Results of fieldwork studies do not necessarily show only congruent patterns of cooperation or agreement but also veins of conflict, disagreements and oppositions forming schismatic patterns. "Culture" is not a harmonious whole but includes patterns of dissociation and competition, and also follows far-flung routes of dispersal and return. What constitutes fieldwork, whether in simple or complex economies and polities, is the business of discerning by observation and inference and learning from participation and conversational elements.

Cross-cultural comparisons, based on reading ethnographies, are done either in a holistic form or in a series of reports about topics that pertain to multiple demarcated social groups in their location and time of observation. Typically groups considered as cultural or multicultural units have close or distant interactions over time. More complex comparisons of the ethnography of dispersed societies, such as described by Wolf (1982), may deal with interactions of local groups embedded in larger or encompassing entities composed of political structures and networks of trade, conflict, exploitation, and resistance, and other interactive dimensions of regions and world systems. Wolf's approach is based on studying networks of interaction over time rather than inferences about variables coded independently in a sample of societies with little concern with their histories and interdependencies.

A cross-cultural sample of groups or societies is based on sociocultural groups located in space and time, such that one or many sets of ethnographic narratives may be read as largely converging descriptions or understandings of what is seen as occurring within and across identifiable boundaries of interaction with other groups, or similar patterns that occur over larger distances. Not all descriptions or understandings will be convergent. A coding project for cross-cultural research employs a strategy for coding variables that form part of the mosaic of types of ethnographic reporting and interpretation that can be abstracted at one extreme as discrete kinds of occurrences within some generic topics (e.g., kinds of kinship terms between specific relationally-defined relatives, or kinds of kinship groups) or in the opposite extreme as points or measures along a continuum (like degree of feuding between kinship groups). Kinds of political groups might be coded by multiple variables for levels of hierarchy and type of authority, for example. Variables may be defined by multiple authors for a particular sample of societies, or by one or a group of authors. They will differ in type or quality (e.g., high vs. low inference) or type of measure (such as single variables versus scales constructed from multiple variables). They may differ as to whether they distinguish features of the ethnographers (a source for data quality tests or discovery of biases) or of the societies studied.

The model (or understanding) of cultures is not one of ephemeral entities. Rather, it includes what Boyd et al. (1997) call coherent memes, with little recombination and slow change; or, that "small elements (words, innovations, components of ritual practices), are linked together in larger, potentially transmittable entities (technological systems, myth, religion), which themselves are collected into
“cultures” that characterize human groups of different scales (kin groups, villages, ethnic groups, and so on). These units can cross-cut one another, and thus the analyst must be explicit in defining the unit being used” (O’Brien et al. 2014:54-55). The end points for two contrastive models of culture, then, are hierarchical systems and cultures as assemblages of many coherent units.

Understanding “structures of homogeneity” is one way that anthropologists have drawn comparisons across cultures, even if this concept does not help solve Galton’s problem. Kroeber, for example, was biased in his belief that shared traits in a local area or region were indices of common histories and their gradual divergences or migratory penetrations, such that one could reconstruct cultural histories from correlations and migrations. In contrast and contention, Driver and Jorgenson examined local cultural similarities (Q analysis correlations) and correlations among variables (R analysis correlations) to argue for spread, evolution, causal, functional and other kinds of cross-cultural findings. They were acutely aware of Galton’s (1889) problem that the statistical significance of correlations, which might reflect causal or functional relationships, are often vastly inflated when clusters of separate homogeneities are present. They did not, however, solve Galton’s problem. Nor do correlational results, scaling, regression analysis (ols), and other statistical or experimental results. Multidimensional scaling and correspondence analysis, for example, simply beg many of the questions as to which are clusters of causal effects, which are functional relationships, which are due to what were initially spatial similarities, or similar due to preferential homophilies?

To achieve the more difficult problems in the social sciences is the operational basis of this book, as applied to cross-cultural and most other forms of observational research. In this sense a solution for Galton’s problem provides a transformation of research practices.

Is it possible, from the data of cross-cultural research, to piece together the processes of human evolution, the alternative developments of social organization and structure, modes of behavior, sexuality, systems of belief, morality, religion, etc., topic after topic? Evolutionary anthropologists have begun to answer such questions. Hamilton et al. (2007), for example, testing the scalar stress hypothesis of Gregory Johnson (1982), found fractal self-similarity of hunter-gatherer group sizes (N=283) that for the nonAustralian foragers tend to approximate multiples of four, 4 being the average size of the nuclear family; 16, that of extended families; 64, a typical size of “dispersed extended family ... groups during the most dispersed phases of the mobility cycle”; 256, “larger groups that are dispersed over larger areas and interact with decreasing frequency, but serve to maintain social ties, conduct trade and information exchange, perform ceremonies and exchange marriage partners;” 1024, “periodic aggregations defined as multi-group socio-economic aggregations occurring at periods usually greater than every year;” 4092, “regional populations defined as the total size of regional ethnic units.” These definitions are taken from Binford (2001). It is not that these branching ratios are evolutionary progressions, since our primate ancestors would also have had
embeddings of social networks at different spatio-temporal levels, although with different aspects of sexual sociality: e.g., monkeys with maternal-based lineages or higher primates with residential cores of males and outmigrating females. These are structural regularities brought on by specific processes. There are also homogeneous variants found in Hamilton et al.’s study of fractal structure. They found a distinctive homogeneous cluster of 56 Australian forager societies, however, where the fractal coefficient is closer to 6, the average size of the polygynous family. This speaks to geographic separation producing alternative evolutionary paths. It is likely that the causes of these two sets of variant findings could be identified statistically if tested for additional distance and language homogeneities using solutions to Galton’s problem using GeoPhylo Regression, described below. There is also a tantalizing pair of clustered homogeneities shown in Hamilton et al.’s supplementary material (table and graphic) for the distributions of foragers on the dimensions of branching ratio and energy availability (higher for the many Australia foragers). This minor pattern was not investigated further.

Programmable solutions (“DEf”) for cross-cultural research were provided by Dow and Eff (Dow 2007, Dow and Eff 2009, Eff 2008, 2009), along with tutorials and applications (Brown and Eff 2010, Dow and Eff 2013, Eff and Dow 2008, 2010, Eff and Rionero 2011). Cross-cultural, along with all other research based on observational data share the generic autocorrelation problems of observational research, to which DEf (GeoPhylo Regression if more descriptively labeled) is posed as a general solution.

2. GeoPhylo Regression with Imputation (Dow-Eff functions in R; GeoPhyloR)
   Kepler Project: Doug White, Tolga Oztan (UC Irvine)
   Ilkay Altintas, Jianwu Wang, Tahereh Masoumi (UC San Diego)
   Malcolm Dow (Northwestern University), Anton Eff (MTSU)
   To get at the problem of nonindependence or Galton’s problem in cross-cultural, experimental, or other observed-variable research, five basic elements in the observations of cases and outcomes of analysis are taken into account. In each case the observations:
   1. exist in space (e.g., a set of coordinates)
   2. exist in history (e.g., devolving from ancestral predecessors, ethnic or language groups, etc.)
   In terms of outcomes, if 1) and 2) are controlled, GeoPhylo Regression may produce:
   3. statistically valid probabilistic Bayesian-like predictors with use of the R library(bnlearn)
   4. alternative probabilistic bnlearn models (Nagarajan, Scutari and Lèbre 2013) further examined by R library(bootstrap) analysis of R library(bnlearn) outcomes.
   5. statistically valid approaches to the problem of imputing missing variables (Rubin 1987, 1996) using principal component analysis that may be used to predict missing values of variable provide researchers with completely-coded variables for their samples.}

6
GeoPhylo Regression with Imputation (Dow-Eff functions in R or GeoPhyloR) analysis aims to restore independent identically distributed (iid) error terms in ordinary least squares (ols) regression using n by n row normalized proximity (W) matrices to define the network of closeness in geography and ancestry for pairs of observations in a sample of n observations (e.g., cultures). It uses a 1st stage ols where y is the dependent variable, X is a matrix of independent variables, W is an n-n weighted proximity matrix with row elements divided by row sums. Then an estimated W*y ~ c + X*W + ε defines a new estimated variable Wy=W*y. Adding Wy to the original ols variables corrects for inflated significance tests in the 2nd stage ols and gives results corrected for autocorrelation, which is ubiquitous in observed variable studies. This approach can be used in any of the sciences in which observed variables are analyzed. More information at: http://intersci.ss.uci.edu/wiki/index.php/DEf.

Useable representative sample datasets are SCCS, LRB, WNAI, EA, XC and AWC (to receive data from SCCS and EA). Once a dataset is selected, the Dependent Variable, Unrestricted Variables and Independent Variables are changed accordingly. Go to http://intersci.ss.uci.edu/wiki/index.php/Dow-Eff_Codebooks to search for variables for each dataset.

3a. Dow-Eff modeling briefly described

In Dow-Eff modeling as a solution to Galton’s problem, there is a first-order OLS regression that identifies a single variable, Wy, which is an autocorrelation predictor specific to a particular regression (ols, logit etc.) model. A second-order OLS then includes independent and potential covariate (“Unrestricted”) variables plus the estimated Wy predictor. Wy and the independent variables have different Rsquared predictions. Model quality is evaluated by inferential statistics. Wy and each independent variable are also evaluated by significance tests. It is not uncommon that the predominant percentage of variance is predicted by autocorrelation (Wy). That does not diminish the relative importance of other independent variable predictors. Some models may be found to have no effect of autocorrelation. In the four DEF equations below, W in Eqn. 4 is a ols weighted-composite, the Xs are columns of independent variables Xi, y is the dependent variable, and estimates from equation 1 are marked by dot above Wy and constants in equation 2 and over Wy in equation 3. Wy and WX are matrix products.

\[ W_y = \alpha_0 + \alpha_i (WX_{i=1,n}) \]  
(Eqn 1).

\[ \dot{W}_y = \alpha_0 + \hat{\alpha}_i (WX_{i=1,n}) \]  
(Eqn 2).

The estimated result is added to the raw independent variables in a second-stage least squares (2sls) regression:

\[ y = \beta_0(\dot{W}_y) + \beta_1 + \sum_{i=2,n+1} (\beta_i X_i) + \epsilon \]  
(Eqn 3).

The W in the equation 1 is derived from solution of a biparametric spatial lag
regression model (Brandsma and Ketellepper 1979; Dow 1983; Dow 2007) that calculates weighted additive effects of distance (D), language phylogeny (L) and potentially other W matrices on the dependent variable y (Eff 2008a):

\[ W_y = X\beta + \lambda_L W_L + \lambda_D W_D + \epsilon \quad \text{(Eqn 4)}. \]

I provide those four equations here, although they appear in chapters by Dow and Eff, so that the average reader can see that these solutions are complex. There is no way to write a back-of-the-envelope solution to autocorrelation, as one commentator suggested. Galton’s autocorrelation problem can be considered to be clustered separately around each society’s similar neighborhoods and ancestries. Accordingly, the process of controlling for autocorrelation is to construct proxy independent variable values of closer spatial neighbors and linguistic ancestries, weight these two proxies to create the best composite proxy for an autocorrelation-based dependent variable, and create the estimated Wy proxy independent variable values to be subtracted from the actual dependent variable (as in equation 3) to define a regression with a estimated Wy control for autocorrelation.7

This book provides instructions to access four cross-cultural research datasets and possibly more (section 12. ) that are consistent with GeoPhylo Regression with Imputation (GeoPhyloR=DEf). The GeoPhyloR in the alternate name for DEf is a reminder that DEf solves for relative weightings of spatial autocorrelation and for phylogenetic autocorrelation. Instructions for data and software with easily accessible modeling inputs are given for options of 1) the R Dow-Eff functions, 2) the Complex Social Science (Co-SSci) Gateway, and 3) the Kepler science workflows (SWFs) for GeoPhylo Regression (Kepler GeoPhyloR). It is this complex of computational tools that provides access for the Transformation of Cross-Cultural Research. New observational samples may be produced for GeoPhyloR and Dow-Eff functions and added to the Co-SSci databases if they conform to these general features, are of general interest, and suitable for DEf/GeoPhyloR analyses.

3b. Complex Social Science (CoSSci) Galaxy and Gateway
The CoSSci Complex Social Science Galaxy and Gateway provides access to DEf (Dow-Eff GeoPhyloR) data and software to researchers, teachers and students in classrooms that have internet access so as to connect to cross-cultural analysis methods orf variables in the databases discussed in 12. A preliminary on-line syllabus is provided at http://intersci.ss.uci.edu/wiki/index.php/Class_instructions. Access to the Galaxy is free, and provides ultra-fast access to a UC Irvine computer.

4. Concepts, Scales of Comparison, Critique of Hermeneutics
Without new statistical tools like GeoPhylo Regression (DEf) we cannot have confidence in many of the things we believe to be true about world cultural or regional comparisons because the framework for analysis that is used has not attended to the fact that societies and cultures are not historically and spatially independent.8 The only type of study with cross-cultural data that I exempted from
Galton’s problem involved studies of the sexual division of labor where Guttman scales were evident and were evaluated by signal detection methods.\(^9\)

In 1982 Dow, Burton and I published two articles on what turned out to reproduce, not surprisingly, the extent of the problem of nonindependence of cases (Galton’s problem) in cross-cultural research. Nonetheless I founded and started to publish, the year that Murdock died, issues of the on-line journal World Cultures in hopes that the SCCS data could be complemented with new methods to deal with corrections for nonindependence. We published another study of effects of nonindependence in 1984, adding Karl Reitz as coauthor, but it offered no solution to Galton’s problem of controls for nonindependence. I published in 1986 the bibliography of ethnographies recommended for coding new projects for the SCCS and in 1988 published my first comparative study using the SCCS, hoping to show that polygyny was a far more complex topic to undertake than admitted by Murdock. Murdock, aiming at reducing the extent of autocorrelation, had coded the breakpoint between limited polygyny and general polygyny at 20%, which reduced the apparent prevalence of general polygyny and reducing the extent of Galton’s problem. I included focal dates for all 186 societies, and a total of 20 dichotomous, ordinal and Guttman scale variables dealing with polygyny and monogamy, including the exact or estimated percentage of men and women polygynously married, without doing further analysis. A score of societies had considerable missing data.

The accomplishment of co-creating the SCCS was fully realized after the development of the Dow-Eff software in identifying problems that when solved would provide for valid results in of cross-cultural research results. Many of Dow and Eff’s articles were published in UC eScholarship open access journals, *World Cultures* and *Structure and Dynamics*. In an article devoted to how to create composite scales, Dow and Eff (2013) used my monogamy variable as their dependent variable in their Social, Evolutionary, and Cultural Psychology article and based their analysis entirely on Eff’s scaling methods, which are a part of DEfR (GeoPhyloR) not yet implemented in the online modeling site at CoSSci.

With controls for autocorrelation, four of the 12 composite scales in Dow and Eff’s (2013) study were independent variable predicting monogamy: low pathogen\_stress, low violence\_stress, hi division\_of\_labor specialization, and hi beneficial\_environment, all positive effects favoring the first wife and that women are decision makers for both the option of monogamy and that of becoming a polygynous wife. None of the following scales had predictive effects for monogamy: male/female resource inequality, female contribution to subsistence, socially imposed monogamy, societal scale, political participation, famine stress, modern economy, and autocorrelation with religion. The Rsquared for Wy and the four significant variables was 0.288, of which 0.170 came from the four dependent variables, not terribly impressive except for the direction of the relationship predicted. Only one finding agreed with evolutionary generalizations by Alexander et al. (1979), that of occupational specialization, which was positively associated
with monogamy and lent positive status to women if we consider males as more specialized than women. Two findings disagreed in the direction of the statistical relationship: those of 2) pathogen stress and 3) harsh environments, contradicting the evolutionists’ view that monogamy occurs under conditions of environmental harshness because this orients men to be “less likely to favor more than one wife.” Their finding 4), beneficial environment, could not be matched to correspond with the arguments of Alexander’s evolutionary biology group. This type of conflicting results might be considered a common occurrence when using Dow-Eff controls for autocorrelation because researchers will often be misled by correlations that do not control for Galton’s problem.

Dow and Eff’s (2014) draft of a new Wiley chapter, “The Network Autocorrelation Approach to Comparative Method: Monogamy in the Pre-industrial World,” improves on their 2013 article on monogamy in using, among other methodological improvements (e.g., bootstrap standard errors), a Box-Cox power-law transformation as their monogamy dependent variable with a scalar power-law parameter λ. These improvements net them a major improvement of Rsq (0.453 over 0.288) and two additional independent variables: technological level/societal scale (p=.008) and modernization/modern economy (p=0.038). The distance/language autocorrelation effects change from 0.50/0.50 to 0.78/0.22, i.e., greater effect of proximity commensurate with modernization as distance-clustered.

With the development of Dow-Eff modeling, the analysis of religion has been an area of great promise for cross-cultural research, with new and solid findings with GeoPhyro Regression (DEf) that contradict earlier findings. These topics are among the chapters of the Companion. In each case the divergence of results might be expected from comparing studies that control for autocorrelation with those that do not. In other words, it is likely to turn out that here, as elsewhere, many of the classical results of cross-cultural research will be overturned given corrections for Galton’s problem. With collaboration of coauthors, excellent results of GeoPhyro Regression were derived for three domains of religion, with dependent variables such as: Beliefs in Reincarnation (a scale composed from many diverse sources), Moral Gods (Swanson 1960, more fully coded by Snarey 1996), discussed in the next section, and Evolution of Gods (Sanderson and Roberts 2008, coding a scale developed by A.F.C Wallace 1996). Wallace (Chapter 2) proposed a fourfold taxonomy of conglomerative religions that tend to be cumulative rather than mutually exclusive: “shamanic religions, containing only shamanic and individualistic cult institutions; communal religions, containing communal, shamanic, and individualistic cult institutions; Olympian religions (polytheistic), containing an Olympian variety of ecclesiastical institutions, along with the other three; and monotheistic religion, with a monotheistic ecclesiastical cult institution, together with communal, shamanic, and individualistic cult institutions” (courtesy Wesley Roberts). The code implies cumulative rather than the mutually exclusive religions of the Swanson (1960) variable.
The A.F.C. Wallace approach to religion is a good place at which to recognize that some of our readers are epistemological relativists, while in contrast (Spiro 1986: 270):

... a scientific generalization is a statement of the kind not that monotheism is universal – and not only because it is false – but that wherever and whenever monotheism occurs, it sustains a systematic relationship to the occurrence of some other specified social or cultural condition [conditions]
By the application of Mill’s canons of similarity and difference, ethnographic studies of different religious belief systems can ascertain to what degree, if any, such regularities obtain.
Taking these generalizations as paradigmatic, it will now be notices ... that (1) scientific generalizations regarding society or culture are statements not of frequencies but of regularities and (2) although these regularities are of universal scope, they pertain not to universals but to differences in social and cultural conditions (in this case religious belief systems).

Spiro (p271) concludes that epistemological relativism is not a descriptive approach but based on prescription against anthropology as a scientific discipline.

Wallace’s (1996) view as a comparativist goes further in a descriptive vein, recognizing that monotheism is not mutually exclusive of many other types, but rather that once a given religion evolves, such as shamanism, it may still be practiced (in some form) in Los Angeles (even if loosely borrowed), and even if excoriated by priests. Yet other cross-culturalists, like Swanson (1960), may consider that religions might be defined exclusively as normative types.

Spiro (1986:271-279) had gone further than Wallace, however, in discussion of the issues of science versus the hermeneutics argument that separates science as the study of “nature” from the study of culture (e.g., Rabinow and Sullivan 1979). Spiro, like Dewey (1929), rejects the notion that atoms, trees and stars are part of nature while mind, culture and human experience are not: a physical/human science dichotomy does not entail a natural/hermeneutic dichotomy: “the former dichotomy is substantive, the latter epistemic.” The claim that meanings cannot be causes contradict the opportunity for cultural meaning systems may “create cultural entities, direct one to do certain things, and evoke certain feelings” (D’Andrade 1984:96-101). “I would claim that the opposition between the hermeneutic tradition has erected between interpretation and explanation is yet another false dichotomy”; “the study of meaning does not entail that anthropology cannot be a scientific discipline” (Spiro p.273). His following 6-page critique of “the Hermeneutic Agenda of Epistemological Relativist” and of its deleterious effects on anthropology as a social science is well worth reading to grasp the merits of comparative research and its current transformation.

The study of kin avoidances is another classical comparative topic. A world sample (between 3-400 societies) of kin avoidances was one of the original topics studied by Tylor (1889) that was the object of of Sir Francis Galton’s statistical criticism because of dense clusters of societies with distinctive homogeneous clusters of features. In discussing this problem, Jorgensen and Driver were so befuddled by this problem that they estimated that it might take sample sizes of 1-2 thousand (Harris 2001:616) to get an effective sample size that would yield statistically significant
results with the packing of so many similar societies in the same regions, especially in Tylor’s extensive inventory of avoidances in Australia.

The chapter here by Oztan and White, in both the SCCS world sample and the Western North America (WNAl) sample, give totally different results not only than those of Tylor but also contradict those discussed by Murdock (1949:273):

Societies fall into two groups with respect to the manner in which they handle incest and other sexual taboos. One group ... [Murdock’s: has] taboos taboos ... so thoroughly instilled by precept and sanctions that they become “second nature.” ... The other group ... apparently succeeds less well in internalizing sexual prohibitions. ... and are compelled to bulwark these with precautionary safeguards such as avoidance roles.

These prejudices and prejudgments are not supported by cross-cultural evidence, such as provided in his (1971) article on avoidances, or as I have coded for his data the 200 societies in his “sampling provinces” worksheets (Murdock 1968) from which the SCCS societies were chosen. Using these data, Driver’s (1965) avoidance codes combined with the extensive WNAl dataset of Jorgensen (1980), and Murdock’s unpublished codes on cross-sex kin behaviors (catalogued by White circa 1970 and referenced as Murdock n.d.), Oztan and White were able to analyze kin avoidances in the SCCS sample and among North American Indians in the other.

Goody and Buckley (1974) noted that throughout the world provinces identified by Murdock, pairs of cross-sex kin avoidances have "nearly identical distribution," perhaps a strong signal that these behaviors once indexed very widely distributed human behaviors. Oztan and White’s Wiley chapter show avoidances to have been an evolutionary step toward complex societies with more broadly extended marriage networks and more complex political alliances. Interestingly, with autocorrelation controls, we were able to replicate very similar findings in both the Western North America and SCCS world samples. Variables from other datasets were used to test Murdock’s hypotheses but no support was obtained for his Yankee-oriented psychological or psychoanalytic hypotheses.

Autocorrelation often distorts statistical findings, especially in exaggerating significance tests, as was the case with Tylor’s study of avoidances (Murdock’s errors as the sexual unconscious were of a different order). They were a far greater problem than many social scientists bargained for. Even the survey sample theorists who calculated "effective sample size" in attempting to correct for inflated significance tests failed to grapple directly with Galton’s problem, which was never fully solved in the sampling literature. For anthropologists the techniques of survey sample statisticians -- adjusting "effective sample size" to a lower number -- was called, by the 1960s, "Galton's Problem" (Naroll 1961) although it was not a solution to Galton’s problem.

5. NA options, Hausman and Wald tests, Residuals, and “To Try” as Diagnostics for Improving GeoPhylo Regression Models: An Experiment for Moral Gods

NA is the R value for missing data. It’s use is not recommended as a means of improving DEf regression models. The example here for SCCS dependent variable
2007, Moral Gods, is exceptional. It was defined but very sparsely coded by Swanson (1960), coded by Murdock for SCCS (v238 “High Gods,” N=170, with 18 cases left as missing) and the Ethnographic Atlas (N=181 in a sample that had 748 coded cases), and fully coded by for the SCCS by Snarey (1996: v2007 "Moral Gods") along with presence of religious missions and various predictive variables. It is a particularly fascinating but also a troubling variable concerning the evolution of religion but nicely illustrative of problems and options for modeling. It is also important as an illustration of Bayesian a prior theory versus ad hoc choice of variables based on both regressions based or not based on GeoPhylo Regression (DEf). Many articles have analyzed Murdock’s “High Gods,” with rather poor results and theory. Brown and Eff (2010) use DEf regression building on theory about large societies (e.g., reviewed by Atkinson et al. 2014), adding several multivariate “large society” scales using Dow and Eff’s (2003) scale construction techniques.

As a Bayesian theoretical basis for building a causal model of morality, evolutionary biologist Alexander (1987: 240) states “this is how moral systems have always operated; the expense of being selfish is why extreme altruism has prevailed within groups in conjunction with severe extrinsic threats.” Biologist turned evolutionary historian Turchin (2005) showed for periods when societies fail that sociopolitical instability and collapse of empires, when due to prior periods of severe economic inequality caused by overpopulation relative to resources, are followed by periods in which religious and prosocial movements often engage in moral reconstruction to counteract the inequalities of failed economic exchange.

Figure 1 takes Turchin’s hypothesis to the level of Bayesian network learning (Nagarajan, Scutari and Lêbre 2013) with R library(bnlearn) graph results for a test of independent variables predicting that moral gods may appear in economic exchange systems having periods of scarcity\(^{10}\) indexed by FxCmtyXWages (fixed communities multiplied by wages: in a Turchin cycle, overpopulation lowers wages while holders of resources benefit from inflation) or AnimXwealth (associating the Silk Road form of trade with camels, horses, and bridewealth exchange, where trader lineages holding pastoral wealth benefit from animals in bridewealth exchanges – the multiplicative variable – that helps to recruit reproductively valuable wives from the lineages of service workers). FxCmtyWages prove to be a predictor of the large set of Christian societies together with a smaller set of more agrarian Islamic societies coded as having Moral Gods, while AnimXwealth is a predictor of a smaller set of pastoral Islamic societies engaged in trade. \textit{This leaves only 10 societies outside Islamic and Christian SCCS societies that are coded as having Moral Gods,} eight of which have Missions (Snarey 1996) at the pinpointing date of observation. When these ten are recoded as NA (missing data – an unusual step), only three were classed in the imputed variable output as likely to have Moral-God beliefs: the Amhara, Darfur, and Masai, all in Africa. Of these only the Masai had no Missions. Although pastoral, the Masai have few camels or horses, so they would be excluded from the predictors of Moral God beliefs. This exclusion gives a new twist to identification of Moral Gods as a category of religion limited to the 46 Abrahamic societies in the sample: Hebrews, Christians and Mohammedans. No other study of
this variable has seen this connection. The imputed sample after imputing the nonAbrahamic societies gives an Rsq=.38 (DpV v2007==4 is not predicted as a dichotomous but as an ordinal dependent variable)\(^1\), a far lower Rsq than for the Evolution of Moral Gods v2013 GeoPhylo Regression where Rsq=.78.

GeoPhylo Regression has an Rsq = 0.25 predictor of the variance in Moral Gods (v2007) and key diagnostics are not significant for various types of model defects (although the Shapiro-Wilkes test does not pass H0: residuals normal), as shown in the tests listed in the paragraph below:

\[
\begin{array}{cccc}
\text{Fstat} & \text{df} & \text{pvalue} & \text{star} \\
0.7246 & 88378 & 0.3947 & \text{n.s.} \\
0.2291 & 65135 & 0.6322 & \text{n.s.} \\
2.5772 & 22950 & 0.1084 & \text{n.s.} \\
\end{array}
\]

Baysian graphic results using library(bnlearn) for the Moral Gods variable are shown in Figure 1 for a few of the variables evaluated prior to the NA testing results. If there is potential causality among these variables (undirected lines with Bayesian priors that define direction of potential causality) it is that FxCmtyWages (Fixed Communities with Wages) and AnimXbwealth (Animal Husbandry with Bridewealth) are the predictors of Moral Gods (SCCS v2007. Higod4) while the lower (nontropical) rainfall variable connects to more complex societies with writing, and writing to religious missions.\(^{11}\)

---

\(^1\) Does the dichotomous variable “Christianity and Islam” versus other religions have predictors like AnimXbwealth and FxCmtyXBwealth? No: only with growth of moral scale.
Figure 1. A small set of variables evaluated by the Nagarajan, Scutari and Lèbre (2013) R library (bnlearn)

Strong *a priori* (Bayesian) theory is of course preferred in cross-cultural research, backed by Hausman, Wald, and other inferential statistical tests, such as whether residual error terms are uncorrelated with independences, as in other sciences.

“To Try”, however, is one of the GeoPhylo Regression (DEf) outputs that are reported in the *.csv output. It lists those variables, if any, which may improve statistical results with regressors that are statistically significant and that add to Rsq predictions, drawing from the entire dataset. Often there are either zero or many candidate variables given by “To Try” options, and each needs to be evaluated as independent in meaning from the dependent variable and in terms of a conceptually strong potential predictor.

Results in Table 1, with additional predictors added to those in Figure 1, shows for Moral Gods how ols regressions jump from Rsq=.28 to .38 with GeoPhylo Regression and its inclusion of the Wy correction for autocorrelation.

<table>
<thead>
<tr>
<th>Table 1: Moral God SCCS variable v2007 for ols and (DEf01d) models with imputation, the latter controlling for autocorrelation with the computed variable Wy; defined by Guy Swanson (1960) and coded John Snarey (1996). Not all variables in the models (^DEf) are significant according to the code: *** p &lt; 0.001 ** p &lt; 0.01 * p &lt; 0.05 . p &lt; 0.10 n.s. These summaries do not include other statistical tests produced by Dow-Eff functions (DEf).</th>
</tr>
</thead>
</table>
| 1 ols <- lm(v2007 ~ SuperjhWriting + FxCmtyWages) #R-squared: 0.1021 ***
2 ols <- lm(v2007 ~ AnimXbwealth + NorainDry) #R-squared: 0.2139 ***
3 ols <- lm(v2007 ~ SuperjhWriting + FxCmtyWages + AnimXbwealth + NorainDry) #R-squared: 0.2844 * (best result)
4 ^DEf<- lm(v2007 ~ Wy + SuperjhWriting + FxCmtyWages) #R-squared: 0.3235 *** |

What is required for a DEf result is the *imputed Wy* <- h$data[, "Wy"] even though Wy is not imputed.

19
5 \^DEf<-lm(v2007 ~ Wy+SuperjhWriting+FxCmtyWages+AnimXbwealth+NorainDry) #R-squared: 0.3819 *** - ****
6 DEF<-lm(v2007 ~ Wy + AnimXbwealth + NorainDry) #R-squared: 0.3646 **** **** <-No variables nonsignificant.
7 DEF<-lm(v2007 ~ Wy+SuperjhWriting+AnimXbwealth+NorainDry) #R-squared: 0.3784 *****(best result) <-No variables nonsignificant.
8 DEF<-lm(v2007 ~ Wy+FxCmtyWages+AnimXbwealth+NorainDry) #R-squared: 0.3729 **** ** <-No variables nonsignificant.

All variables are significant for six of the eight models, and four of these have a Wy term. In such cases the Bayesian Model Averaging (Raftery, Hoeting, Volinsky, Painter, Yeung 2014) is useful: "Bayesian Model Averaging is a technique designed to help account for the uncertainty inherent in the model selection process (Wintle et al., 2003), something which traditional statistical analysis often neglects. By averaging over many different competing models, BMA incorporates model uncertainty into conclusions about parameters and prediction. BMA has been applied successfully to many statistical model classes including linear regression, generalized linear models, Cox regression models, and discrete graphical models, in all cases improving predictive performance." BMA "has become widely accepted as a way of accounting for model uncertainty, notably in regression models for identifying the determinants of economic growth (Eicher et al. 2011).

Regressions such as those in Table 1 appear at the end of a series of workflows -- successive processing starting with raw variables with missing data, then data imputation, then modeling predictors of dependent variables, model evaluation, then a multivariate analysis that checks in graphical form for potential causal networks within the over set of variables used in the earlier stages of modeling, as in Figure 1, then the simple ols models in Table 1 using the fully imputed variables, and then the DEf regression modeling in Table 1, using the autocorrelation variable Wy computed by the Dow-Eff functions. This can be done directly in R, but for researchers and students, as well as classroom settings, this is all accomplished more easily with the open access site Complex Social Science Galaxy constructed for the Wiley project by Argonne National Labs (and University of Chicago) computer scientists who design Galaxy interfaces for the sciences with NSF funding through the ECSS Extended Collaborative Support consortium, which is a feature of XSEDE, the Extreme Science and Engineering Discovery Environment that is funded by NSF for the sciences, of which Anthropology is now a member. Here, we had ECSS support from Lukasz Lucinski, Thomas Uram and Ráchana Ananthakrishnan. Virtual Computers operate in the use of the CoSSci open access resources at no cost for this project, and make the initial steps in modeling as simple as making variable-name entries in a screen with windows for the dependent variable, independent variables, other variables to be explored, and new variables that are constructed from ones already in the database. Maps of variables are their autocorrelation clusters are provided according to the names of variables. Higher-order tests like those in Table 1 are currently calculated in R but are being programmed in workflows with help of the Kepler project team (2. ) at UCSD, again pro bono as this project is one of their
user groups. In a later stage of CoSSci workflow development, these steps will be provided through workflows at high performance computing sites when needed for modeling with large numbers of variables.

There have been many studies of variants of the Moral Gods variable, the v2007 (Higod4) variable representing a recode of all the SCCS societies for Murdock’s Ethnographic Atlas coding of this variable (SCCS v238. High Gods). Studies using v238 include an excellent article by Brown and Eff (2010) which builds on earlier publications that view moral god religions like Islam and Christianity as predicted by independent variables that emphasize large and highly populated societies. Our SCCS v2007 model is one that emphasizes the theory that large societies with exchange systems that go through historical oscillations in which economic collapses occur periodically (Turchin 2005), and the independent variables are those predictive of complex economic exchange system indicators of potential economic collapse. In that sense these to models predict opposing aspects of Turchin’s historical model of complex societies. GeoPhyloR (DEf) modeling does not yet contain functions for path analysis, but when path analysis is added, an unpublished study by White and his PhD students (White, Oztan, Gosti and Feng n.d.) indicates that the “large systems” model variables are mediated by the “potential collapse” variables of the Turchin model. If this is correct, the Brown-Eff model has Rsq=0.578 while the Turchin model has Rsq=0.250 yet both are correct in terms of potential path analysis model (using DEf imputed variables) where the latter model is the primary predictor. Hopefully a path analysis component will eventually be added to the GeoPhyloR Kepler workflows (Altintas et al. 2004, Ludäscher et al. 2005) if not directly to DEf.

6. Comparison Post-1969 SCCS
An estimate of the severity of Galton’s problem discussed by Dow, Burton, and White (1982:197), shown with simulations that resemble cross-cultural theory test data (White 2003, Eff 2004), is that higher significance levels are increasingly exaggerated even in small samples (N=30) and even maximum likelihood procedures of estimating significance tests must be used very conservatively. These results were ignored by Ember and Ember (1998:678): "We suggest that those who worry about Galton's problem misunderstand two requirements of statistical inference -- that sample cases be independent and that the measures on them should be independent (Kish 1965, 1987...). Independence of cases means only that the choice of one case is not influenced by the choice of any other (which random sampling guarantees)." This is doubly false: Kish (1965, 1987), the Dean of survey sampling, was greatly concerned with statistical corrections for clustering of similar cases drawn in samples, which is a problem of autocorrelation; and random sampling does not prevent the clustering of similar cases, as is well known to spatial geography and survey sampling, and nicely demonstrated in Murdock and White. Loflin disabused many researchers of Ember’s belief that probability samples (e.g., HRAF's 60 society samples) are free of Galton’s problem, and also that the SCCS sample is free of autocorrelation. Ember’s assumptions conflict with and ignore Loflin’s (1972) statistical explanations and the foundational bases of statistics.
In the last quarter of the century, in spite of resistance to recognizing Galton’s problem, many researchers contributed coded data to the SCCS database. Between 1970 and 2000 archaeologist Lewis Binford (2001) finished 30 years of work in publishing a monumental cross-cultural study of 339 forager societies. The database included five hundred coded variables, with more from and to come from his colleague Amber Johnson. Jorgensen (1980) followed suit, focusing on Western American Indians, analyzing relations among societies as well as among variables in the Driver tradition (Driver and Massey 1957).

The object of this Companion, then, is to finally put comparative research on a solid footing rather than in the limbus of ad-hoc methods that so often result in wrong answers to statistical and research questions. It is better to alert readers to a multitude of misconceptions here and in the larger literature as to how to control the statistical errors created by autocorrelation. There is no advantage to the researcher to hide behind false solutions such as random samples, ignoring all the corrections to Galton’s problem that are actually necessary. As we will see in the chapters by Dow and Eff, the issues and statistical computations are complex, although not difficult to implement with proper software. The proper type of software, GeoPhylo Regression (DEF), is not available in commercial packages such as SPSS and Stata and would be difficult to implement correctly because each dataset requires new matrices for solution of autocorrelation and new summary principal component datasets are needed to solve imputation for missing data.

7. **Autocorrelation controls and HRAF/eHRAF ethnographic texts and samples**

Ember and Ember (1998:679) state: "recent mathematical and computer solutions don’t require a special sampling strategy (which is true), nor do they require expensive, time-consuming controls." Rather, "all you have to do is test statistically for the possibility that proximity or common ancestry accounts for a result (Dow et al. 1984; Burton et al. 1996)." Statistical analysis that would "account for" autocorrelation, however, is neither provided nor claimed in either the Dow et al. or Burton studies cited, so this dismissal of autocorrelation is misleading. Finding clusters is one thing; controls for how they affect regressions and other models, or affect significance tests, is quite another. Solutions are all the more urgent as the depth, quality, and number of cross-cultural databases advances. The errors, misconceptions and facile dismissals that HRAF has disseminated in the cross-cultural research community do not mean that the Wiley Companion to Transformation of Cross-Cultural Research is at odds with HRAF: to the contrary. Although HRAF has no expertise in inferential statistics it is an invaluable repository of ethnographic text, and it has created two valuable 60-society probability samples, which together make up a larger probability sample. Considerable textual data is available for these samples that is retrievable from the HRAF text files, including the eHRAF electronic files. "Searching across cultures for particular kinds of information is facilitated by the unique indexing system (a controlled vocabulary) that HRAF has developed and refined over more than 50 years, the Outline of Cultural Materials or OCM (Wikipedia:Human Relations Area Files 2014)," which builds on work by
Murdock et al (1938). "In contrast to most subject-indexing which is done at the document level, HRAF has its indexers subject index at the paragraph level."

Paragraph retrieval by OCM topics can be coordinated with either or combined HRAF samples.

In 2014 HRAF adopted a proposal to improve the interface between texts and SCCS coded variables by filling in the roughly 1/3rd of the societies in each of its probability samples with closely equivalent SCCS societies, replacing cases that were not in SCCS while retaining the properties of the probability samples (regional, language, culture area). In coming years, the ethnographic texts for these 40 to 50 SCCS societies will be added to the eHRAF sample. Comparativists will then be able to locate more textual data useful in coding new variables for both the SCCS and the two probability samples, resulting in two new and one combined achievable HRAF sample datasets and the potential for new variables for cases in the SCCS.

8. Development of DEf (GeoPhylo Regression with Imputation: GeoPhyloR)

From 2007 to the end of the decade in which the Wiley Companion is published, the goal here is to see the potential quality of Cross-Cultural Research transformed, from faulty reliance on descriptive statistics which failed to deal with Galton’s problem to advanced inferential statistics that control for autocorrelation efficiently (Dow 2007), as implemented by Eff, using two stages of OLS, and which imputes missing values of variables. Eff and Dow’s (2009) first primer for R set off a flurry of research amongst the authors preparing this Wiley Companion so as to help provide an open access site for modeling cross-cultural data with a sophisticated OLS regression that could be tested for adequacy of its solution to Galton’s problem. We were able to put together a fuller set of databases and codebooks, examples, manuals, teaching strategies, and the background data to support imputation and solutions to autocorrelation.

This book turns in coming chapters to authors Dow and Eff, statisticians and programmers of the R software that underlies the present "Transformation of Cross-Cultural Research." As we advance to further chapters we will have examples of cross-cultural modeling (regression, logistic analysis, plausible causal modeling for networks of variables, Akaike’s Information Criterion (AICc), possibly observed variable path analysis and temporal panel analysis), all of which will depend on basic key outputs (imputed variables and statistical tests) of the Dow-Eff models (GeoPhylo Regression).

9. Culture and Phylogeny

Biological anthropologists (e.g., Borgerhoff Mulder, Boyd, Nunn, O’Brien et al.), archaeologists (Buchanan), and British anthropologists excelled in the 1990s and the following decade in exploring evolutionary phylogenetics in cross-cultural research. Given a spate of phylogenetic anthropological studies, leading researchers Mace and Pagel (1994) argued that "cross-cultural comparative studies must be based upon the identification of independent events of cultural change. Once this principle is applied, it becomes apparent that it is in fact groups of closely related
cultures that are potentially the most informative for testing cross-cultural hypotheses. Constructing phylogenies of cultures and placing upon them independent instances of cultural elements arising or changing is an essential part of this task.” The books of Shennan (2002) on genes, memes and archaeology, the 13 chapters by different authors in Mace, Holden, and Shennan (2005), and numerous other articles lay out this view in detail.

Language family phylogenetics plays a role in tracing contrastive models of societal history, as does Shennan (2002:95-98) in tracing the Austronesian language family. Here "The pattern apparent in linguistic relationships are integrally tied to the movements, contacts, and activities of language speakers" (Kirsch and Green p.1054), although "the genetic information is not so clear" (Shennan 2002:277), reflecting archaic features "that must have originated in eastern Indonesia."

Figure 2 shows for the 'express train' model (Gray and Jordan 2000) tracing evolutionary trajectories that follow linguistic spread from Taiwan-Philippines in a seaward radiation to SE Asia, Madagascar, Melanesia, Polynesia and New Zealand, as against the 'entangled bank' model (Kirsch and Green 2001), who argue from triangulation of archaeological, ethnographic and linguistic evidence that proximity in the language tree says very little about geographic proximity.

![Figure 2: Express-train model of Austronesian Colonization of the Pacific (after Gray and Jordan 2000; adapted from Brien 2014). 1; Taiwan, 2; Philippines, Chamorro, Palau; 3, Borneo, Indonesia, Malay; 4, Sulawesi; 5, central Malayo-Polynesian; 6, Southern Halmahera/western New Guinea; 7, Near Oceania; 8, Remote Oceania; 9, central Polynesia; 20, eastern Polynesia. Bryant et al. (2005) explains NeighbourNet analysis, with various examples, as a network construction that can recover known reticulations in linguistic data. Between 1994 and 2005 (Mark Pagel 1994; Pagel, Meade and Barker 2004; Pagel and Meade 2005), developed "continuous" and "discrete" algorithms for phylogenies used in the Mace, Holden, and Shennan (2005) articles. Five of these chapters use Bayesian analysis for phylogenetics, including the Gray and Jordan (2000) approach above. Greenhill and Gray (2005) compare the two competing Austronesian hypotheses using MCMC Bayesian phylogenetic analysis as opposed to NeighbourNet analysis. Their conclusion is that while the "entangled bank"
If the hypothesis is rejected, there are four other competing models for the Austronesian language family phylogenies that they propose to evaluate with better data.

Phylogenetics for multiple cultural features is a much harder problem than a simple one-variable phylogeny using Neighbor-Net, e.g., for a single language group. That of Indo-European (I-E) is shown in Figure 3. Fortunato and Mace (2009) investigate the likely phylogenetic evolution of polygyny-monogamy, wealth transfer, dowry, and bride wealth for I-E and find that “dowry with monogamy represents the most likely ancestral state,” supported by the separate temporal evolution of bride wealth and dowry. Bride wealth with polygyny and dowry with monogamy represent relatively stable states of coevolution (p245). Diverging “[d]irect vertical transmission” of property from parents and children took the form of both inheritance and dowry for daughters; “dowry in particular served as a way to control their status in marriage to husbands of appropriate standing (Goody 1976)” (p246). They suggest that further aspects of Goody’s theory can be investigated using “Bayesian MCMC inference, reviewed in Huelsenbeck et al. 2001; and Holder and Lewis 2003.”

**Figure 3**: Phylogenetic diagram of the Indo-European Language Family, from Ethnologue (Grimes 2002). Reprinted with permission from Eff (2004:30). Many of these societies are in Murdock’s Ethnographic Atlas (EA) and Atlas of World Cultures (AWC:1961). Fortunato and Mace (2009) find evidence for monogamy at the Turkish-region root societies of the Indo-European language family.

“Once language trees have been constructed, they can be used to examine other aspects of culture” (Brien et al. 2014:49). Examples given are those of Curie et al. (2010) and Curie and Mace (2011) for societies in Southeast Asia and the Pacific. Fitting a dynamical modeling of phylogenetic rise and fall of ethnolinguistic groups coded in Murdock’s Atlas for the variable Jurisdictional Hierarchy Beyond the Local, which varies from 1 to 4 (which they interpret as band -> tribe -> chiefdom -> state). Hierarchy rose and fell in both that order and reverse order, but only in jumps up or down a single level, never 2 or 3, skipping levels. The model held up well when fitted
to the branch tips of ancestral phylogenies. The 2011 study showed co-evolution with wider hereditary social stratification in neighboring societies.

Bayesian methods for the construction of phylogenetic network models are now widely used for comparative linguistics. Contemporary reviews of computational phylogenetics discuss phylogenetic methods that provide exchange protocols (Bast 2013), as did Sullivan and Joyce (2005) a decade earlier. These overviews suggest appropriate ways to do model selection for distance matrices and phylogeny reconstruction for both linguistic and genetic phylogeny as well as visualization of multi-route spatial autocorrelation. Nichols and Warnow (2008) provide useful tutorials.

A classic phylogenetic result is the finding of Holden and Mace (2003, 2005) for Africa, tracing the Bantu language phylogeny, namely that cattle drive out matriliny resulting in patriliny or other forms of descent. In the next section we attempt to replicate those results with GeoPhylo Regression (DEf), which has a phylogenetic component.

Can phylogenetic findings be replicated with GeoPhylo Regression (GeoPhyloR/DEf)? Here we explore this question by attempting a replication of Holden and Mace’s (2003 “Spread of Cattle led to the loss of matrilineal descent in Africa”) and, more specifically, their (2004:220-223) findings in: “The Cow is the Enemy of Matrlny.” For Africa, cattle appear to drive out matriliny in 26 of 30 cases of cattle in the Bantu language family, contrasted to other forms of descent in 18 non-matrilineal cases out of 38 with no cattle (Fisher exact test, 2-tailed= 0.0009). It is important to note that the SCCS sample (N=186) is less suited than a larger sample (e.g., AWC:1961) for the task of replicating Holden and Mace’s (2003, 2005) phylogenetic results with GeoPhylo Regression (DEf) because the SCCS sample is purposefully thinned out to minimize autocorrelation while remaining a representative world sample.

Figure 4 replicates Holden and Mace’s (n.d.) map of Bantu language family migrations. The latter’s 20 societies with matriliny and no cattle are mostly in the tropical forest central zone (Red in Holden and Mace n.d.). Just under those cases are four matrilineal cases (Black in n.d.) with cattle. Northwesterly (Green in n.d.) are 21 societies with patriliny and no cattle. In the Northeast and South (Blue in Holden and Mace n.d.) are those with patriliny and cattle.
**Figure 4:** Bantu migrations relevant to “The Cow is the Enemy of Matriliny” in Bantu Africa (Holden and Mace n.d.). Original figures show Green (Northwest) and Blue colors (Northeast and South) the societies with Patriliny and “Mixed or No Cattle” or “Mixed or Cattle”, respectively and show in Red (central zone) and Black (lowest zone) Matriliny with “No Cattle” and “Cattle,” respectively.

Table 4 shows results of an attempt at replication of Holden and Mace’s (2004) “The Cow is the Enemy of Matriliny” model using data from the Ethnographic Atlas (EA) of 1248 societies. Among the EA (V=166) variables are codes for matriliny as a form of descent and for bovines (which includes cattle) as a prominent type of animal husbandry. The results of GeoPhylo Regression (DEF) are shown in Table 2. Bovines are, indeed a (temporal) predictor of loss of matriliny, controlling for autocorrelation, which is present and highly significant at \( p_{val} < 0.0001 \), on a par with Holden and Mace’s (2003, 2005) findings for Africa. In Table 2, however, this finding holds for all 1248 world societies in the Ethnographic Atlas.

**Table 2:** GeoPhyloR model predictors controlling for autocorrelation for Matriliny.
Ethnographic Atlas: “The Cow is the Enemy of Matriliny” run at smi 5,7
Source: MatrilinealEA8.xls.png
For the GeoPhylo Regression (DEf) project (Transforming Cross-Cultural Research), these are exciting findings. One of the big questions of evolutionary theory is whether certain features evolve by vertical (e.g., language phylogeny) or horizontal transmission, and here we have an answer for matrilineality: while 80% of the autocorrelation in these data are predicted by language phylogeny, 20% are predicted by geographic proximities, i.e., distance. This is extremely important because we can measure the extent to which vertical autocorrelation (language phylogeny) occurs in comparison to horizontal autocorrelation (distance or ecology). In this case there is no effect of similar environments on autocorrelation (the Wy term as explicated in section 2. GeoPhylo Regression). The Wy autocorrelation factor (a single predictive variable) is the most significant effect (pval, hcpval, and bootval <0.00001) in the model results. These findings open further questions about under what conditions do the erasures of matriliny follow a spatial trajectory versus directive changes within language families. Here, the capability of GeoPhylo Regression (DEf) to produce maps – still in development – can be very suggestive and lead to further statistical analyses.

My version of “The Cow is the Enemy of Matrliny” in Table 2 holds across the entire globe, not just for the Bantu language family phylogeny of Africa. DEf, unlike Holden and Mace’s (2003, 2005) phylogenetic analyses, also searches for additional variables using Eff’s “To Try” DEf function, which lists potential independent variables. In this case precipitation in that part of the year with warmest climate (bio.18) is a positive predictor of matrilyn as is lower Net Primary Productivity variable, indexing regions of refuge for matrilyn. After controlling for autocorrelation (variable Wy, combining a 80% of vertical language phylogeny with 20% of horizontal distance effects), the GeoPhyloR model in Table 2 shows Rsq=0.24 for regressions that might influence evolutionary effects of bio.18, mnnpp, and bovines on the stability or disappearance of matrilyn.
How do language phylogeny models relate to the phylogeny component of DEf? The first can handle any number of variables located in the language tree so that single and multiple variables may appear on the tree and there may be transitions such as “Cows are the Enemy of Matriliny,” or combined sets of variables that have a probabilistic effect on what come next in the tree. The trees of GeoPhylo Regression with Imputation (DEf; GeoPhyloR), however, depend on the specific set of variables in the Regression, dependent and independent. Table 2, for example, has not only the Bovines variable but also regional primary production and a climate variable. With a very large sample, all the predictors are highly significant.

Analyzing the same model as in Table 2 (EA sample: Matriliny dependent variable), this time with SCCS variables in Table 3, we get a different result: the predictors of Matriliny include more variables (Matrilocality, Low Kin Group Vengeance, and Female contribution to agriculture), and a higher Rsq=0.271 but there is no autocorrelation among the set of variables in the model.

Table 3: ‘Female contribution to agriculture & Matrilocality are Friends of Matriliny’

<table>
<thead>
<tr>
<th>Rmodel</th>
<th>coef</th>
<th>stdcoef</th>
<th>VIF</th>
<th>relimp</th>
<th>pval</th>
<th>hcpval</th>
<th>star</th>
<th>desc</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.9242</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>0.0424</td>
<td>0.0015</td>
<td>**</td>
<td>NA</td>
</tr>
<tr>
<td>Bovines</td>
<td>-0.0213</td>
<td>-0.0509</td>
<td>1.1857</td>
<td>0.0105</td>
<td>0.4646</td>
<td>0.3594</td>
<td></td>
<td>Low Bovines</td>
</tr>
<tr>
<td>Matrilocality</td>
<td>0.3897</td>
<td>0.4540</td>
<td>1.0366</td>
<td>0.2165</td>
<td>0.0000</td>
<td>0.0000</td>
<td>***</td>
<td>Matrilocality</td>
</tr>
<tr>
<td>v2008</td>
<td>0.0262</td>
<td>0.1318</td>
<td>1.0364</td>
<td>0.0224</td>
<td>0.0434</td>
<td>0.0143</td>
<td>**</td>
<td>Low Kin Group Vengeance</td>
</tr>
<tr>
<td>v821</td>
<td>0.0018</td>
<td>0.1215</td>
<td>1.1609</td>
<td>0.0201</td>
<td>0.0904</td>
<td>0.0894</td>
<td>*</td>
<td>Female Contribution to Agriculture</td>
</tr>
<tr>
<td>Wy</td>
<td>-0.0154</td>
<td>-0.0027</td>
<td>1.1493</td>
<td>0.0016</td>
<td>0.9695</td>
<td>0.9529</td>
<td></td>
<td>Network lag term</td>
</tr>
</tbody>
</table>

Variable v2008 in this model reflects a key problem with Bovines: they are easily stolen, which activates Kin Group Vengeance, mostly taken by males, hence inimical to Matriliny, but it is not that Cattle are directly the Enemy of Matriliny (not a significant effect). Matriliny is supported by Female Contribution to Agriculture and Matrilocality. Here Matriliny predictors have Rsq=0.271. Yet this model is not commensurate with language phylogeny since the Wy is insignificant. GeoPhyloR has the capability of going further variables linked to a given model. Suppose we posit and test for linked variables where “Cattle are the Enemy of Matriliny” not directly but indirectly, as in No Bovines (Low Kin Group Vengeance) → allow female contribution to agriculture → (Table 3) matriliny, while subsequently (following the language tree phylogeny): Bovines → lowered female contribution to agriculture → low matriliny. Here we can learn evolutionary sequences from the language tree.
Table 4: ‘Bovines are the Enemy of Female contribution to agriculture’

<table>
<thead>
<tr>
<th>Rmodel</th>
<th>coef</th>
<th>stdcof</th>
<th>VIF</th>
<th>relimp</th>
<th>pval</th>
<th>hcpval</th>
<th>star</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-52.1959</td>
<td>NA</td>
<td>NA</td>
<td>0.0032</td>
<td>0.0001*</td>
<td></td>
<td></td>
<td>(Intercept)</td>
</tr>
<tr>
<td>Bovines</td>
<td>-4.0950</td>
<td>-0.1477</td>
<td>1.1568</td>
<td>0.0592</td>
<td>0.0539</td>
<td>0.0265*</td>
<td></td>
<td>No Bovines</td>
</tr>
<tr>
<td>v2008</td>
<td>-2.3740</td>
<td>-0.1712</td>
<td>1.0660</td>
<td>0.0237</td>
<td>0.0176</td>
<td>0.0187**</td>
<td></td>
<td>Low Kin Group Vengeance</td>
</tr>
<tr>
<td>Wy</td>
<td>2.6000</td>
<td>0.4710</td>
<td>1.1613</td>
<td>0.2275</td>
<td>0.0000</td>
<td>0.0000***</td>
<td></td>
<td>Network lag term</td>
</tr>
</tbody>
</table>

The causal structure of variables between Bovines and Matriliny is more complex than attested by Holden and Mace (2004). This set of independent variables also warn against too-simple conclusion relating phylogenetic findings and causality, but Tables 3 and 4 show examples where models of changes (and temporal order) in evolutionary language phylogenies (ELP) are supplementary to GeoPhyloR.

The fact that in ELP and GeoPhyloR as used here the data pertain to the same representative sampling universe is potentially importance if these two methods are jointly examined. If would be ideal if the larger but not too large AWC Atlas of World Cultures were populated with SCCS variables (N=562 V=166 EA + SCCS 1943) that could be used in imputation of missing values (the EA Ethnographic Atlas sample is too large for imputations for N=186 SCCS cases to an EA sample of 2109). That would allow ELP and GeoPhyloR methods to be examined jointly.

I discuss what might appear as a failed replication of the Holden and Mace (2004) phylogenetic findings possible due to a flawed sample that is not representative of any particular sampling universe in Sections 13. and 22. Appendix B. “Notes on a failed model replication due to lack of shared representative sample” might illustrate why dataset XC fails to replicate a phylogenetic model predicting matriliny but those issues are more complex.

11. NeighbourNet/Neighbor-Net Diagrams
Greenhill and Gray (2005:50) depart from ordinary phylogenetic analysis in testing their population dispersal hypotheses with NeighbourNet analysis, an agglomerative method for the construction of planar phylogenetic networks (Bryant, and Moulton 2004, 2005). They argue that the "entangled bank" hypothesis about Austronesian languages "is not an adequate explanation of the data" but also that 'express train' model is not a strong conflicting signal but a lack of clear signal in the dataset. NeighbourNet analysis was further developed by (Huson and Bryant 2013), with core ideas of reticulation developed into SplitsTree4 analysis (Huson
and Bryant 2013), which computes both planar and potentially unrooted phylogenetic networks ("neighbor-nets"). Rooted trees are a special case of neighbor-nets, which we use here because they are relatively simple to implement. The Nexus input format (Maddison, Swofford, Maddison 1997) is extremely simple.15 Neighbor-nets are in some cases an improvement on taking the rooted tree as axiomatic to phylogenetics (some trees might have a cluster or roots) and is an approach that may easily permeate outside of Britain.

Some North American anthropologists have adopted phylogenetic methods, although many in the biogenetic sciences use specialized software to model genetic phylogenies. Waddell and Tan (2013) provide an example that includes genetic contrasts between humans, Neanderthals and other archaic humanoids. Highly regarded model selection methods include likelihood ratio tests (LRTs) of goodness of fit between models and data, provided that care is taken to avoid order of model comparisons, and Akaike information criteria (AIC), which can be used to find an optimal balance between over-fitted and under-fitted models. The most widely google-cited application in evolutionary studies has evolved from NeighbourNet (Huson and Bryant 2006) to their SplitsTree4 (2013) analyses that build off of neighbor-net algorithms. The former’s abstract, however, notes that “even when evolution proceeds in a tree-like manner, analysis of the data may not be best served by using methods that enforce a tree structure but rather by a richer visualization of the data to evaluate its properties, at least as an essential first step. Thus, phylogenetic networks should be employed when reticulate events such as hybridization, horizontal gene transfer, recombination, or gene duplication and loss are believed to be involved, and, even in the absence of such events, phylogenetic networks have a useful role to play. [Their] article reviews the terminology used for phylogenetic networks and covers both split networks and reticulate networks, how they are defined, and how they can be interpreted. Additionally, the article outlines the beginnings of a comprehensive statistical framework for applying split network methods. [They] show how split networks can represent confidence sets of trees and introduce a conservative statistical test for whether the conflicting signal in a network is treelike. Finally, this article describes a new program, SplitsTree4, an interactive and comprehensive tool for inferring different types of phylogenetic networks from sequences, distances, and trees”.16

Chapter authors in the Wiley Companion are able to use Huson and Bryant’s updated (2013) SplitsTree4 software17, as I do here, to analyze square intersocietal (W) matrices, as used in GeoPhylo Regression (DEF) to examine the structure of distance, language and ecological autocorrelation. Since these W matrices are designed for complete datasets, only one set of computations is needed.18 (That is, there is no need in the data structure to normalize square W matrices to row sums of 1, as with Wy). The DEF Wy method of potential control for autocorrelation could be supplemented by neighbor-net analysis of language phylogeny matrices, plus similarity data on global distance and ecological data, available from DEF datasets and R code. Dow-Eff model variables can also be used to create distance matrices, getting a neighbor-net or SplitsTree structure and possibly even remapping it geographically. Our first use for visualizing proximal distance autocorrelation is shown below in Figure 5 for 186 SCCS societies.
Figure 5 shows that distance replicates the geographic interconnectivity structure in location of societies, showing multiple shortest paths between them, but would not be expected to have a phylogenetic structure. Rather, one sees highly connected blocks of the closer distances, as opposed to outlying regions. SplitsTree can also apply to groups of variables in the DEf datasets and chapter-level studies. There are goodness of fit tests for NeighbourNet as well as SplitsTree structures. Although similarity data can be constructed and evaluated for appropriate phylogenetic analysis on a world scale, as in Figure 5, global models of substantive phylogenies (e.g., political, economic, religions, social structure, and other or finer-level composite datasets) may not have the precision of the British phylogenetic research using regional data. As more cross-cultural and linguistic data become available, however, that difference in precision could begin to disappear. It is important to note, however, that a global neighbor-net of languages will have multiple disconnected trees that show language autocorrelation in different regions.

Because DEf W matrices are computed for all societies in an entire dataset, and they are not model specific, one efficient way to use the distance (as in Figure 5), language and ecological similarity data is to compute them only once during a single run of the software and compute and save the output using Huson and Bryant's (2006) enhanced SplitsTree (2013) software for each variable separately and possibly even in pairs or for two or three variables. This provides the potential for a visualization and a metric for comparison of how other variables or packets of variables are keyed to phylogenetic networks. This is already done in constructing the Wy control variable used in DEf, and can provide a measure of the effective sample size controlling for autocorrelation, but can also measure and show where more detailed phylogenetic modeling is required, as in the British research tradition.
Figure 5: SCCS Neighbor-net of Intersocietal Distances (computed using SplitsTree)

Figure 6 shows the SCCS language tree in phylogenetic form, with prominent language groups for northeast Asia, Africa, and the New World, with the latter separated by a single rather arbitrary link to the rest of the world.
Figure 6: SCCS language phylogeny coded by Anthon Eff from Ethnologue world languages and graphed by SplitsTree4 Neighbor-net.

Figure 7 shows much larger trees for similar environments than for language phyla, but there are few proximities among them. That entails that there should be very few “clustered homogeneities” for ecologies.
Figure 7: SCCS eco-phylogeny for 19 BioClim variables were created using BIOCLIM variables http://fennerschool.anu.edu.au/publications/software/anuclim.php or those from used by Worldclim http://www.worldclim.org/. They are graphed by SplitsTree4 Neighbor-net but visual inspection shows no regional clustering in spite of densities of environmental tree-like clusters of that exceed the densities shown for language in the SCCS.

11b. Convex hull autocorrelation clusters
Figure 8 shows the result of Anthon Eff’s mkmapng function for SCCS v238 (Moral Gods as coded by Murdock for Swanson’s Higod variable as modeled with DEf by Brown and Eff (2010).
Figure 8: Visual representation of Brown and Eff’s (2010) results for the Moral Gods DEf regression showing convex hulls of five societal regions with separate multiple contiguous regions of autocorrelation. Convex hull regions in the Circum-Mediterranean show concentrations of Christian and Islamic moral god monotheism, those in Oceana, and North and South America show no high gods or high gods with no moral engagement with humans.

12. DEf/GeoPhylo Regression Datasets: SCCS, LRB, WNAI, EA, AWC, eHRAF
GeoPhyloR (DEf) offers a true transformation of cross-cultural research not only because of a generic solution to Galton’s autocorrelation problem and the imputation of coded values of variables for societies where these values are missing but because of results such as those in Table 2, which employs the Ethnographic Atlas (EA) variables to test a model obtained from Language Phylogenetics.

In the list of datasets below, codebooks are listed by URLs, and those that are simply proposed (AWC, eHRAF) are listed with bold type:
https://dl.dropboxusercontent.com/u/9256203/SCCScodebook.txt
https://dl.dropboxusercontent.com/u/9256203/LRBcodebook.txt
https://dl.dropboxusercontent.com/u/9256203/WNAIcodebook.txt
https://dl.dropboxusercontent.com/u/9256203/EAcodebook.txt
•XC (N=371 V=2657) is an amalgam of variables from the EA with additional societies from at least two other datasets from SCCS, LRB, and WNAI19. Superficially it is not a representative sample (See Section 13. ): e.g., 198 of its societies (53%) are in North America, but Eff argues that DEf (GeoPhyloR) has the effect of shrinking
those 198, on the basis of similarities, into much fewer types of societies that may be representative of pre-modern societies. That type of reduction should also hold for the EA and its reduced subset, SCCS. One can expect some anomalies in XC results, however, if certain kinds of variables are very poorly represented in some regions, such as matriliny in Africa. Rather than argue “no use of the XC sample” for researchers or students because of violation of representativeness one can also argue that if XC were expanded to a larger XC2 sample that reduces the criteria to EA plus ONE or more variable from SCCS, LRB and WNAI then XC2 would have a larger net of variables more likely to include those variables that would otherwise be missing (in section 22. B, in Table 5), for example, where very few matrilineal societies happened to be available in the full XC sample).

*AWC(+SCCS) (Atlas of World Cultures N=562 V=166+1943), selected by Murdock (1981) as a high-quality EA subset with matching EA society identifiers, could be amalgamated with SCCS variables to be suitable for doing phylogenetic analysis.

*eHRAF (electronic Human Relations Area Files) this could combine the two HRAF 60 society random samples into a single 120 society sample and would be more useful for societies not yet in HRAF if closely similar societies in the SCCS sample are substituted for those cases so as to create a version with complete SCCS codes.).

Note that random sampling in choice of societies, as for the eHRAF samples, does not solve for Galton’s problem. Substitution of SCCS societies that closely resemble societies in the HRAF random samples to replace societies that are not already in the HRAF sample does not bias the results of DEf analyses.

*Supplementary variables for all samples. All of these samples are or can be supplied with supplementary climatic, socioeconomic, ecological, soil, geographic, landcover, landforms, primary productivity and other data assembled by Anthon Eff.

Returning to possibilities of a new AWC sample, Murdock (1981:5) selected the AWC sample as: “a selected sample of this coded material [variables in the N=1254 EA sample] in a form designed to be maximally useful to scholars in the behavioral sciences,” i.e., a subset of best described representative societies in the EA. These societies “presumably number the majority of those whose cultures are most fully described in the ethnographic literature.” Creating the AWC dataset is proposed for completion before the Wiley Companion volume is published. This would enlarge, for example, the number of variables that are likely to be added to those in Table 2 as predictors of matrilineality, answering the question, for example: are other variables besides cattle involved in the decline of matriliny? (This is already partly answered in Table 3.)

Anthon Eff and I have worked to nominate cases for a prospective AWC sample that has 562 societies in the EA and 173 in the SCCS, 166 of which match the coded EA data. The ratio 29% or 166/572 is much better for purposes of imputation. A new AWC would improve on the fact that most of our researchers and nearly all of our students have avoided use of our XSEDE Science Gateway, for which the Complex Social Science website, a Galaxy site that directs the researcher or student either to one of two fast virtual machines at UC Irvine, building models for SCCS variables in
1.5 minutes, or to the UCSD/SDSC Trestles computer (with HPC high performance computing with 20 second job completion but a queue of 20 minutes). An AWC sample would open new possibilities for HPC computing with cross-cultural data.

The attraction of virtual machines is that one can do many of the successive steps in DEf modeling in a series of 1½ minute steps, and be done within a class period. Because the simpler modeling work is iterative, adding more or sounder variables step by step, few researchers or students use the HPC facilities at present. With a new AWC, sample, designed as a large dataset in terms of number of societies and number of variables, the time to job completion for Trestles would be much quicker than for the use of Dow and Eff’s native R or the virtual machines at UC Irvine. With AWC the expansion of the statistical power of DEf (GeoPhyloR with Imputation) would enable the student or researcher to expand Table 2 and discover, by yet another avenue, whether there are additional variables associated with increase or decline of matriliny, and whether GeoPhyloR can replicate language-based phylogenetics as a source of evolutionary modeling.

Ideally, since the AWC sample includes the best-coded societies of the EA and includes many of the SCCS societies that are in that portion of the EA, this strategy may work because the expanded AWC would be a best-coded subsample of the EA. GeoPhyloR models for variables in the AWC should take only 10-20 minutes to compute on CoSSci. This would easily permit students to test replications of phylogenetic models (such as those of Holden and Mace 2004) during class time.

The AWC subset could prove to be an important addition to the DEf or GeoPhyloR project of “Transforming Cross-Cultural Research” thanks to the combination of providing: 1) imputation of missing values for variables that require or merit them, 2) the solution for autocorrelation, and 3) optimal expansion of variables by imputation from the commensurate data of both the Standard Cross-Cultural Sample (SCCS) and the Ethnographic Atlas (EA).

13. Principles of Constructing Representative Cross-Cultural Samples
A representative sample is subset of a statistical population that reflects the members of an entire population, and in the present context, of known and well described societies. Large representative samples are ideal in cross-cultural research where “representation” has to do with the type of society: e.g., world cultural samples (EA, SCCS), heavily prestate with their capitals, such as Rome, Angkor or Cuzko, included proportionally; cross-national samples; forager societies (LRB); regional samples (WNAI); or random samples of one of these kinds, including the eHRAF world samples or combined sample. When samples are composed of a mix of different types of societies, such as the XC sample of EA, SCCS, LRB and WNAI, samples might be weighted to achieve representation (Kish 1965, DuMouchel and Duncan 1983) but whether weighting is best done by region or other selection criteria is open to question. E.g., should the XC sample (where all societies selected have EA variables) be weighted by ratios of societies from SCCS, WNAI, or LRB
alone, or including combinations SCCS-WNAI, SCCS-LRB, WNAI-LRB and SCCS-WNAI-LRB.

“Survey weights are numbers associated with the respondents that specify the weight or influence the various observations should have in analysis. **Weights are used to adjust for unequal probabilities of selection in the survey sampling design**, unequal response rates between subsets of the intended sample, differences between the population covered by the survey frame or list and the target population, and differences between sample estimates and desired numerical controls. **The final survey weight can usually be thought of as a measure of the number of population units represented by the response.**”

http://www.esourceresearch.org/tabid/374/default.aspx

Methods of calculation are found in DuMouchel and Duncan (1983). But does failure to use sampling weights lead to biased results in GeoPhylo Regression? Is that the cause of the Table 5 results in Appendix B (Notes on a failed model replication due to lack of shared representative sample)?

Or would GeoPhyloR (DEf) normally compensate for unequal weighting of regions in a dataset like XC? Would a slight alteration of the XC dataset, the XC2 sample (adding societies from EA that are also in one or more of the SCCS, LRB and WNAI samples) solve the problem of missing cases for the matriliney variable? -- as discussed in 12.

The Ethnographic Atlas (EA) dataset is particularly well suited to testing varieties of autocorrelation because its 1248 societies span the entire globe. The EA and SCCS are representative and commensurate world samples. For many variables DEf should “shrink” societies in EA to results more akin to SCCS with a much smaller “effective sample size,” closer to N=175.

It might be optimal, for expanded the variables in the EA sample, to add the variables from the V=2109 coded variables in the N=186 SCCS into an enlarged Ethnographic Atlas (EA) sample, expanding its limited but well coded V=166 variables to something over V=2200 variables, including supplementary climatic, socioeconomic, ecological, soil, geographic, landcover, landforms, primary productivity and other data assembled by Anthon Eff.

The problem with this might be that the capacity for imputation of missing data is limited because there are only 183 societies in SCCS coded for EA variables out of 1248. This leaves only 14% of this expanded sample from which to make imputations.

If we need to expand the variables of the EA sample used in Table 2 to include additional variables from the SCCS, however, Anthon Eff and I have worked to nominate in our list of datasets below a new prospective sample, AWC (Atlas of World Cultures; Murdock 1981) which has 562 societies that are in the EA and 173 in the SCCS, 166 of which match the coded EA data. The ratio 29% or 166/572 is much better for purposes of imputation.

Murdock (1981:5) selected the AWC sample as: “a selected sample of this coded material [variables in the N=1254 EA sample] in a form designed to be maximally useful to scholars in the behavioral sciences,” i.e., a subset of best described representative societies in the EA. These societies “presumably number the majority of those whose cultures are most fully described in the ethnographic literature.”
Creating the AWC dataset is proposed for completion before the Wiley Companion volume is published. This would enlarge, for example, the number of variables that are likely to be added to those in Table 2 as predictors of matrilineality, answering the question, for example: are other variables involved in the decline of matriliney?

The AWC subset could prove to be the most important addition to the DEf or GeoPhylo Regression project of “Transforming Cross-Cultural Research” thanks to the combination of providing: 1) imputation of missing values for variables that require or merit them, 2) the solution for autocorrelation, and 3) optimal expansion of variables by imputation from the commensurate data of both the Standard Cross-Cultural Sample (SCCS) and the Ethnographic Atlas (EA).

A new AWC would alter the fact that most of our researchers and nearly all of our students have avoided use of our XSEDE Science Gateway, for which the Complex Social Science website, a Galaxy site that directs the researcher or student either to one of two fast virtual machines at UC Irvine, building models for SCCS variables in 1.5 minutes, or to the UCSD/SDSC Trestles computer (with HPC high performance computing with 20 second job completion but a queue of 20 minutes). The attraction of virtual machines is that one can do many of the successive steps in DEf modeling in a series of 1.5 minute steps, and be done within a class period. Because the simpler modeling work is iterative, adding more or sounder variables step by step, few researchers or students use the HPC facilities at present. With a new AWC, sample, designed as a large dataset in terms of number of societies and number of variables, the time of Trestles to job completion would be much quicker than for the use of Dow and Eff’s native R. With AWC the expansion of the statistical power of DEf (GeoPhylo Regression with Imputation) would enable the student or researcher to expand Table 2 and discover, for example, whether there are additional variables associated with increase or decline of matriliney.

GeoPhylo Regression (DEf) offers a true transformation of cross-cultural research not only because of a generic solution to Galton’s autocorrelation problem and the imputation of coded values of variables for societies where these values are missing but because of results in Table 2, which employs the Ethnographic Atlas (EA) variables. Ideally, we could employ a third strategy of data analysis that takes the AWC sample that includes the best-coded societies of the EA and includes many of the SCCS societies that are in that portion of the EA. This strategy may work because the SCCS is a representative subsample of the most culturally independent cases of the EA (Murdock and White 1969) and the AWC is a best-coded subsample of the EA. A variant of this strategy might work, using the XC sample if DEf shrinks similar societies in the direction of a representative sample such as SCCS or the EA world cultures. GeoPhyloR models for variables in the AWC Atlas of World Cultures should take only 10-20 minutes to compute on CoSSSc. This would easily permit students to test replications of phylogenetic models (such as those of Holden and Mace 2004) during class time.

14. Substantive questions with issues and answers
Among the questions we ask of specific models developed and analyzed here and by chapter authors are those designed to question the consequences of our "Transformation of Cross-Cultural Research."

1) One of our most important results (in 10. ), potentially expanding the GeoPhyloR (DEf) method, was that GeoPhylo Regression with Imputation (DEf) could replicate the findings of phylogenetic analysis, as in Holden and Mace’s “The Cow is the Enemy of Matriliny” (Table 2). The phylogenetic approach (Holden and Mace 2005:217), shows that “using a phylogenetic comparative method that can also test the direction of evolution, or which of two co-varying traits changes first (Pagel 2002, Holden et al, 2005),” is simply a function of where each society and its values of variables fall in the temporally ordered phylogenetic tree. In (11. ) we showed how we might begin to extract from the GeoPhyloR/DEf dataset the phylogenetic language trees of the SCCS, which we can also do for the phylogenetic language trees of the EA. We used the EA dataset and variables to test phylogenetic models, but also showed in (12. ) how we might construct an AWC sample with the best-described societies numbering about half of the EA (N=562) but adding an additional V=circa 2200 SCCS variables. With these added steps of analysis, we can both 1) tell “which of two co-varying traits changes first”, for example, and identify further 2) whether there are any other SCCS variables which are “the Enemy of Matriliny”, or 3) “the Friend of Matriliny.” Such results can be analyzed to determine a fuller phylogenetic tree, with temporal ordering of branches, which tell of branchings in where matriline increases over time and others where it decreases.

2) The enhanced ability to do worldwide evolutionary modeling with time-ordered phylogenetics for individual or sets of variables is a powerful result that can be taught and used by researchers and would be a widely-used highlight of the Wiley Companion. The High Performance Computing linked to our CoSSci open access site also makes these analyses easy to run in a 10-12 minute session that fits into a classroom situation (assuming that an AWC dataset of best described societies in the EA might be created, as by Anthon Eff for the XC dataset).

3) What is it that GeoPhyloR (DEf) does to reduce autocorrelation? 1) It creates independent variable proxies from spatially nearby neighbors and societies with similar languages (the autocorrelation zone) that are predictive of the weighting of dependent variable in each society’s autocorrelation zone. That prediction is added to the usual regression equation for those independent and dependent variables as the control for autocorrelation. In addition, 2) “multiply structured homogeneities” may occur as a second aspect of Galton’s problem of autocorrelation when the results in (1) also turn out to be separately clustered around each society’s neighbors and ancestries. This possibility is shown in the mkmapping of “convex hulls” shown on maps of variables in the DEf R or CoSSci output, such as Figure 8.

4) Under what conditions is distance a stronger autocorrelation effect than language or the reverse? These effects are weighted in GeoPhyloR output data by Rsquared: In the Matriliny example, 80:20 for EA in favor of language. The weightings of
distance vs. language in all GeoPhyloR analysis allows the question to be investigated as to what “meta” conditions produce these imbalances.

5) Is phylogenetic analysis compatible with the study of behaviors and beliefs rather than genetics? Yes. “The Cow is the Enemy of Matriliny” is an example.

Do cultural variables carry phylogenetic information? Yes:
O’Brien et al. (2014): Biologists also view such things as a bird’s nest (i.e., deeply rooted behavior) as part of its phenotype.

Does horizontal transmission in culture prevent phylogenetic structure? O’Brien et al.: it depends on the relative rates of vertical and horizontal transmission.

Are language, biology, and culture concordant? O’Brien et al.: Any supposed cultural-historical “genetic unit” needs to be defined on independent terms (Romney 1956:57) as to whether language, descent and culture are closed and homogeneous units; i.e., very unlikely and not providing a basis for racial categories.

At what scale(s) can cultural phylogeny be examined? O’Brien et al.: Multiple. Human populations, cultures and languages are not simply analytic fictions but coherent entities, historically enduring although slowly-diverging phenomena despite births, deaths, immigration, and the like, and sociocultural entities, as with individuals, have ancestors, descendants, relatives and patterns of hierarchical descent (Terrell, 2001).

6) As shown in (4.) for the Moral Gods example, OLS and GeoPhyloR (DEf) models converge with low levels of autocorrelation (which should occur as well in Logit models (logistic analysis) with dichotomized variables.


8) What classes of variables seem to be well-modeled by some of the methods above? Eff argues for dependent variables that are fully coded, while other project contributors find useful the study of models with the dependent variable incompletely coded.

9) Can the imputed variables be used to model networks of variables? Yes.
Akaike’s Information Criterion (AICc) for single dependent variables? Yes.
Bayesian learning graphs? Yes.
Observed variable path analysis using imputed variables? Hopefully yes.

The following are experimental and bear on the convenience of modeling methods for researchers and students.
How do models using R code directly compare to those using:
Galaxy entry of original and computed variables? Works except for certain transformations of variables.
Quick Virtual Machine version (UC Irvine): Excellent, 1.5 minute runs for SCCS, WNAI, and LRB.
Queue-based High Performance: Excellent, but 20 minute runs.
Kepler Science Workflows: Excellent for AWC, runs all R variables, including those constructed from other variables, optimal for use with CoSSci.
iPython modeling within Kepler by Ilkay Altuntas: Experimental.
iPython Notebooks overlay for iPython and Kepler workflows: should make for easier classroom and homework for students.

15. Culture and Economic Growth: An intensive frontier of cross-cultural research
Economist Frederick Pryor (2005, and in earlier work) pioneered the use of the SCCS and Ethnographic Atlas cross-cultural data in studies of economics, including the effect of the plow on gender roles and many other topics. More recently, Eff, Dow and Rionero (Eff and Rionero 2011, Eff and Dow 2008, 2010) analyzed parental investment and growth, markets and prosocial behavior, and market integration and prosocial behavior using Dow-Eff (GeoPhyloR) software that controls for autocorrelation.

A major current debate in economics is over the long-term effects of culture on economic development and growth. Key articles make extensive use of cross-cultural data and are concerned with appropriate research methodology. Spolaore and Wacziarg (2013) review this literature, which shows that economic development is affected by "traits that have been transmitted across generations over the very long run." They establish an extensive research framework to evaluate which kinds of traits are involved in these effects, whether cultural (behavioral and symbolic transmission) or biological (epigenetics or genetic distances), in different ways. Acemoglu and Robinson (2011) show that man-made political and economic institutions underlie the economic success of nations. Michalopoulos and Papaioannou (2013) find that the best predictor of current economic performance, proxied by light-intensity maps, was pre-colonial institutions of different ethnic groups, coded for Africa from Murdock's Ethnographic Atlas. Using the Atlas data again, Michalopoulos, Naghavi and Prarolo (2012) show that contemporary intra-country Muslim representation is higher in regions of unequal endowments of agricultural potential and proximity to pre-Islamic trade routes. Studies of changes in women's roles from pre-industrial times by Alesina, Giuliano and Nunn (2013) and Giuliano (2014) use cross-cultural and national codes to argue, like Pryor, that differences in gender equality can be related back to a history of using the plow and its male-oriented division of labor.

Economists are very good with statistical models that include instrumental variables as controls for estimating causal relationships when controlled experiments are not feasible, somewhat analogously to solving Galton’s problem. Instrumental variables often specify a specific context, such as the effects of melatonin on sleep which may be considered exogenous, for example, when as a medication it can be taken exogenously irrespective of normal processes of sleep. Single or multiple variable controls for "context effects" are not so easy to find in cross-cultural studies. There, it is the more generalized controls for autocorrelation or clustering (especially proximity, language phylogeny and phylogenetically based evolutionary models) that can be controlled, as in Dow-Eff (GeoPhyloR) modeling. The modeling done by economists in the Culture and Economic Growth/Development paradigm does not as yet use such controls, and their models are often vulnerable to Galton’s problem. The question is, with the convergent use of cross-cultural or cross-national databases, will economists adopt the new modeling tools developed, used, and evaluated in the Wiley Companion to the Transformation of Cross-Cultural Research?

16. Kepler Science Workflows and Embedded Software

The downloadable Java-based Kepler Project self description is as follows:

"The Kepler Project is dedicated to furthering and supporting the capabilities, use, and awareness of the free and open source, scientific workflow application, Kepler. Kepler is designed to help scientists, analysts, and computer programmers create, execute, and share models and analyses across a broad range of scientific and engineering disciplines. Kepler can operate on data stored in a variety of formats, locally and over the internet, and is an effective environment for integrating disparate software components, such as merging "R" scripts with compiled "C" code, or facilitating remote, distributed execution of models. Using Kepler’s graphical user interface, users simply select and then connect pertinent analytical components and data sources to create a "scientific workflow”—an executable representation of the steps required to generate results. The Kepler software helps users share and reuse data, workflows, and components developed by the scientific community to address common needs."

"The Kepler software is developed and maintained by the cross-project Kepler collaboration, which is led by a team consisting of several of the key institutions that originated the project: UC Davis, UC Santa Barbara, and UC San Diego. Primary responsibility for achieving the goals of the Kepler Project reside with the Leadership Team, which works to assure the long-term technical and financial viability of Kepler by making strategic decisions on behalf of the Kepler user community, as well as providing an official and durable point-of-contact to articulate and represent the interests of the Kepler Project and the Kepler software application. Details about how to get more involved with the Kepler Project can be found in the developer section of this website."

"Kepler is a java-based application that is maintained for the Windows, OSX, and Linux operating systems. The Kepler Project supports the official code-base for Kepler development, as well as
Further description of Kepler Workflows is provided by Ludäscher et al. (2005) and Barseghian et al. (2010). The Kepler package, when downloaded, has access to The Kepler Getting Started Guide, Actor Reference and User Manual.

Interest in Wiley Companion helped to create a “Wiley Open User Group” Kepler Project. The SDSC Workflows for Data Science Center of Excellence "WOrDS" group showed how to use the Dow-Eff R code (GeoPhyloR) for a simple, complete, and well documented DEf model (Moral Gods) created as a test workflow for GeoPhyloR modeling. Assembling Kepler user options to change the independent variables and the name of the output file required only two hours of expertise to map the DEf program dependencies and make the two options accessible as successive workflows. A next step was to create input workflows for generic use of the workflows for any given DEf dataset, dependent variable, and independent variables, plus the additional, covariate (Unrestricted) variables and program output. Early tests showed that the output produced by these Kepler workflows was identical to that produced by DEf models computed with R. GeoPhyloR workflows added 1) new computed variables 2) scale construction and 3)...

A further step will enable the library (bnlearn) to produce graphics of networks of variables with links that show possible causal links, such as hypothesized between certain independent variables with a given dependent variable (or possibly others). In this case new R code enters the workflow that takes the output of DEf (the imputed variables and a computed Wy) as new input. Examples of still further possibilities include an output of distance, language and ecological matrices used for autocorrelation, or combinations thereof, and combinations of imputed variables from which inter societal distances can be computed and output for analysis with the Neighbor-net (SplitsTree4) software by Huson and Bryant (2013). Kepler workflows can provide prototypes for NSF-funded projects that enable workflows to be added to the core on-line capabilities of the Co-SSci Galaxy software, which does not require downloading software. Members of our user community can also produce workflows, as did Michael Fischer in producing a workflow that checks if the correct R packages have been installed to a DEf or Kepler script.

Contemporary software like Java-based Kepler can run software written in other languages, just as Kepler can run inside more encompassing packages like Eclipse, which can integrate the use of other software packages. SDSC leader of the WOrDS group is investigating making use of Kepler easier by embedding workflows in iPython visualization software, and iPython in iPython Notebooks so as to provide a more effective research, teaching, student and classroom environment.

One experiment to carry out is whether the GeoPhyloR (DEf) package will provide complementarity or divergence between SplitsTree models and information from Wy matrices trained separately on linguistic and proximity autocorrelation as in.
Table 3. “The Cow is the Enemy of Matrlny” model in Table 4 augers well for the convergence of GeoPhyloR with standard phylogenetics.

Table 3: Autocorrelation results for several models carried out as Kepler Workflows

2. EvoGod2.olsresultsNewEvoGod13.ew Ev2013.01d Sanderson Religion
3. Ev2007MoralGod5.14.4SDSC
4. v626quintet2 Dow & Eff Functions DEf01d Reiss v626quintet

<table>
<thead>
<tr>
<th>dist</th>
<th>lang</th>
<th>ecol</th>
<th>Rsq</th>
<th>Listed as Ilkay Altintas models with Kepler.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>0.00</td>
<td>--</td>
<td>.36</td>
<td></td>
</tr>
<tr>
<td>0.28</td>
<td>0.38</td>
<td>0.34</td>
<td>.50</td>
<td></td>
</tr>
<tr>
<td>0.70</td>
<td>0.30</td>
<td>--</td>
<td>.72</td>
<td></td>
</tr>
<tr>
<td>0.50</td>
<td>0.50</td>
<td>--</td>
<td>.55</td>
<td></td>
</tr>
</tbody>
</table>

17. A 120 year improvement gap

By the time the Wiley Companion to the Transformation of Cross-Cultural Research is published (e.g., 2017) it will open the door to 120-year improvements (counting from Tylor’s error in 1889 and including the improvement by GeoPhylo Regression over earlier types of phylogenetic modeling) not only in core social science and anthropological research but to the study of observed variables generally. This applies to observational sciences with the notable exception of those based on observations that are so ubiquitous that statistical inferences are unnecessary, the phenomena do not involve complexity, and can be robustly measured directly. As an exception, Einstein’s theory of how gravity is warped by mass offers a deeper example that in physical terms is so complex that it required not only exquisitely careful measurement (of the movement of stars before, during and after an eclipse) but also a reshaping of theory rather pre-Newtonian types of Ptolemaic corrections. Simple billiard-ball physics does not require attention to complex inferential statistics: given initial conditions billiard balls interact locally with outcomes approximating entropy and eventual equilibrium without further outside effects. Systems with long-range interactions disequilibrates local closure governed by short range equilibrium outcomes. Most aspects of the biological and social sciences are complex, and cannot be understood without understanding complex interactions. Clearly, not only are better models required at the level of regression or logic, but higher order interactions that can only be understood either with complex networks of interdependent variables (Bayesian networks, complex dependencies in time series), or a whole new understanding based on theories of constraints rather than mechanisms (Deacon 2012, Venditti, Meade and Pagel 2010). Vendetti et al., for example, show from phylogenetic data that 78% of their sample of mutations among species are imposed by a single random mutation that creates a new constraint on reproductive interactions. This is not the case for humans and some of their common ancestors where complex cooperative interactions may have depended on extra somatic complexes that we have come to call "cultures" (Pagel 2011), characteristic of complex interactions within and between groups. The use of sounder inferential statistics employed to impute missing data and solve the problems of autocorrelation, at the level of better ols or logit regressions are not a
final solution for social science modeling if these lead to better foundations in modeling more complex networks of variables, or models of the dynamics of complex systems of constraints, they may provide avenues for deeper discoveries. These views, whether correct or not, are those of a social anthropologist turned complexity scientist and not those of the kind of positivist often associated with quantitative methods.

It is all the more understandable and relevant that controls for Galton’s problem were not achieved with calculation of effective sample size, whether in survey sampling (Hanson and Hurwitz 1953, Kish 1965, 1987), cross-cultural research (Naroll 1961, 1965, 1970, White and Murdock 1969), or other fields of study in the social and natural sciences. What was needed over the 120 year period 1889-2009 and only finally solved by Dow and Eff (2009) was a measure of the actual effects of variables with clustered similarities on outcomes, and not just a calculation of effective sample size controlling for clustering.

18. Teaching a Seminar, Class, or Tutorial with CoSSci

Go to the CoSSci site yoursrel and press to play the "Co"-"Sci" histories YouTube = https://vimeo.com/79056519 that explains in 2 minutes how you or your students can and should login and then change the name of your modeling history (models that don’t work can be erased) to a save-online site to copy and refer to in the future. Teachers can record their students’ saved CoSSci coursework and permit you to view their sites (and invite experienced outsiders to view and make suggestions). You can view the 20 minute YouTube that explains how to set up and run a model with "DEf01d" analysis, choose the (easiest) SCCS dataset, and fill in the blanks that appear with a dependent variable, independent variables, and a few "covariates" (Independent variables in Unrestricted model): those that seem from the SCCS codebook\(^27\) to be strong variables coded for most of the 186 cases. Then go to the bottom of the page and press EXECUTE, and if the right hand column turns red press the blue button and then the “stderr” check for errors in the model and rerun. Once the right column is green press the button that looks like a diskette, which will send Mac csv output to "show downloads" at upper right. Click to open that *.csv file, copy it to .xls to save formatted output, use Screen Save to copy your tables, and insert that graphic into your word text. This will show the variables that were significant. Next search for "To Try" in the csv, which may contain new variables that may be checked in the codebook to see if they make sense, and you may then do a model with some variables deleted and some added. Repeat this whole operation (through *.csv output files) until you have a stable model and then evaluate what they imply as a potential model where the regression equation makes sense.

If we implement the AWC (N=562, V= circa 2100) subset of the Ethnographic Atlas (N=1267, V=166), see (10.), we are in the range of “big data” requiring HPC compute time in classes (circa 10-20 minutes). As noted in (12.), it also becomes possible, with use of Neighbor-Net analyses (11.) to do full scale evolutionary scenarios with GeoPhylo Regression focusing on phylogenetics.

47
Also to be studied is whether, just as Table 2 results could be used to display with Neighbor-Net the phylogenetic branches (Figure 6) where societies turn to “Cattle and No Matriline,” something similar could be done with Neighbor-Net distance pathways (Figure 5). Which nodes for societies link to other nodes where some feature, possibly as a result of other features, is borrowed over a certain distance or through more than one possible route of diffusion?

19. Conclusions
Survey sampling – like “big data” – deals with data as it comes in raw form: percentages, correlations, regression, multidimensional scaling and other forms of data analysis that do not correct for autocorrelation – Galton’s Problem, which distorts findings to overemphasize significance tests and often distorts the true relationships among variables. Ordinary regression analysis, for example, gives distorted results if the error terms are correlated with the predictor variables, a sign of autocorrelation. For cross-cultural studies, we can only make an overall guess, but our estimate is that more than 50% of all statistical findings are seriously wrong because they are statistically invalid without corrections for autocorrelation. The physics of simple systems do not have to deal with this problem: they typically have so many observations that significance tests are unnecessary and equations can be fit to the data. Modeling complex systems usually requires taking autocorrelation into account.

The purpose of the Wiley Companion is to provide easy and extensive access to researchers, instructors, and students – at all levels – to fundamentally important statistical tools that are not available in commercial statistical packages that deal with samples of observed variables – especially in cross-cultural research – and may not be accessible elsewhere for another decade. That is because W matrices are required to control for autocorrelation and completely coded subsets of a database are required in the best forms of missing value imputation. An open-access R-based package also has greater flexibility in allowing improvements to the code.

Many of the sciences now have specialized “Gateways” with science workflows (SWFs) specialized to their needs and their latest implementation. Rather than options for single methods (regression, logit, etc.) workflows allow a range of specialized options that can be visualized for different configurations of analysis. Dow and Eff’s mkmapping procedure, for example, not only provides world maps of cross-cultural variables but also convex hull boundaries around sets of observations (societies) that have clustered homogeneities, as in Figure 8. DEf output of a dataset that has been imputed can be routed to workflows such as library(bnlearn) that can calculate and visualize (as in Figure 1) potentially causal relationships as Bayesian graphs. Our cross-cultural databases (SCCS, EA, LRB, WNAI) are evolving into larger configurations (XC, AWC) with more auxiliary datasets (e.g., ecological).

Our CoSSci (Complex Social Science) Gateway, freely available for classroom use, is enabled to use HPC (high performance computing) on the Trestles supercomputer to implement more complex, evolving applications with initial NSF funding and
eventual funding by user groups, as in the other sciences. GeoPhylo Regression is now also implemented in the Kepler Science workflow open-source project begun in 2003 (Altintas et al.) and completed in 2014, with an ongoing user group. This enables a team of researchers to experiment with new forms of cross-cultural and related analysis, and submit proposals for user-group funding.

The effects of particularistic and multiple sets of clustered similarities on outcomes to observed-variable research questions are highly deleterious in inflating significance tests and randomly distorting parameter values of statistical models. These effects of autocorrelation are not solved by a simple calculation, such as adjusting “effective sample size” in survey sampling (Hanson and Hurwitz 1953, Kish 1965, 1987) or in cross-cultural research (Naroll 1961, 1965, 1970). Valid statistical methods that affect outcomes to these autocorrelation problems are required for valid research outcomes. Distortions of research findings without autocorrelation controls are ubiquitous in observational studies generally. They reoccur in the social, natural and evolutionary sciences. GeoPhylo Regression (Dow and Eff 2009, DEf) controls for Galton’s problem, and gives statistical solutions to these foundational research problems.

The Dow-Eff GeoPhylo Regression methods include testing whether a Wy correction term for autocorrelation is sufficient to render unbiased results of inferential statistical analysis, i.e., solving for Galton’s problem. Eff summarized what Galton pointed about autocorrelation (clusters of separate homogeneities in a sample of observed variables) as the uncertainty as to whether “similarity between cultures could be due to borrowing, could be due to common descent, or could be due to evolutionary development; he maintained that without controlling for borrowing and common descent one cannot make valid inferences regarding evolutionary development” (Eff 2004). Multiple conclusions that follow from GeoPhylo Regression (DEf) are of major importance for cross-cultural research and for the problem of autocorrelation controls in observed variable datasets generally.

1) For phylogenetics, we have discovered how to use imputation of missing data for variables in datasets such as SCCS, EA, LRB, WNAI, and, best of all, the possibility of an as yet unimplemented AWC (an N=562 best-described subset of the N=1267 societies of the Ethnographic Atlas variables to which an additional 2100+ variables can be borrowed from the SCCS. Once this is done then two small extensions to DEf could locate the positive and negative effects shown in the DEf csv output and output produced in-memory by execution of DEf model, e.g., as to where variables tend to change at certain nodes in a phylogenetic tree. One new methodological step would be a function that extracts a Neighbor-Nets language tree (11. ) from the AWC dataset geographic coordinates. The other is a function that would locate where each society falls in the phylogenetic tree in relation to certain values of its dependent variable. This is possible because “using a phylogenetic comparative method ... can also test the direction of evolution, or which of two co-varying traits changes first (Pagel 2002, Holden et al, 2005).” In essence, we should be able to identify evolutionary trees for different dependent variables, and possibly identify
where some dependent variables fall for given language trees. If there are shortcomings in the use of AWC with respect to fully-coded variables for use in imputation of missing data, a new coding project could rectify these shortfalls.

The capability of cross-cultural research – all of the variables in our GeoPhylo databases, whether for foragers, specific regions, or world samples – to begin to estimate evolutionary time series for sociocultural variables, an approach that began largely in Britain with physical anthropology and language phylogeny reconstruction of relative time series, represents a massive scientific advance. The fact that GeoPhyloR (DEf) is able to recover language phyla reconstruction through time is a major step that allows students and researchers to learn and enter into this kind of research, allowing for a proximal development to implement, given DEf findings, and powerful phylogenetic analyses that go beyond the types of analysis we see in Figure 6, i.e., mapping the GeoPhylogenetic DEf results into the inferential statistics with time series.

2) Controlling with DEf for both distance and phylogenetic effects, where applicable, can tell us not only that for some percentage of the autocorrelation effect is due to distance, but help us to estimate where these effects are occurring. Comparing for two societies the normalized rows in the W matrix for distance and directed language effects and noting their relative percentages should allow identification of pairs of societies that are affecting one another through proximity as compared to changing values of variables as peoples migrations create a language tree.

3) For econometricians, the use of instrumental variables is a partial attempt to control for Galton’s problem, but the GeoPhyloR (DEf) solution is based on both particularistic autocorrelation networks and differentiated homogeneous clusters rather than control variables. It is likely to be difficult, once the DEf network solution to autocorrelation is employed, to find that instrumental variables are better controls that network-based controls.

4) There may be no better network predictors of autocorrelation than those proposed by Dow and Eff (DEf GeoPhyloR).\(^2\) Dow, in fact, chose network predictor methods as the best approach to the solution of Galton’s problem. That allows the possibility that there might be additional W matrices, such as second order \(W^2\) and third order \(W^3\) matrices (i.e., indirect paths of influence) that might be added to further control Galton’s problem (Brandsma and Ketellapper 1979).

Combined with imputation of missing variables, and the possibility of using samples of imputed variables for the study of complex networks of variables, the DEf approach, after 120 years of inability to deal with comparative observed-variable studies (discounting the partial ability of phylogenetic research to provide a half-solution without dealing with borrowing and effects of distance), finally enables social scientists to deal with the complexities of real-world systems. Economists have begun to recognize that sociocultural systems have long and complex footprints that demand inclusion of past histories, difficulties in cross-cultural
borrowings, and evolutionary developments. Econometricians, however, have not yet absorbed these new inferential approaches. The past does indeed shape our present and our future but only now do social scientists have the opportunity to deal with these complexities, which can be more realistically discovered with GeoPhylo Regression (DEf). Hopefully this Wiley companion and its use in research and the classroom will sharpen our opportunity to use these new resources productively.

20. Acknowledgements
Thanks to Peter Bearman, Lilyan Brudner, Scott White and Anthon Eff for commentary, to Halbert White for early guidance, and to Malcolm Dow and Anthon Eff for the creation of DEf/GRPI software. The Santa Fe Institute gave essential support in this work over five years in hosting successive meetings of the Causality/Robustness Working Group, with members Henry Wright, Amber Johnson, Peter Turchin, B. Tolga Oztan, Giorgio Gosti, Elliott Wagner, and Ren Feng, with occasional participation from Marcus Hamilton, Seth Lloyd, Scott White, Laura Fortunato and, remotely, Chris Boehm, John Snarey and Anthon Eff. Many thanks to the Max Planck Institute for Mathematics in the Sciences, Leipzig, where Jürgen Jost and Nihat Ay hosted a two-week meeting of our working group where Ren Feng, B. Tolga Oztan, and Giorgio Gosti were able to attend.

21. References:
Bacon, Margaret K.; Herbert Barry III; Irvin L. Child; Charles R. Snyder. 1965 "A cross-cultural study of drinking", Quarterly journal of alcohol studies (special issue), Supplement 3.


http://eclectic.ss.uci.edu/~drwhite/pw/NetworkAutocorrelationASimulationStudy.pdf


http://www.jstor.org/pss/644404


Holden, Claire J., Ruth Mace. n.d. The cow is the enemy of matriliney: Using phylogenetic methods to study cultural evolution in Africa. pdf of a powerpoint.


Huson, Daniel H., and David Bryant. 2006. Application of Phylogenetic Networks in Evolutionary Studies. Molecular Biology and Evolution 23(2):254-267. Center for Bioinformatics (ZBIT), Tubingen University, Tubingen, Germany (SplitsTree4
http://mb.e.oxfordjournals.org/content/23/2/254.full.pdf
http://www.maths.otago.ac.nz/~dbryan/Papers/05Networks.pdf


Naroll, Raoul. circa 1978. personal communication.


Waddell, Peter J., Xi Tan. 2013. New g%AIC, g%AICc, g%BIC, and Power Divergence Fit Statistics Expose Mating between Modern Humans, Neanderthals and other Archais.


White, Douglas R. 1985a. Entailment Analysis. World Cultures 1#3. Art. 7. (Codeveloped with Michael Burton)


White, Douglas R. 2014. How Not to do Cross-Cultural Research. For submission to World Cultures eJournal [eScholarship].


22. Appendices
A. Misunderstandings about Galton’s Problem
B. Notes on a failed model replication: Due to lack of shared representative sample?

A. Misunderstandings about Galton’s Problem
A common misunderstanding about autocorrelation, which applies in the context of cross-cultural research (as well as in econometric modeling with instrumental variables, likely to be accepted easily among anthropologists), is stated by Ember and Ember (1998:678): “Raoul Naroll and others have considered Galton’s problem a serious threat to cross-cultural research. They have devised several methods to test for the possible effects of diffusion and historical relatedness (Naroll 1970). The concern behind these methods is that statistical associations couldn’t be causal if they could be attributed mostly to diffusion (cultural borrowing) or common ancestry.” This is simply untrue. Autocorrelation effects and potentially causal effects can be segregated: one set of effects does not obliterate the other. This is true
for econometric models that include cultural variables and is true for anthropological models that focus on cultural variables.

In the post-1969 period, Goodenough (1970) muddied the issues for comparativists by giving priority to a linguistic differentiation between the "etics"²⁹ of cross-cultural descriptive observations and the "emics"³⁰ of fieldwork in understanding native categories. At the comparative level, Melvin Ember (1971), who became the HRAF president after Naroll, misunderstood the statistical effects of autocorrelation, asserting that autocorrelation did not exist so long as the sample was randomly chosen. Loflin disliked many researchers of Ember’s belief that probability samples (e.g., HRAF’s 60 society samples) are free of Galton’s problem, and also that the SCCS sample is free of autocorrelation. Ember (1971) omitted to mention Murdock and White’s suggestions that autocorrelation “runs tests” along a path of adjacent societies can measure the extent of Galton’s problem, but they do not “solve” the problem other than reducing the estimation of effective sample size n to what for the SCCS were ridiculously low values.

My comments in the preface on Naroll’s, Ember’s and Bernard’s misunderstandings of the statistical complexities of Galton’s problem, are intended as examples that give voice to the complexities of cross-cultural research. Acceptance of false descriptions of Galton’s problem, e.g., by Bernard (1987:515, 533), as well as Ember, along with unnecessarily finicky critiques by Naroll, may have led prominent anthropologists like Eric Wolf to reject the validity of cross-cultural research. Others, such as Harris (2001:615,627), argued for its validity when properly conducted.

B. Notes on a failed model replication? Due to lack of shared representative sample?
One strength of DEf is that with representative samples it evens out discrepancies in number of sample cases in different relatively homogeneous clusters. Murdock’s Ethnographic Atlas, for example, has 412 relatively homogeneous cultural groups and 200 cultural provinces that aggregate the most similar cultural groups. Within the 412 clusters and the 200 provinces the numbers of societies vary. DEf computations sort out the effects of distinct clusters of homogeneity.

Like all observation-based research comparative research requires representative samples of cases that are subsets of a statistical population that accurately reflects the members of the entire population, in the present context, of known and well described societies. In the one case of the XC sample, which is not representative, a complex set of weights can be employed within the regression framework so that representativeness is achieved even though the sample has a very irregular distribution of its continental census of societies (47 Africa, 57 Asia, 12 Australia, 7 Europe, 198 North America, 16 Oceania, 34 South America). XC is an amalgam of all the variables of the EA and at least one of the other datasets: SCCS, LRB, and WNAI (only a three SCCS societies were not coded for variables in the EA sample and were given only an EA identifier).
The use of the XC sample without appropriate weightings shows what happens when a phylogenetic replication study is done that fails to adjust for the unrepresentativeness of the sample, which might have been fixed with weightings. Here, however, Africa happened to be represented for the Matriliny variable by only four societies, all in Southern Africa, in spite of a dataset size of N=371 societies and several thousands of variables (V=2657). The sample’s unevenness is because of an excess of North American Indians from the WNAI samples and from the LRB sample of foragers. Its use unweighted is thus not recommended for a test of “The Cow is the Enemy of Matriliny.” Table 5, however, shows naïve results of my own for predicting Matriliny without noting that the datased was badly skewed because of the lack of even representation of world cultures.

Prior to my testing of a model of Holden and Mace’s (2003, 2005) “The Cow is the Enemy of Matriliny” as shown in Table 2, I used variables in the imputed XC sample, chosen simply by using variables from the XC codebook, i.e., at the codebook url, https://dl.dropboxusercontent.com/u/9256203/XCcodebook.txt. Among the relevant variables initially selected for testing this (potentially) phylogenetic model was constructed by retrieving the following R code, where the first element of the variable name places the variable as belonging to XC but the additional command names old or new XC variable for retrieval into the XC dataset:

```r
setDS("XC")
dx$XC.v43<-if(dx$SEA.v43==3)*1 # EA DepVar<--Descent: Matrilineal of Major Type
dx$XC.v39<-if(dx$SEA.v39>3)*1 # EA Plow: Animals and Flw Cultivation
dx$XC.v1127<-if(dx$SCCS.v1127) # SCCS Plow-positive: Crop Type Plow-positive or -negative
dx$XC.v40<-if(dx$EA.v40==7)*1 #n=69. EA Bovines: Predominant Type of Animal Husbandry
dx$XC.v41<-if(dx$EA.v41) #n=58. EA Milking
dx$XC.v42<-if(dx$EA.v42>5)*1 # EA Agriculture: Subsistence Economy
dx$XC.v44<-if(dx$SEA.v44) # EA Metalworking
dx$XC.v109<-if(dx$SCCS.v109) # SCCS v109 Female Soil Preparation
dx$bi.8 # BIOCLIM: Mean Temperature of Wettest Quarter (dgC*10)
dx$mnnpp # Mean Net Primary Production within 50 km radius dx$XC.v6<-
dx$XC.v6.7<-if(dx$SCCS.v6==7)*1 # n=46. Bovines as Animals- Domesticated
dx$SCCS.v244<-if(dx$SCCS.v244==7)*1 # n=68. Bovines Predominant Type of Animal Husbandry
https://dl.dropboxusercontent.com/u/9256203/XCcodebook.txt
http://intersci.ss.uci.edu/wiki/index.php/Ev70.d1DEf01fMatriCattlePlow (XC) R code
```

After calling these and other variables and naming the dependent variable of matrilineality, dpV <- "EA.v43.d3", a final list of variables was arrived at after a good deal of initial work. The last of these variables is the presence of cattle. RiV <- c("EA.v44","XC.v109.19","mnnpp","XC.v1514"). Other variables are significant predictors of matriliny, not just for Africa, but also for 371 societies worldwide, including missing data imputations for all other variables. The phylogenetic tests performed by DEf for effects of distance, language and ecology are thus, again, not just for Africa, but 371 societies worldwide.

The results of this worldwide GeoPhylo Regression are a complete bust. If this were a valid sample, with societies that are proportional to types of those in the SCCS versus the Ethnographic Atlas, then they might be showing, if valid, that language phylogeny is not significant (zero effect) but that distance autocorrelation and
ecological autocorrelation account for 85% and 15% of the autocorrelation effect, respectively. The basis of these results, however, does not correspond to the preconditions of representativeness for validity of results. In this spurious and unrepresentative sample few societies in the XC sample are in Africa fewer still of those are matrilineal.

**Table 5.** Distance Autocorrelation and Language Phylogeny tested as Causes of Matrilineality in the XC (N=369) sample. Cattle “appear” to have no effects on driving out Matrilineality once distance autocorrelation is controlled. Other predictors include Female Contribution to agricultural Soil Preparation and Planting, a negative effect of Primary Production within a 50 kilometer radius, and a positive association with Metal Working, which is almost all male worldwide.

One of the main effects in this pseudo model is SCCSv109, a Female Soil Preparation scale that when shown as a table(dx$XC.v109) = (values 49 66 27 14 17 10 with 10 the number of societies in which women do most or all the preparation of the soil for planting), the prediction of matriliney is quite significant (pval=0.005). The problem with this sample, however, is that there are not enough African societies, and similarly for mostly under-representations for other regions. This test, then, is likely to be inappropriate and statistically invalid because of the way the sample is constructed. Here, however, the discussion about a possible XC2 sample is relevant from section 11.

The following ordinary least squares output shows results of an ols model for these same variables. The adjusted Rsq, has a surprisingly non-significant pval=0.02, which is low, possibly because of the unweighted version of the construction of the XC sample that was used, or possibly because and XC2 sample was not used.

```
lm(formula = dx$v43.d3 ~ dx$v40.d7 + dx$bio.18 + dx$mnnpp)
Residuals:
    Min     1Q   Median     3Q    Max
-0.38548 -0.14694 -0.12017 -0.06462  0.95445
Coefficients Estimate    Std. Error t value Pr(>|t|)
(Intercept)  1.088e-01  1.701e-02  6.398 2.30e-10  ***
dx$v40.d7  -6.801e-02  2.018e-02 -3.370 0.000776  ***
dx$bio.18  1.483e-04  3.555e-05  4.171 3.27e-05  ***
dx$mnnpp  -1.955e-02  1.037e-02 -1.884 0.059836 
```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.33 on 1138 degrees of freedom
(123 observations deleted due to missingness)
Multiple R-squared: 0.02702,  Adjusted R-squared: 0.02445
F-statistic: 10.53 on 3 and 1138 DF,  p-value: 7.772e-07

23. Footnotes

1 “Galton invented the use of the regression line (Bulmer 2003, p. 184)... and “is responsible for the choice of r (for reversion or regression) to represent the correlation coefficient.” [Clauser 2007] “In the 1870s and 1880s he was a pioneer in the use of normal distribution to fit histograms of actual tabulated data.” Wikipedia: Francis Galton. Clauser, Brian E. 2007. The Life and Labors of Francis Galton: A review of Four Recent Books About the Father of Behavioral Statistics. 32(4):440-444.

2 Survey subsamples include interviewing members of families, who are much more similar than people sampled from a larger universe; oversampling types of individuals that are rare in the general populations; or oversampling random clusters of people who live closer together in order to reduce the time required to travel from site to site in the process of interviewing.

3 Survey samples are designed to accurately collect numbers that represent a population. Their effective sample size corrections adjust only for sample design, i.e., subgroups that are over or under-sampled. Thus when analyzing sample data, whether a probability sample or representative sample, clusters of homogeneity within the sample – autocorrelation or Galton’s problem – are not taken into account. Ethnographic samples also aim at representation, but often in terms of societies that are well described. A representative sample of “all societies,” whether or not studied ethnographically, would contain a majority of societies with little or no data. In the past it made a huge difference whether a cross-cultural representative sample contained as many ethnographically studied societies as possible (e.g., Murdock’s Ethnographic Atlas) or a sample containing as many such societies representative of relatively independent well described societies. With controls for autocorrelation (e.g., Dow and Eff software) these contrastive sample should produce highly similar results provided that both are representative of the same criteria, e.g., world cultures, foragers, or regional sample.

4 Medical studies often have a very different construction: they typically sample from one or more hospitals from a given region, reducing diversity; they often attempt to recruit a control group to match the homogeneities or diversities of the patients treated, and they use a specific form of randomization for “potential outcomes” that is claimed to eliminate selection bias as to which subjects were recruited to the hospital and which not (Angrist and Pischke 2009:12-15). Tsai and Peace (2013), backed by many others who follow the work of causality theorist Judea Pearl (2009), show first that “Although subjects are randomized to treatment groups in clinical trials, this does not imply the same degree of randomization among sub-populations of the original trials.” The faulty part of such quasi-experimental inferences is equivalent to the cross-cultural research problem of the
extent of independent diversity versus fundamental homogeneous clusters (aka Galton’s problem). True, there are more distinct homogeneity clusters in a cross-cultural sample than in most medical studies where patients are compared to a control sample. But to do controlled comparisons in medical research properly, a researcher would need to do fieldwork on patient and nonpatient populations to capture the full extent of homogeneous clusters that invariably create a version of Galton’s problem that can lead to major errors of inference in statistical results.

5 Again, including controlled experiments as in medical control groups.
6 With the precaution that others of Rubin’s (1973, 1974, 1977, 1991) statistical approaches (see Angrist and Pischike bibliography) do not solve problems of Bayesian probabilistic causality as do mathematical approaches by Judea Pearl (2009) and Nagarajan, Scutari and Lèbre (2013).
7 There are not necessarily any generalized regional clusters to be considered as an essential part of solving Galton’s problem.
8 I must confess that given the severity of nonindependence of societies that I illustrated in the Standard Cross-Cultural Sample (Murdock and White 1969), which rendered significance tests meaningless, I was discouraged about doing substantive studies in cross-cultural research up to 1988 because no means of correcting Galton’s problem were available. Later (White 1993) I showed the serious extent of Galton’s problem up to that date in the variables of the Standard Cross-Cultural Sample using Moran’s coefficient. Eff (2004:161) repeated this exercise a decade later with very similar results: 44% of the variables at that time had a significant Moran’s coefficient for autocorrelation.
9 The reason is that the statistical tests for entailment analysis (White 1985c) were based on signal detection analysis that compared actual entailment frequencies against comparable results randomizing the data while keeping marginal constant.
10 Periods of scarcity in societies with exchange economies are the converse of Norenzayan’s (2013) and many other’s predictions about “Big God” religions (see Atkinson et al. 2014 for a book review). I.e., it is not that Christian and Islamic societies are “Big God” religions because they are “Big God” societies but because they are economically weak and vulnerable economies that, when they crash, they produce inequalities sufficiently extreme and intolerable that preachers of more egalitarian moral religious principles come to be accepted. Alexander (1987:240): “this is how moral systems have always operated; the expense of being selfish is why extreme altruism has prevailed within groups in conjunction with severe extrinsic threats.”
11 BIOCLIM variables must be rounded to integers before they are recognized in library(bnlearn) → Reminder: Anthon Eff July 26 2014. This could be done to improve each of the databases.
14 Wikipedia:Computational phylogenetics.
The core of the Nexus format is a square matrix where only the n rows are named (no spaces, commas or parentheses) followed by distances in n space-separated columns.

Wikipedia:SplitsTree.

SplitsTree4, downloadable at http://splitstree.org

Because DEFW matrices (e.g., distance, language, and ecological similarity data) are computed for all societies in a dataset, and they are not model specific, it is efficient way to compute them once during a single run of the software and compute and save the output using Huson and Bryant’s (2006, 2013) SplitsTree software for each variable separately. This provides the potential for a visualization and metric for comparison of how other variables or packets of variables are keyed to phylogenetic networks. This is already done in constructing the Wy control variable used in DEf, and can provide a measure of the effective sample size controlling for autocorrelation, but can also measure and show where more detailed phylogenetic modeling is required, as in the British phylogenetic research tradition.

Only three SCCS societies were not coded for variables in the EA sample.

Certainly this would be better than a weighted-sample-size solution of XC based on sample segments from SCCS, WNAI, LRB, SCCS-WNAI, SCCS-LRB, WNAI-LRB and SCCS-WNAI-LRB ratios.

D. R. White’s expertise in survey sampling was learned during two years as Project Co-Director for Methodology in the Irish National Language Surveys project, 1971-1973. He directed four national surveys (general population, surveys of the 1st and 6th school tiers of schoolchildren, and the Civil Service) working with Glenda Cimino of the Michigan Survey Research Center and interviewers of the Irish ESRI (Economic and Social Research Institute), a project funded at Ministry Levels, Irish Republic.

One could add to the EA all those variables from other datasets, especially those of the SCCS (V=2109), which includes only one society in each of the 186 homogenous clusters in the Ethnographic Atlas. To include societies in that WNAI (North America) and LRB (forager) samples is inappropriate, however (see 12. Regarding representative samples), because foragers or North American samples will skew and destroy the fact that only the EA and SCCS samples are representative of world cultures, with the SCCS the best described relatively independent societies in the EA.

One would think that network theory would have tests for the effects of multiple homogeneous clustering, but it does not. “We lack an information theory of data structures (e.g., graphs, sets, social networks, chemical structures, biological networks)” (Szpankowski 2012).

Etics: behavior studied from a position outside the system.

Skyhorse (2003), given Goodenough’s genealogical network dataset, eventually solved the "emic/etic" controversy over Chuukese residence rules by means of
network analysis: couples were found to reside with husband’s or wife’s matrilineage relatives depending on the pragmatics of which relatives had garden land that would devolve to the couple.

31 Emics: behavior studied from within the system.