Optimal Data Analysis
Vol. 6 (January 2, 2017), 5-8

2155-0182/10/$3.00

Novometric Analysis of Transition Matrices to Ascertain Markovian Order

Paul R. Yarnold, Ph.D.
Optimal Data Analysis, LLC

The American National Election Panel Study modeled transitions in social class identification occurring between 1956, 1958, and 1960.[1] Visual analysis suggested "…that transition probabilities linking class identifications in 1958 and 1960 *vary with 1956 identification*" (italics added). Transition tables were compared using Goodman's chi-square procedure: χ^2=98.2, df=2, p<0.0001. Based on visual examination and non-disentangled[2-5] omnibus chi-square findings, the null hypothesis of a first-order Markov model was rejected in favor of the alternative hypothesis that the underlying temporal process is a second-order Markovian (pp. 13-15). In contrast, novometric analysis[5-12] indicated that a first-order Markov model is appropriate for these data.

Data used in this application[1] are presented in Table 1 (SAS™ code used to construct the data set is given in the Appendix). Social class identification in each year was dummy-coded as Working=1, and Middle=2, for all observations.

Table 1: Social Class Identification Transitions Across 1956, 1958, and 1960 Measurements[1]

		1960 (Class Variable)	
1956	1958	Middle	Working
Middle	Middle	216	70
Working	Middle	56	75
Middle	Working	42	92
Working	Working	47	549

The first analysis treated social class in 1956 and 1958 as attributes (consistent findings resulted if attributes were treated as ordered or categorical), and social class in 1960 as the class variable.[5] An exploratory enumerated-optimal classification tree analysis (EO-CTA[13-18]) identified a single optimal model: if 1958 social class =Middle, predict 1960 social class=Middle; otherwise predict 1960 social class=Working. The confusion matrix for this model in training (and in jackknife) analysis is presented in Table 2: as seen, the model correctly predicted the actual class status of 7 of 8 observations self-identified as Working class, and of 6 of 8 observations self-identified as Middle class.

Table 2: Confusion Table for EO-CTA Model

Actual	Predicted Class		
Class	Working	Middle	Sensitivity
Working	641	145	81.6
Middle	89	272	75.4

2155-0182/10/$3.00

Because this was the only optimal model identified, it thus is also the globally-optimal (GO) model in this application.[7,11] The model was statistically reliable (exact p<0.0001); it yielded relatively strong ESS=56.9 (for 10,000 bootstrap iterations, exact discrete 95% CI for *model* ESS=50.6-63.2; for 10,000 Monte Carlo experiments, exact discrete 95% CI for *chance* ESS=0.1-6.0); and it achieved stable (identical) classification accuracy in training and leave-one-out (one-sample jackknife[19-21]) analysis. For this model D=1.51 (exact discrete 95% CI= 1.17-1.95).[7,11]

In contrast to the conclusion reached on the basis of visual examination and omnibus chi-square—that 1960 class status identification was related to both 1956 *and* 1958 self-classifications (suggesting a second-order Markovian), novometric analysis found that only the self-classifications recorded in 1958 predicted the self-classifications that were made two years later. The novometric result thus supports the use of a first-order Markov model.

Accordingly, a second novometric analysis was conducted predicting 1958 class self-identification (class variable) as a function of 1956 self-identification (attribute). As before, treating the attribute as ordered or categorical did not affect the result. And, conducting a confirmatory analysis—replicating the prior model, versus an exploratory analysis, also did not affect results presently. Using EO-CTA a single GO model was identified: if 1956 social class= Middle, then predict 1958 social class=Middle; otherwise predict 1958 social class=Working.

Table 3 is the confusion matrix for this model in training and jackknife analysis: the model correctly predicted the actual class status of 7 of 8 observations self-identified as Working class, and 7 of 10 observations as Middle class. The model was statistically reliable (exact p< 0.0001); achieved moderate-to-relatively strong ESS=50.2 (for 10,000 bootstrap iterations, exact discrete 95% CI for *model* ESS=43.9-56.5; for 10,000 Monte Carlo experiments, exact discrete 95% CI for *chance* ESS=0.1-5.8); and had stable classification accuracy in training and leave-one-out analysis. For this model D=1.98 (exact discrete 95% CI=1.54-2.56).

Table 3: Confusion Table for EO-CTA Model

Actual	Predicted Class		
Class	Working	Middle	Sensitivity
Working	596	134	81.6
Middle	131	286	68.6

The exact 95% CIs for ESS and D of the two novometric models overlap, indicating the models do not differ significantly with respect to predictive accuracy normed against chance (ESS) and also against parsimony (D). Taken together the novometric analyses offer moderate to relatively strong evidence of a reproducible first-order Markov model in this application.

References

[1]Markus GB (1979). *Analyzing panel data*. Beverly Hills, CA: Sage (pp. 13-15).

[2]Yarnold PR (2016). UniODA *vs*. chi-square: Describing baseline data from the National Pressure Ulcer Long-Term Care Study (NPULS). *Optimal Data Analysis*, *5*, 24-28.

[3]Yarnold PR (2016). CTA *vs*. disintegrated chi-square: Integrated *vs*. piecemeal analysis. *Optimal Data Analysis*, *5*, 118-120.

[4]Yarnold PR (2016). CTA vs. non-disentangled omnibus chi-square: Comparing samples (not) selected for study participation. *Optimal Data Analysis*, *5*, 154-157.

[5]Yarnold PR, Soltysik RC (2014). Globally optimal statistical classification models, I: Binary class variable, one ordered attribute. *Optimal Data Analysis*, *3*, 55-77.

2155-0182/10/$3.00

[6]Yarnold PR, Soltysik RC (2014). Globally optimal statistical classification models, II: Unrestricted class variable, two or more attributes. *Optimal Data Analysis*, *3,* 78-84.

[7]Yarnold PR, Soltysik RC (2016). *Maximizing predictive accuracy*. Chicago, IL: ODA Books. DOI: 10.13140/RG.2.1.1368.3286

[8]Yarnold PR, Linden A (2016). Novometric analysis with ordered class variables: The optimal alternative to linear regression analysis, *Optimal Data Analysis*, *5,* 65-73.

[9]Yarnold PR, Bennett CL (2016). Novometrics *vs*. correlation: Age and clinical measures of PCP survivors, *Optimal Data Analysis*, *5,* 74-78.

[10]Yarnold PR, Bennett CL (2016). Novometrics *vs*. multiple regression analysis: Age and clinical measures of PCP survivors, *Optimal Data Analysis*, *5*, 79-82.

[11]Yarnold PR, Linden A (2016). Theoretical aspects of the D statistic. *Optimal Data Analysis*, *5*, 171-174.

[12]Yarnold PR (2016). Using novometrics to disentangle complete sets of sign-test-based multiple-comparison findings. *Optimal Data Analysis*, *5*, 175-176.

[13]Linden A, Yarnold PR (2016). Using data mining techniques to characterize participation in observational studies. *Journal of Evaluation in Clinical Practice*, *6*, 839-847.

[14]Yarnold PR (1996). Discriminating geriatric and non-geriatric patients using functional status information: An example of classification tree analysis via UniODA. *Educational and Psychological Measurement*, *56*, 656-667.

[15]Yarnold PR, Soltysik RC, Bennett CL (1997). Predicting in-hospital mortality of patients with AIDS-related *Pneumocystis carinii* pneumonia: An example of hierarchically optimal classification tree analysis. *Statistics in Medicine, 16*, 1451-1463.

[16]Soltysik RC, Yarnold PR (2010). Automated CTA software: Fundamental concepts and control commands. *Optimal Data Analysis, 1*, 144-160.

[17]Yarnold PR, Bryant FB (2015). Obtaining an enumerated CTA model via automated CTA software. *Optimal Data Analysis*, *4*, 54-60.

[18]Yarnold PR (2015). Optimal statistical analysis involving a confounding variable. *Optimal Data Analysis*, *4,* 87-103.

[19]Linden A, Yarnold PR (2016). Using machine learning to identify structural breaks in single-group interrupted time series designs. *Journal of Evaluation in Clinical Practice*, *6*, 855-859.

[20]Linden A, Yarnold PR, Nallamothu BK (2016). Using machine learning to model dose-response relationships. *Journal of Evaluation in Clinical Practice*, *6*, 860-867.

[21]Yarnold PR, Soltysik RC (2005). *Optimal data analysis: A guidebook with software for Windows*. Washington, DC, APA Books.

[22]Bryant F, Harrison PR (2013). How to create an ASCII input data file for UniODA and CTA software (Invited). *Optimal Data Analysis*, *2*, 2-6.

[23]Ebert TA (2016). Getting started with ODA software: A short tutorial (Invited). *Optimal Data Analysis*, *5*, 235-240.

Author's Notes

Analyzed data are publically available, and no conflict of interest was reported.

2155-0182/10/$3.00

Appendix

SAS™ Code used to Construct (Reproduce[1]) the Data File for Analysis by ODA Software[7,16,17,22,23]

```
data real;
infile datalines;
input 1956 1958 1960;
cards;
1 1 1
;
Data example;
Do n=1 to 216;
put '2 2 2';
end;
Do n=1 to 70;
put '2 2 1';
end;
Do n=1 to 56;
put '1 2 2';
end;
Do n=1 to 75;
put '1 2 1';
end;
Do n=1 to 42;
put '2 1 2';
end;
Do n=1 to 92;
put '2 1 1';
end;
Do n=1 to 47;
put '1 1 2';
end;
Do n=1 to 549;
put '1 1 1';
end;
Output;
Run;
```