1. Introduction

"Galton's problem," named after the nineteenth century English statistician Francis Galton, occurs whenever the observations in a regression model may be related through borrowing or descent. The problem was first stated in 1888, when Galton gently criticized a study presented to the Royal Anthropological Institute by Edward Tylor, the prominent anthropologist. Tylor had attempted to show that simple societies typically define kin relationships matrilineally, while complex societies define kin relationships patrilineally. Tylor was trying to state this as a general evolutionary rule, and he proceeded by collecting evidence for over 300 societies and showing the associations (or as he called them, "adhesions") between traits in tables and diagrams (Tylor 1889). Galton, who was present at Tylor's lecture, pointed out that the association between complexity and patrilineality need not be functional, but rather could be due to borrowing or descent. For example, societies with high complexity and high patrilineality could have obtained those characteristics because the traits were inherited from a single common ancestor, which just happened to have high values for both traits. Without controlling for borrowing and common descent, Galton maintained, one cannot make valid inferences regarding functional relationships (Stocking 1968: 175; Tylor 1889: 270).

Galton's comments have stimulated much thought in cross-cultural research, and there exists a certain amount of pessimism as to whether it is even possible to define independent cultural units for statistical analysis (e.g., Gatewood 2000). Others view Galton's problem as an opportunity to tease from the data relationships of borrowing and descent, in addition to functional relationships, though the researcher must take pains to discover the independent cultural units unique to each analysis (e.g., de Munck and Korotayev 2000; Korotayev and de Munck 2003; Mace and Holden 2005). Still others are more sanguine, and argue that the appropriate independent cultural unit is the "society," defined as a community speaking the same language located in a single territory, and that sampling techniques used to create ethnographic databases help ensure that these units are independent (Ember and Ember 2000). Nevertheless, autocorrelation studies of variables in the Standard Cross-Cultural Sample (Dow and Eff 2008; Eff 2004a), and of variables in international datasets (Eff 2004b) show that relations of both borrowing and descent are widespread in these data. Any serious effort to estimate econometric models using cross-cultural or international data must adopt appropriate techniques to test and correct for Galton's problem.

A particularly promising approach to Galton's problem was developed in a series of papers published in the 1980s. Working with ideas originally from geography, these anthropologists used spatial statistics and spatially lagged models to study the relationships among societies, and incorporated those relationships in regression models (Loftin 1972; Loftin and Ward 1983; White, Burton, and Dow 1981; Dow, Burton, and White 1982; Dow, White, and Burton 1982; Dow, Burton, Reitz, and White 1984).

In a regression context, Galton's problem causes spatially autocorrelated errors: societies linked through borrowing or through common descent will have similar error term values. The estimated coefficients will be biased (due to the absence of the true influences attributable to borrowing and descent) and the standard errors will be biased, so that inference is impossible (Anselin 1988). Correcting this problem requires the use of spatial weights matrices; one insight of the literature from the 1980s was that language phylogeny matrices could be used for relationships of descent, and geographical distance matrices for relationships of borrowing. Incorporating both of these effects into a single model is most easily accomplished with a biparametric spatial lag model (Brandsma and Ketellepper 1979; Dow 1983; Dow 2007):

$$y = X\beta + \lambda_L W_L y + \lambda_D W_D y + \varepsilon \qquad (1)$$

Where y is an *nx1* vector, **X** is an *nxk* matrix of independent variables, β is a *kx1* coefficient vector, W_L and W_D are the row-standardized weight matrices, for language and distance respectively, and ε is the error vector. The scalars λ_L and λ_D are the spatial lag parameters, allowing an estimate of the effects of common descent or cultural borrowing on y. The spatial lags y_L=W_L y and y_D=W_D y are correlated with the error term ε, and therefore endogenous—a problem most easily corrected using two-stage least squares (Dow 2007).

Spatial weight matrices for other relations (subsistence types, cultural complexity, religion, ecology, imperialism, trade, etc.) can easily be introduced (Dow 2007; Dow and Eff 2008; Eff 2004a; Eff 2004b). It is an empirical question whether other forms of spatial dependence should be considered in any given modeling exercise. For some forms of dependence a spatial lag may not be needed: religious or topographical influences (de Munck and Korotayev 2000) can be modeled by introducing dummy variables (e.g., a variable equal to one when the society is Muslim, and equal to zero otherwise).

The biparametric spatial lag model never quite caught on within cross-cultural studies, and was also ignored in economics, despite the fact that any regression model with nations as individual observations will have meaningful cultural similarities among some of the observations. The neglect of Galton's problem can probably be attributed to the limited availability of software for spatial statistics, as well as to the lack of attention given to the creation of weight matrices, especially matrices measuring non-geographic distance among nations. The software problem has now been solved, particularly with the development, by

Roger Bivand and his collaborators, of the “spdep” package for R (Bivand 2007). This paper seeks to initiate a discussion of the weight matrix problem by documenting a method for producing a weight matrix based on language phylogeny. The method is applied to produce two matrices: one for the 186 societies in the Standard Cross-Cultural Sample; the other for 216 contemporary nations. The matrices are freely available for interested researchers to download.

2. Language phylogeny as cultural proximity

Language phylogenies classify languages into categories each of whose members can be hypothesized to have spoken a common ancestral language. Thus, for example, the members of the Germanic branch of Indo-European are classified together because each of the languages is believed to be a descendant of a single language 2,500 years or so ago. The classification is based on a putative genetic relationship, not on lexical similarity, and in fact the Germanic language English has borrowed so heavily from Romance languages that from the perspective of lexical similarity it may be even more similar to a Romance tongue like French than it is to the Germanic language German.

Whereas phylogenies delineate the links among genetically related languages, a *Sprachbund* classifies a group of genetically unrelated but spatially proximate languages that have become similar due to borrowing—a phenomenon seen in South Asia, Ethiopia, the Balkans and elsewhere. Extensive borrowing from other languages can make it difficult to reconstruct genetic relationships. The further back in time genetic relationships are drawn, the more uncertain they become, and most historical linguists are unwilling to suggest super-families that tie together the current widely recognized language families. A minority, however, are more daring, and one of the most widely discussed super-families—called “Nostratic”—is hypothesized to contain Indo-European, Afro-Asiatic, Dravidian, Uralic, and Altaic (Ruhlen 1994).

Empirical work has shown that a great deal of culture is vertically transmitted together with language, so that speakers of languages descended from a common ancestor will often have many similar cultural traits (Mace and Holden 2005: 116). In cases where a native language is replaced by a new language, one generally sees “the adoption of most or all of a whole range of elements” from the culture represented by the new language (Mace and Pagel 1994: 552). Nevertheless, language is only one part of culture, and speakers of different languages may find themselves sharing many cultural traits (such as Bosnian Muslims and Turkish Muslims), while speakers of the same language may differ in important aspects (such as Bosnian Muslims and Bosnian Orthodox) (de Munck and Korotayev 2000).

Since language is usually transmitted from parent to child, as are genes, much attention has been given to unraveling the extent to which language phylogenies match phylogenies created from biometric data. Some studies have claimed a fairly close match (e.g., Cavalli-Sforza et al. 1994), others deny a match (Bateman et al. 1990), while yet other studies find conditional agreement (e.g., Nettle 2003). The potential agreement between language phylogenies and genetic phylogenies has intrigued some archaeologists (e.g., Bellwood 2005; Renfrew 2001) who hypothesize that very large language phyla stem from languages spoken in Neolithic hearth areas, and that demic expansion of these early farmers led to the worldwide spread of descendant languages. Bellwood (2005: 216-217) maintains that the Indo-European, Dravidian, and Afro-Asiatic language families all descend from a language spoken in the Neolithic hearth area in the Fertile Crescent. He suggests that the Uralic language family has its source in late Paleolithic peoples in Europe, and that its similarities with Indo-European are due to borrowing. Language isolates and small phyla are therefore likely to descend from foraging peoples who managed to survive demic expansions of farmers, and major phyla such as Austronesian are the linguistic traces of those demic expansions.

Current research is examining the archaeological, genetic, and linguistic evidence for prehistoric population movements, and a consensus view of the relatedness among world populations may eventually emerge. While there remain problems with genetic data (Fix, this volume), a steady body of evidence is accumulating and methods are becoming more sophisticated (Cann 2001: 1744). At the same time, phylogenetic tools developed in cladistics are being applied to language and other cultural data, deepening our knowledge not only of language chronologies, but of the vertical and horizontal transmission of cultural traits (Mace and Holden 2005; Lipo et al. 2006).

For now, though, the more prudent approach is to avoid grouping well-recognized language phyla into super-families such as Nostratic, instead adopting the traditional language taxonomic system, characterized by multiple families, each containing many languages but without any connection among the families. Similarity measures can therefore be calculated for languages that are members of the same family, but languages in different families are assumed to be completely unrelated.

3. Extracting proximity from a language phylogeny

A language phylogeny has a network structure, forming a tree with a single root and multiple branches with contemporary languages located at the tips. Each fork in the tree is an ancestral language, shared by all of the languages located tip-ward from it. Moving from a tip toward the root, one passes through successive ancestral languages, shared by a successively larger proportion of the tree's other tips, until one reaches the root.

Figure 1 gives an example showing a few of the languages in the Indo-European phylum. Arrayed around the periphery of the graph are nodes representing distinct living languages (Danish, French, Dari, etc.). The graph is directed, with links pointing away from living languages back to ancestral languages. Stepping from the node of a living language takes one to higher taxonomic levels, each of which can be considered (somewhat loosely) to be an ancestral language from which all of the connecting living languages descended. Thus English and Standard German both connect (through a directed path) to a node labeled “West Germanic”, and this can be interpreted as the ancestral language from which both English and German evolved.

The distance from the tip to the root, measured in chronological time, is the same for each contemporary language. But a difficulty arises, since the number of nodes on the path from tip to root will be different for each language. Languages with more living relatives will have more nodes in their path to the root, since each node represents a fissioning event in which several languages split off from a common ancestor. The path from an Austronesian language to its root, for example, may pass through as many as 13 nodes, while a language isolate will move in one step to the pan-human root. Therefore, one cannot use path length as the distance from a language to its root. What makes this even more difficult is that there are no estimates for the chronological distance between ancestral languages in these phylogenies,[1] so there is no way to translate path length into chronological time.

[1] Current research using phylogenetic techniques from cladistics has produced estimates of chronological distance within some of the larger language phyla, such as Indo-European and Austronesian (Mace and Holden 2005: 117-118).

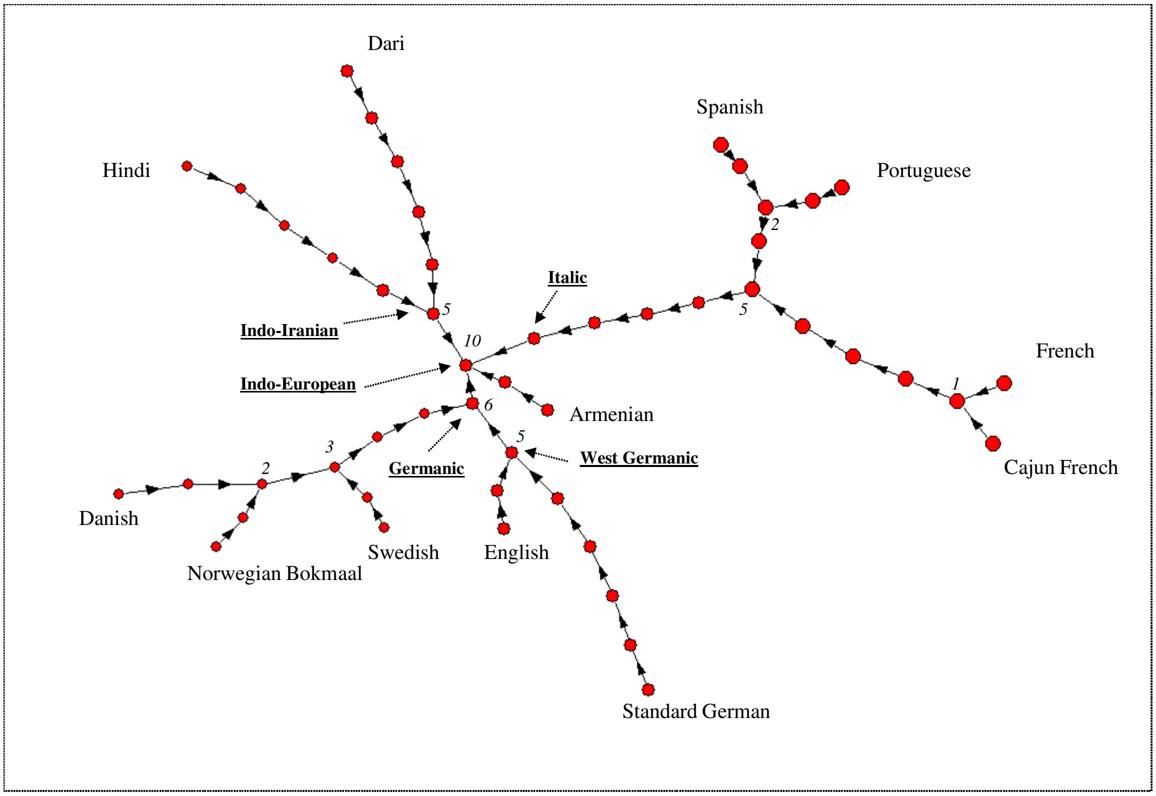

Figure 1: The relationships among 12 languages in the Indo-European phylum. The 12 selected languages are on the periphery of the digraph. Links point toward higher taxonomic levels, with all nodes ultimately connected to the node labeled Indo-European. The numbers indicate for selected nodes the maximum path length leading to that node. The taxonomy is from Grimes 2000.

In Figure 1, each node is assigned a number equal to the largest number of steps it takes for a living language to reach that particular node. For example, it takes English two steps to reach West Germanic, and it takes Standard German five steps: West Germanic is assigned the number five, since that is the maximum number of steps it takes to reach West Germanic from a living language. The Indo-European root is assigned the number 10, since the maximum path length leading to that node (from French and Cajun French) takes 10 steps. The pan-human root (not shown) is set as one step beyond the Indo-European root, so that the maximum path length leading to the pan-human root is 11.

From this number we create a measure of distance to the pan-human root:

$$d_{ir} = (n_r - n_i)/n_r \tag{1}$$

where n_r is the maximum path length leading to the pan-human root, and n_i is the maximum path length leading to node *i*. The distance to the pan-human root from a node is thus the distance yet to be travelled as a percentage of the total distance traveled. For any of the living languages in the above graph, distance is (11-0)/11=1. The example of English shows how this works: moving one step toward the root from the English node (to an unlabeled node), distance is (11-1)/11; moving two steps, to the West Germanic node, distance is now (11-5)/11; moving three steps, to the Germanic node, distance is (11-6)/11; moving four steps, reaching the Indo-European root, distance is (11-10)/11; and moving five steps, reaching the pan-human root, distance is (11-11)/11.

Distance to the root can then be used to create a proximity measure between any two living languages. The proximity between each language is the distance from the root to their nearest common ancestor. Thus the proximity between English and Standard German is 6/11 (the distance from the pan-human root to West Germanic, their nearest common ancestor), the proximity between English and Danish is 5/11 (the distance from the pan-human root to Germanic, their nearest common ancestor), and the proximity between English and Armenian is 1/11 (the distance from the pan-human root to Indo-European, their nearest common ancestor). The proximity between English and any language not in the Indo-European phylum would be zero. The proximity between English and itself is one.

The resulting proximity measure between any two languages will lie between zero and one. A distance measure can be created from the proximity measure by setting *distance=(1-proximity)*. A distance measure $d(x, y)$ between x and y is *metric* (i.e., behaves in the way we informally think distance should behave) when it satisfies three conditions (Rektorys 1969: 998):

1. Non-negativity: $d(x, y) \geq 0$, and $d(x, y) = 0$ if and only if $x = y$
2. Symmetry: $d(x, y) = d(y, x)$
3. Triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$

For our measure (converted to distance) these three conditions hold. The first two are fairly evident. For the first: if a language can have proximity of one only to itself, then it can have a distance of zero only to itself. For the second: proximity between language x and language y is found not by their own characteristics, but by the characteristics of a third node (that third node's distance from the root), and therefore the proximity must be the same for each of them, so that symmetry exists.

The third condition is also met. Consider any set of three nodes (x,y,z) in a tree graph with a pan-human root. Each *pair* of nodes in the set of three has a nearest common ancestor, and the set of nearest common ancestors for the set of three nodes will contain either one or two members. When there is one common ancestor, then $d(x, y) = d(y, z) = d(x, z)$, and the triangle inequality is obviously met. When there are two common ancestors, there are two proximity measures, and one would be greater than the other, due to the tree-like structure. Suppose $d(x, y) = d(x, z)$, and $d(x, y) > d(y, z)$. Then all three triangle inequalities hold: $d(x, z) \leq d(x, y) + d(y, z)$; $d(x, y) \leq d(x, z) + d(z, y)$; and $d(y, z) \leq d(y, x) + d(x, z)$. Therefore the distance is metric, with the practical implication that the measure will be well-behaved, in that it acts as we think a distance measure should act.

4. Comparison with previous attempts at linguistic proximity matrices

The pioneering constructions of weight matrices based on language phylogenies implicitly required that each language traverse the same number of nodes in its path to the root. White, Burton, and Dow (1981), and Dow, Burton, Reitz, and White (1984) defined distance as it traditionally is in graph theory: the length (measured in number of links) of the shortest path in the undirected graph connecting the two languages (Scott 2000: 68). Similarity is then calculated as the inverse of distance. This method was also used by Eff (2004a).

In Figure 1, one can find a few instances where the earlier procedure leads to undesirable results. For example, the number of links between English and German is seven, while the number of links between English and Armenian is six—which would imply that English is more closely related to Armenian than to German. The earlier procedure works well only when all branches in the phylogenetic graph have the same number of links. Since branches in a detailed taxonomy vary in the number of links they contain, one is forced to remove detail. The new procedure presented here has the advantage of allowing the full detail of linguistic phylogenies to be considered when converting the phylogeny into a weight matrix.

5. Linguistic proximity among contemporary world nations

The Summer Institute of Linguistics produces a database called *Ethnologue*, cataloging information on all of the world's languages. The database presents both taxonomic relationships and the number of speakers in each country (Grimes 2000). A few modifications were made to the *Ethnologue* taxonomy. The most important was to eliminate "Creole" as a separate family, and instead to treat each creole language as a sister language to the language from which it derives most of its vocabulary. In addition, unclassified languages and deaf sign languages were eliminated, thus dropping nearly 50 million speakers worldwide from the data.

After these modifications, 6,017 languages remained. While following the conservative approach of ignoring possible linguistic super-families such as

Nostratic, it nevertheless seemed prudent to use one very controversial language super-family: Amerind, postulated by Joseph Greenberg (1987). Without this super-family, the many different American Indian language phyla would have been unrelated to each other. Since it seems obvious that American Indians are more closely related to each other than to Old World peoples, this step seemed not only permissible, but necessary. Na-Dene and Eskimo are the only two New World language families that do not fit into the Amerind super-family.[2]

For each of the 6,017 languages the proximity was calculated to each of the other 6,017 languages, using the method described above of finding the distance of their common ancestor from the pan-human root. This gives a 6,017 by 6,017 matrix **S**, where each element s_{ij} is the proximity of language i to language j.

The linguistic proximity w_{rk} between countries r and k is calculated as follows:

$$w_{rk} = \sum_i \sum_j p_{ik} p_{jr} s_{ij} \tag{2}$$

where p_{ik} is the percentage of the population in country k speaking language i, p_{jr} is the percentage of the population in country r speaking language j, and s_{ij} is the proximity measure between language i and language j. Thus, every language in country r is compared to every language in country k. Intuitively, the measure gives the expected similarity of the languages spoken by two persons, one drawn at random from each country. High values of w_{rk} will occur only when both countries have a high percent of their population in similar languages.

Note that country r's self-similarity w_{rr} provides a measure related to *Ethno-Linguistic Fractioning* (Easterly and Levine 1997). That measure is based on the Herfindahl Index[3] and is the probability that two persons drawn at random from the population speak *different* languages. The diagonal of the language phylogeny matrix, however, gives the *expected similarity* of the languages spoken by two persons drawn at random from that country's population.

Perhaps surprisingly, expected similarity is often higher when drawing speakers from two different countries than when drawing both from the same

[2] The main differences between Eff (2004b) and the present derivation of language similarity is that the former considered only the thousand or so languages with more than 1,000 speakers, did not use Amerind as a super-family, and calculated cultural proximity among only 152 of the 216 nations considered here.

[3] $ELF = 1 - \sum_k p_k^2$, where p_k is the percent of the population speaking language k.

country. For example, imagine that country *A* speaks only one language—Spanish—while country *B* has 60 percent of the population speaking Spanish and the other 40 percent speaking a completely unrelated language, Quechua. The expected similarity between *A* and *B* is w_{AB}=.6*1*1+.4*0*0=0.60. The expected similarity between *B* and *B* is w_{BB} =.6*.6*1+.6*.4*0+ .4*.6*0+.4*.4*1= .36+.16=0.52. In fact, if a country has more than 50 percent of its speakers in a single language (and that language is unrelated to the country's other languages), expected similarity of two speakers within the country will be lower than with a country that has 100 percent of its population in that single language.

The elements w_{rk} of proximity matrix **W** are symmetric, and will lie between zero and one. Converted to distance ($distance_{rk}$=1- w_{rk}), examination of every off-diagonal triad in the matrix shows that the triangle inequality holds with only a few trivial exceptions.[4] Nevertheless, the first condition for a metric space (non-negativity) is violated because the distance of a country from itself is not always equal to zero. Thus, even though the matrix of distances among languages is metric, the matrix of distances among countries is *not* metric: the distance measures don't behave exactly in the way we conceive that distance should behave.

Nevertheless, since the problem is limited to the diagonal, and the diagonal is set to zero for use as a weight matrix, the *weight matrix* would behave like any metric weight matrix. In addition, the weights have a clear intuitive meaning: the expected similarity of the languages spoken by two people, one drawn at random from each of the two countries.

Before using the proximity matrix **W** as a spatial weight matrix, the diagonal should be set to zero, and the matrix then standardized by dividing each element w_{rk} by the row sums, so that rows sum to one. The matrix is available online in an Excel 2003 workbook.[5]

6. Linguistic proximity in the Standard Cross-Cultural Sample

One of the most widely used data sets in cross-cultural research is the Standard Cross-Cultural Sample (SCCS). George Peter Murdock and Douglas R. White initiated the development of the SCCS as an attempt to mitigate Galton's

[4] Of the 1,656,360 unique triads, 16 violated the triangle inequality. In all of these cases, the distance from Chile, Colombia, Spain, or Nicaragua to El Salvador, is very slightly longer (1.1E-16) than going from Chile, Colombia, Spain, or Nicaragua to either Cuba or Uruguay, and from there to El Salvador.

[5] http://www.mtsu.edu/~eaeff/downloads/cntryprox.xls Since the diagonal provides an intuitive measure of cultural diversity for each country, it has been retained in the spreadsheet.

problem. Their objective was to select a sample of societies relatively independent from each other, reasoning that relations of borrowing and descent would be weak among the sample members. Murdock and White divided a set of over 1,200 ethnographically known peoples into 186 groups of closely related cultures. They then selected one especially well-documented society from each of the 186 groups (Murdock and White 1969). The societies are quite varied, ranging from hunter-gatherers to industrial peoples. When scholars wish to conduct a statistical study over some feature of social life, they comb the ethnographies for each of the 186 societies, and code the variables they wish to use. Over nearly four decades, the work of many researchers has led to the accumulation of about 2,000 variables in the SCCS. The journal *World Cultures* acts as the repository of the SCCS.

Despite the effort to create a sample with independent societies, Galton's problem remains an issue in the SCCS (Eff 2004a; Dow and Eff 2008). Correct use of the SCCS requires that the researcher use the methodology developed by Dow, Burton, and White in the 1980s, and further developed by Dow (2007). One important tool needed is a proximity matrix based on language phylogeny.

The *Ethnologue* data are the source for the SCCS weight matrix, as they were for the international weight matrix. Two societies speak Creole languages (the Saramacca and the Haitians) and these are converted to sister languages to English and French, respectively. To handle the problem of language isolates, a type of language super-family is employed, similar to George Murdock's use in his *Ethnographic Atlas* of the concept of "language continent" (Murdock 1967). All language phyla from the western hemisphere are assigned to the super-family "New World", and all from the eastern hemisphere are assigned to the super-family "Old World."[6]

Proximity between each of the 186 languages is then calculated as described above: for each node, distance from the pan-human root is calculated as in Equation 1; the proximity between any two languages is then given as the distance from the pan-human root of their nearest common ancestor. The resulting matrix meets all of the criteria for being metric.

Before using the proximity matrix **W** as a spatial weight matrix, the diagonal should be set to zero, and the matrix then standardized by dividing each element w_{ij} by the row sums, so that rows sum to one. A spreadsheet, available online, gives the completed matrix.[7] Included, in the rightmost columns, are the *Ethnologue* taxonomy, and the hemispheric super-family.

[6] This suggestion comes from Malcolm M. Dow.

[7] http://www.mtsu.edu/~eaeff/downloads/SCCSprox.xls In the spreadsheet, the diagonal has been set to zero, but the matrix is not row standardized.

7. Summary and caveat emptor

This paper serves to document the release of two proximity matrices based on the language phylogenies found in *Ethnologue* (Grimes 2000). The method for calculating language proximity was detailed. One matrix gives language proximity among the 186 societies of the Standard Cross-Cultural Sample, and the other gives the language proximity among 216 contemporary nations. Both matrices, formulated as spatial weights matrices, meet the criteria for being metric, and are therefore well-behaved. The matrices are available online.

The weight matrices produced in this paper are based upon current views of language history in linguistics. For a number of reasons, the weight matrices provide only imperfect approximations for relations of descent among the SCCS societies or among nations. First, language is just one part of culture, and many cultural traits (for example, those associated with religion) may descend from predecessor cultures along pathways quite independent from language (de Munck and Korotayev 2000). Second, language itself develops through both borrowing and descent, and it is often controversial whether similarities among languages are due to one process or the other (for example, the relation of Tungus to Turkic). Third, our method for calculating language proximity does not include information about the chronological age for nodes in the graph, and the resulting proximity measures are therefore only relative and coarse-grained. Fourth, language phylogenies cannot reliably be inferred far back into the past, so there is as yet no accepted way of relating language isolates and major phyla to a pan-human root.

Nevertheless, a variety of studies have shown that weight matrices based on language phylogenies do in fact capture meaningful patterns in cross-cultural and international data (Dow and Eff 2008; Dow 2007; Eff 2000a; Eff 2000b). It is therefore reasonable to use these matrices (in conjunction with matrices for geographic proximity) to address Galton’s problem within the context of biparametric spatial lag models. In this context, significance tests readily show whether or not the spatial lag term based on language belongs in the model, so the lag will only be included when it has explanatory power. Thus, there is little danger of using the language-based weight matrices when inappropriate. The more serious danger is that one might overlook the true relationships of dependence: when dependence is due to religious, trade, imperial, or other ties, it would be appropriate to use weight matrices based on these other relationships to construct spatial lag terms in the model. Users of matrices based on language phylogenies should therefore exercise caution and think carefully about alternative sources of dependence in their data.

REFERENCES

Anselin, Luc. 1988. *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer Academic Publishers.

Bateman, Richard, Ives Goddard, Richard O'Grady, V. A. Funk, Rich Mooi, W. John Kress, Peter Cannell. 1990. "Speaking of Forked Tongues: The Feasibility of Reconciling Human Phylogeny and the History of Language." *Current Anthropology* 31:1-24.

Bellwood, Peter. 2005. *First Farmers: The Origins of Agricultural Societies*. Blackwell Publishing, Ltd.

Bivand, Roger, with contributions by Luc Anselin, Olaf Berke, Andrew Bernat, Marilia Carvalho, Yongwan Chun, Carsten Dormann, Stéphane Dray, Rein Halbersma, Nicholas Lewin-Koh, Jielai Ma, Giovanni Millo,Werner Mueller, Hisaji Ono, Pedro Peres-Neto, Markus Reder, Michael Tiefelsdorf and and Danlin Yu. 2007. *spdep: Spatial dependence: weighting schemes, statistics and models*. R package version 0.4-4.

Brandsma, A. S., and R. H. Ketellapper. 1979. "A biparametric approach to spatial autocorrelation." *Environment and Planning A* 11:1-58.

Cann, Rebecca L. 2001. "Genetic Clues to Dispersal in Human Populations: Retracing the Past from the Present." *Science* 291, no. 5509 (March 2): 1742-1746.

Cavalli-Sforza, L. Luca, Paolo Menozzi, Alberto Piazza. 1994. *The History and Geography of Human Genes*. Princeton: Princeton University Press.

de Munck, Victor. 2000. "Introduction: Units for Describing and Analyzing Culture and Society." *Ethnology* 39:279-292.

de Munck, Victor, and Andrey Korotayev. 2000. "Cultural Units in Cross-Cultural Research." *Ethnology* 39:335-348.

Dow, Malcolm M. 2007. "Galton's Problem as Multiple Network Autocorrelation Effects." *Cross-Cultural Research*. 41(4): 336-363.

Dow, Malcolm M. 1984. "A Biparametric Approach to Network Autocorrelation: Galton's Problem." *Sociological Methods & Research* 13:201.

Dow, Malcolm M., Michael L. Burton, and Douglas R. White. 1982. "A Simulation Study of a Foundational Problem in Regression and Survey Research."*Social Networks*. 4:169-200.

Dow, Malcolm M., Michael L. Burton, Karl Reitz, and Douglas R. White. 1984. “Galton's Problem as Network Autocorrelation.” *American Ethnologist*. 11(4): 754-770.

Dow, Malcolm M., and E. Anthon Eff. 2008. “Global, Regional, and Local Network Autocorrelation in the Standard Cross-Cultural Sample.” *Cross-Cultural Research* 42(2): 148-171.

Dow, Malcolm M., Douglas R. White, and Michael Burton. 1982. “Multivariate Modeling with Interdependent Network Data.” *Behavior Science Research*. 17(2):216-245.

Easterly, William and Ross Levine. 1997. “Africa's Growth Tragedy: Policies and Ethnic Divisions.” *The Quarterly Journal of Economics.* 112(4): 1203-1250.

Eff, E. Anthon. 2004a. “Does Mr. Galton Still Have a Problem?: Autocorrelation in the Standard Cross-Cultural Sample.” *World Cultures.* 15(2): 153-170. (http://www.mtsu.edu/~eaeff/downloads/EffsWC15no2.pdf)

Eff, E. Anthon. 2004b. “Spatial and Cultural Autocorrelation in International Datasets”. MTSU Department of Economics and Finance Working Papers. February 2004. (http://econpapers.repec.org/paper/mtswpaper/200401.htm)

Ember, Melvin, and Carol R. Ember. 2000. “Testing Theory and Why the 'Units of Analysis' Problem Is Not a Problem.” *Ethnology* 39:349-364.

Gatewood, John B. 2000. “Distributional Instability and the Units of Culture.” *Ethnology* 39:293-304.

Greenberg, Joseph H. 1987. “Language in the Americas: Author's précis.” *Current Anthropology*. 28:647-652.

Grimes, Barbara (editor). 2000. *Ethnologue: Languages of the World* (Fourteenth edition). Dallas: SIL International.

Korotayev, Andrey and Victor de Munck. 2003. "Galton's Asset and Flower's Problem: Cultural Networks and Cultural Units in Cross-Cultural Research". *American Anthropologist* 105 (2): 353–358.

Lipo, Carl P., Michael J. O’Brian, Mark Collard, Stephen J. Shennan (editors). 2006. *Mapping Our Ancestors.* New Brunswick, New Jersey: Aldine Transaction.

Loftin, C. 1972. "Galton's Problem as Spatial Autocorrelation: Comments on Ember's Empirical Test" *Ethnology* 11: 425-35.

Loftin, C. and S. Ward. 1983. "A Spatial Autocorrelation Model of the Effects of Population Density on Fertility" *American Sociological Review* 48: 121-128.

Mace, Ruth, and Clare J. Holden. 2005. “A Phylogenetic Approach to Cultural Evolution.” *Trends in Ecology & Evolution* 20:116-121.

Mace, Ruth and Mark Pagel. 1994. “The Comparative Method in Anthropology.” *Current Anthropology*. 35(5): 549-564.

Murdock, George. P. 1967. *Ethnographic Atlas: A Summary*. Pittsburgh: The University of Pittsburgh Press.

Murdock, George P. and Douglas R. White. 1969. “A Standard Cross-Cultural Sample.” *Ethnology* 9:329-369. (http://repositories.cdlib.org/imbs/socdyn/wp/Standard_Cross-Cultural_Sample/)

Nettle, Daniel, and Louise Harriss. 2003. “Genetic and Linguistic Affinities between Human Populations in Eurasia and West Africa.” *Human Biology*. 75:331-44.

Rektorys, Karel. 1969. *Survey of Applicable Mathematics*. Cambridge: MIT Press.

Renfrew, Colin. 2001. “At the Edge of Knowability: Towards a Prehistory of Languages.” *Cambridge Archaeological Journal.* 10(1): 7-34.

Ruhlen, Merritt. 1994. *On the Origin of Languages: Studies in Linguistic Taxonomy.* Stanford, California: Stanford University Press.

Scott, John. 2000. *Social Network Analysis: A Handbook* (second edition). London: Sage Publications, Ltd.

Stocking, George W. Jr. 1968. “Edward Burnett Tylor.” *International Encyclopedia of the Social Sciences*. David L. Sills, editor, New York, Macmillan Company: v.16, pp. 170-177.

Tylor, E. B. 1889. “On a Method of Investigating the Development of Institutions; Applied to Laws of Marriage and Descent.” *The Journal of the Anthropological Institute of Great Britain and Ireland*. 18:245-272.

White, Douglas R., and Michael L. Burton, Malcolm M. Dow. 1981. “Sexual Division of Labor in African Agriculture: A Network Autocorrelation Analysis.” *American Anthropologist.* 83:824-849.