Commentary on Sokal, Oden, and Thomson's "A Problem with Synthetic Maps" (1999)

JOHN NOVEMBRE[1]

The demographic impact of the origin of agriculture in Europe has been a long-standing area of interest and controversy in human genetics. For contemporary students of human population genetics, reviewing past work on this question can serve as a wonderful case study that illuminates different approaches that one can take to study historical processes by using genetics and can highlight the inherent challenges. One chapter in that history involves Robert Sokal's contributions, including his critique of synthetic maps.

Robert Sokal became keenly interested in this question, as well as other potential factors structuring genetic diversity in Europe, and he worked with great rigor and insight. His work spanned genetics (e.g., Sokal et al. 1991), linguistics (e.g., Sokal et al. 1990), and also included the analysis of a large-scale ethnohistorical database that he created (e.g., Sokal et al. 1991). Whereas most of his work on European human genetic diversity took place in the late 1980s and early 1990s, in 1999 he published a short piece in *Human Biology* revisiting and critiquing the analysis of synthetic maps (maps based on principal components analysis, PCA) that Cavalli-Sforza and his colleagues started using in the late 1970s (Cavalli-Sforza et al. 1994; Menozzi et al. 1978; Piazza et al. 1991). This made for a lively exchange in the pages of *Human Biology*. The original piece (Sokal et al. 1999a) was published with a response from Cavalli-Sforza's group (Rendine et al. 1999), and later in the same year Sokal published a letter critiquing the Rendine et al. (1999) response (Sokal et al. 1999b).

Previous to the 1999 exchange, both Sokal and Cavalli-Sforza's groups had, in separate papers, agreed that population genetic data from Europe support the existence of clinal patterns in allele frequencies that align with the expansion of agriculture, and that such clinal patterns support the demic diffusion hypothesis. Cavalli-Sforza's group argued this point most strongly by using synthetic maps, showing that in their data the PC1 map in Europe has a northwest/southeast gradient that mirrors the expansion of agriculture (e.g., Menozzi et al. 1978). In contrast, Sokal's group used an analysis based on distance metrics (Sokal et al. 1991). Specifically, they assessed how genetic distances correlate with the timing of the origin-of-agriculture and found a correlation, even after accounting for the partial correlation due to geographic distance. So, whereas both groups

[1]Interdepartmental Program in Bioinformatics, Dept. of Ecology and Evolutionary Biology, University of California, Los Angeles, 610 Charles E. Young Dr East, Los Angeles, CA 90095-7239. Correspondence to John Novembre, e-mail: jnovembre@ucla.edu.

Human Biology, October 2012, v. 84, no. 5, pp. 623–626.

ultimately were in agreement about their conclusions regarding clines, they differed in how they reached those conclusions.

In Sokal et al.'s 1999 critique of synthetic maps, the focus is on a more general problem in population genetic studies: the incompleteness of most samples. For logistical reasons, it is challenging to get a perfectly uniform sampling over the study area. The dataset that both Sokal and Cavalli-Sforza's group analyzed had sampling locales that were geographically clustered and intervening areas that went unsampled. As Sokal highlighted, Cavalli-Sforza's team dealt with the incomplete data by spatially interpolating the observed allele frequencies along a complete grid across Europe and then running the PCA to produce the synthetic map. This approach falls into a category sometimes referred to by statisticians as using "pseudodata"—where some initial data are used to infer an intermediate variable, and then analysis is run on those intermediate values. The intermediate variables are pseudodata in the sense that they are treated as real data when they are not directly observed.

The problem of taking such an approach in a spatial context is not simply that the underlying uncertainty in downstream inferences may be underrepresented. As Sokal and colleagues made clear in their paper, spatial interpolation may falsely create autocorrelation in the data, and, "The extra spatial autocorrelation falsely enhances or distorts true trends in such data." To quantify the effect, they conducted numerical experiments with data from a previous publication (Sokal et al. 1989); they found that interpolation contributed 72% of the variance seen in the first principal component; or in terms of Moran's spatial autocorrelation statistic (I), the smoothing contributed 65% of the observed spatial autocorrelation. The critique is that this extra autocorrelation would make spatial patterns seem more important in a dataset then they are in fact.

As a pithy example to drive home the point, Sokal and colleagues created a mock dataset from a limited geographic sample and with no spatial autocorrelation in it. They then interpolated data along a full grid, and they showed how the resulting data had significant levels of spatial autocorrelation. Moreover, they showed in Figures 3 and 4 of the paper how the synthetic maps produced from such pseudodata mimic the synthetic maps of Cavalli-Sforza and colleagues. In their words, "Even data values that are entirely random with regards to spatial position (i.e., values that exhibit no true spatial autocorrelation or pattern) yield apparent trends and patterns, that once these data have been interpolated or smoothed, invite ethnohistorical interpretation by the unwary."

In their response, Rendine et al. (1999) were skeptical that the effects of smoothing were detrimental for their analyses because the dataset they used was larger, and the spatial interpolation method was different from those analyzed by Sokal. They argued the effect on autocorrelation was weak, and they considered it a "low cost to pay if the benefit is to retrieve information from a higher number of genetic markers," though acknowledging that "this may be a matter of opinion." Sokal and colleagues followed up the Rendine et al. (1999) response with a letter in *Human Biology* later that year. In their letter, "Problems with

Synthetic Maps Remain," they critique the Rendine et al. (1999) response and maintain that there are serious problems with interpolation before generating synthetic maps. They argue that the central issue of the effects of interpolation was not adequately addressed by Rendine et al. (1999), and that there remains cause for concern.

Recently, the use of PCA in population genetics was reinvigorated in the context of genome-wide association studies by Alkes Price and colleagues (2006). One major difference of more recent applications of PCA is that the method is applied on individual-level genotype data, rather than population-level allele frequency data. Also, in part, because of the availability of large samples, contemporary studies frequently use PCA without interpolation (e.g., see citations within Novembre and Ramachandran 2011). Accompanying this more widespread use, several studies on the behavior of PCA as a methodology have helped elucidate more about the properties of the method (DiGiorgio and Rosenberg 2012; Francois et al. 2010; McVean 2009; Novembre and Stephens 2008; Patterson et al. 2006). These studies make clear that Sokal's initial concerns about potential misinterpretation of PCA analyses were valid, and not just because of spatial interpolation of data prior to analysis. As a result, there is now a more subtle appreciation of PCA's benefits and pitfalls by its users.

As a broader lesson for population geneticists, Sokal's critique of the application of PCA to make inferences about demic diffusion indicates the importance of considering potential statistical missteps in an analysis chain. The details of how data are processed before analysis are often crucial, and that is no less true in today's era of next-generation sequencing and especially in low-coverage data. Sokal also appreciated how the study of population history is inherently multidisciplinary. It stands on a platform built from the diverse forms of evidence brought forward by archaeology, genetics, linguistics, and history. Sokal's legacy reminds us that each plank of that platform must be rigorously inspected for the structure to hold strong.

Acknowledgments The author wishes to thank both Charles Taylor for an enlightening conversation on Sokal's personality and intellectual background and Alex Platt for comments on a draft of this comment.

Received 31 October 2012; revision accepted for publication 6 November 2012.

Literature Cited

Cavalli-Sforza, L. L., P. Menozzi, and A. Piazza. 1994. *The History and Geography of Human Genes.* Princeton University Press.

DeGiorgio, M., and N. A. Rosenberg. 2012. Geographic sampling scheme as a determinant of the axis of genetic variation in principal components analysis. *Mol. Biol. Evol. Advanced Online Access.*

François, O., M. Currat, N. Ray et al. 2010. Principal component analysis under population genetic models of range expansion and admixture. *Mol. Biol. Evol.* 27(6):1257–1268. http://dx.doi.org/10.1093/molbev/msq010.

McVean, G. 2009. A genealogical interpretation of principal components analysis. *PLoS Genet.* 5(10):10.

Menozzi, P., A. Piazza, and L. Cavalli-Sforza. 1978. Synthetic maps of human gene frequencies in Europeans. *Science* 201(4358):786–792.

Novembre, J., and S. Ramachandran. 2011. Perspectives on human population structure at the cusp of the sequencing era. *Ann. Rev. Genomics Hum. Genet.* 12(July):245–274.

Novembre, J., and M. Stephens. 2008. Interpreting principal component analyses of spatial population genetic variation. *Nat. Genet.* 40(5):646–649.

Patterson, N., A. L. Price, and D. Reich. 2006. Population structure and eigenanalysis. *PLoS Genet.* 2(12):20.

Piazza, A., S. Rendine, E. Minch et al. 1995. Genetics and the origin of European languages.pdf. *Proc. Natl. Acad. Sci.* 92:5836−5840.

Price, A. L., N. J. Patterson, R. M. Plenge et al. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38:904–909.

Rendine, S., A. Piazza, L. Menozzi et al. 1999. A problem with synthetic maps: Reply to Sokal et al. *Hum. Biol.* 71(1):15–25.

Sokal, R. R. 1991. Ancient movement patterns determine modern genetic variances in Europe. *Hum. Biol.* 63(5):589–606.

Sokal, R. R., R. M. Harding, and N. L. Oden. 1989. Spatial patterns of human gene frequencies in Europe. *Am. J. Phys. Anthropol.* 80(3):267–294.

Sokal, R. R., N. L. Oden, P. Legendre et al. 1990. Genetics and language in European populations. *Am. Naturalist* 135(2):157–175.

Sokal, R. R., N. L. Oden, and B. A. Thomson. 1999a. A problem with synthetic maps. *Hum. Biol.* 71(1):1–13.

Sokal, R. R., N. L. Oden, and B. A. Thomson. 1999b. A problem with synthetic maps remains. *Hum. Biol.* 71(3):447–453.

Sokal, R. R., N. L. Oden, and C. Wilson. 1991. Genetic evidence for the spread of agriculture in Europe by demic diffusion. *Nature* 351(6322):143–145. http://dx.doi.org/10.1038/351143a0.