W. Webb Sprague
UC Berkeley
2008-09-22

Sequence analysis and micro-macro demographic connections.

Just as modern thermodynamics depends on the interrelation of aggregate behavior described by classical thermodynamic theory - which describes temperature, pressure, entropy, etc - and its causes at the micro level - the velocity and interaction of particles - so the field of demography should be attempting to draw the causal connections between individual decision making and aggregate demographic behavior. Lifecourse models provide a possible connection: such models are amenable to aggregation and simulation, as is evidenced by the excellent work done with them in population biology, as shown in Caswell (2001). Such models are also of extreme importance in their own right, with a huge corresponding literature, for example Elder (1994) and Shanahan (2000). If exemplar life courses can be formally characterized, their frequencies and variance structures described, and their relation to culture and psychology theorized, the discipline of demography would be advanced greatly towards connecting micro behavior to aggregate demographic results. This paper builds on the work of Aassve et al. (2003) in characterizing lifecourse trajectories using hierarchical clustering algorithms, applying these clustering techniques to the NLSY79, and explicitly arguing for a micro-macro sociological connection between the lifecourse and aggregate demographic data.

Hierarchical clustering techniques (see Brieman, 1984, for an introduction) offer a rigorous method to build life history typologies and categories.   Besides providing the hingepoint for micro-macro models as described above, these clusters yield two additional payoffs. First, the relative frequencies of category membership provide an independent variable that drives rates and other important aggregate outcomes. Equally important, a robust typology of life courses provides a starting point to understanding the influence of cultural forces on human behavior; category membership here functions as a dependent variable, and allows us to measure the likelihood to follow a trajectory based on cultural context. The importance of culture, broadly construed, on demography is now a matter of consensus, especially after the work of the Princeton Fertility Project (Coale & Watkins 1986, see also Fricke 2003 for a more recent comment on culture and demography). There are also excellent arguments that these forces can be examined fruitfully at the life course level (Caspi 1998), with its capturing of individual context, personal memory, and developmental history.

However, before we can cluster the NLSY79 data, it must be recoded in order to make it readable by the clustering algorithms. The coding scheme employed attempts to capture pattern of change in the lifecourse – such as the occurrence of a new marriage or partnership – rather than representing occupation of a given state, like whether married or not. The four variables we track are: cohabiting partnership (including marriage), employment, residence, and fertility. We code a change in each year with a 1, and no change with a 0. This approach is taken for two reasons. For one, it simplifies the coded data, as there are only four possible states for each variable – rather than trying to capture, for example, the number of total children in the lifecourse so far. Secondly, this coding scheme captures instability in employment, marriage, residence, and fertility, an instability which often corresponds to negative outcomes in education, wealth, and well-being.

Each life is thus organized into a 27 by 4 matrix, with each row corresponding to a year (starting at 1979 and ending at 2006, the last release of the data). If there is a change in a variable – a cohabiting person moves out, a new job is taken, or a new residence is occupied – a number 1 is put in that cell; if there is no change, a 0 is entered. These matrices are transformed again into 27 by 1 vectors by replacing each row with the sum of the elements of that row raised to the power: $V_i = \sum_{j=1}^{4} (2^j * M_{i,j})$, where V_i is the i'th row in the lifecourse sequence vector, and $M_{i,j}$ is the i'th row and j'th column in the lifecourse sequence matrix. Years for which there is no data are set to zero; this does not bias the results because all lifecourse sequences have zeros in the same columns and thus these zeros do not affect the clustering. Respondents who have died or censored are not analyzed.

We measure distances between lifecourse sequence vectors using Optimal Matching Analysis (OMA), a common sequence distance metric introduced by Abbott (1995); the resulting distance matrix is used to drive the clustering software. In general, to calculate a pairwise distance between two sequences, the number of minimum transformations (insertion, deletion, and substitution) necessary to transform one sequence into the other is tallied, each transformation is assigned a cost, and these costs are summed. The cost of a single substitution is determined empirically by calculating the substitution matrix for each element in V_i using the same technique used to calculate the BLOSUM50 matrix described in Durbin et al (1998:16). With our lifecourse data, insertions and deletions will not be penalized as each lifecourse sequence vector is padded with zeros for all empty years, and incomplete lifecourses are discarded. The distance matrix representing all possible pairs of sequences can be calculated using Rohwer & Potter’s (2002) TDA, given the substitution matrix and the lifecourse sequence vectors as input. This distance matrix is calculated on a training set from the NLSY79.

This distance matrix is then used as input to the R/SPLUS cluster library, which yields a data structure containing the derived clustering criteria, including a list of the resulting exemplar lifecourse sequences (the algorithm can be set to yield between five and seven clusters). This clustering criteria is applied to the remaining NLSY79 data, yielding a table of lifecourse ID numbers, date of birth, and a code for cluster membership. This table is then processed to calculate frequencies of occurrence of each type of lifecourse, and these frequencies are used to calculate a set of simulated lives, and this set is used to derive a cohort. Aggregate rates are (finally) calculated from this cohort and compared to aggregate fertility rates of similarly aged residents of the US at the same time.

Although more work can be done in this vein, especially with respect to coding lifecourses and calculating distances, we show one possible way to connect micro and macro demographic measurement using methods that already have theoretical traction, as well as begin to create the infrastructure for further research.

REFERENCES

Aassve A., Billari F.C., Piccarreta R. (2003) Sequence analysis of BHPS life course data, Book of Short Papers, CLADAG 2003, Classification and Data Analysis Group Italian Statistical Society, Bologna: 22-24 Settembre 2003, 13-16.

Abbott A. (1995) Sequence analysis: New methods for old ideas. Annual Review of Sociology, 21, 93-113.

Billari F.C., Piccarreta R. (2001) Life courses as sequences: an experiment in classification via monothetic divisive algorithms, in Advances in Classification and Data Analysis, S. Borra, R. Rocchi, M. Vichi, M. Schader (Eds.), Springer Verlag, Berlin and New York, 351-358.

Breiman L., Friedman J.H., Olshen R.A. and Stone C.J. (1984) Classification and regression trees, Wadsworth, Belmont, CA.

Caspi, Av. (1998) Personality development across the life course. Handbook of child psychology 3: 311-388.

Caswell, Hal. (2001) Matrix Population Models: Construction, analysis, and interpretation. Sunderland, MA: Sinauer Associates.

Coale, Ansley, and Susan C. Watkins. (1986) The Decline of fertility in Europe since the Eighteenth Century as a chapter in demographic history. In The Decline of Fertility in Europe, ed. Ansley J. Coale and Susan C. Watkins. Princeton: Princeton University Press.

Durbin, Richard, Sean Eddy, Anders Krogh, and Graeme Mitchison. (1998) Biological sequence analysis. Cambridge: Cambridge University Press.

Elder, Glen H. (1994) Time, human agency, and social change: perspectives on the life course. Social Psychology Quarterly 57, no. 1: 4-15.

Fricke, T. E. (2003) Culture and causality: an anthropological comment. Population and Development Review 29, no. 3: 470-479.

Piccarreta R. (2003) CART for distance or dissimilarity matrices, Studi Statistici 80, Istituto di Metodi Quantitativi, Università Commerciale “L. Bocconi”, Dicembre 2003.

Rohwer G., Pötter U. (2002) TDA user’s manual, Ruhr-Universität-Bochum.

Shanahan, Michael J. (2000) Pathways to adulthood in changing societies: Variability and mechanisms in life course perspective. Annual Review of Sociology 26: 667-69

SHORT ABSTRACT: Life course analysis is of great interest in its own right, but also has also promise as micro base for explaining aggregate demographic results. However, a formalism describing the pathway between typical lifecourses and aggregate results has not yet been developed. This theoretical gap is especially important as cultural and developmental forces are now acknowledged to have a huge impact on demographic behavior, and these forces are most clear at the life course level. I address this theoretical need by using sequence analysis clustering of lifecourses to yield a typology of life course trajectories. My approaches to categorization are based on Piccarreta & Billari (2003), who build on Abbott's (1995) work on sequence analysis. I create hierarchical clusters from NLSY79 with the "cluster" library in R. This approach shows promise at formally connecting micro and macro behavior, as well as elucidating the interaction of cultural processes with the life course