Optimal Data Analysis
Vol. 6 (January 3, 2017), 9-12

2155-0182/10/$3.00

Novometric Comparison of Markov Transition Matrices for Heterogeneous Populations

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

The American National Election Panel Study modeled transitions in social class (Working versus Middle) identification occurring between 1956, 1958, and 1960, and although a first-order Markov model was judged unsatisfactory, insufficient measurements were available to validate higher-order models.[1,2] The sample was thus stratified with respect to whether respondents' perceived their financial condition to have worsened or improved between data collection waves, in order to "…significantly decrease within-strata and increase between-strata heterogeneity of transition rates".[1] The probability of working class identification in 1958 was hypothesized to be greatest for respondents who "…felt their financial status declined, identification in 1956 held constant", and was confirmed by visual examination. Non-directional omnibus chi-square analysis showed "stratification of the sample by perceived financial change (yielded) significantly different transition matrices" (χ^2=14.9, df=2, p<0.01).[1] In contrast, novometric analysis found social class identification in 1958 was predicted by social class identification in 1956, and not by perceived financial status.

Data used in this application are presented in Table 1 (the Appendix lists SAS™ code used to reconstruct the data set[1]). Social class identification (Working=1, Middle=2) and perceived change in financial status (Worse=0, Better=1) were dummy-coded. The novometric analysis treated social class identification in 1956 and perceived change in financial status as attributes (consistent findings resulted if attributes were treated as ordered or categorical), and social class in 1958 as the class variable.[2-10]

Table 1: Social Class Identification Transitions and Perceived Change in Financial Status[1]

		1958 (Class Variable)	
Finances	1956	Middle	Working
Better	Middle	93	23
Better	Working	26	95
Worse	Middle	28	22
Worse	Working	7	66

2155-0182/10/$3.00

Paralleling earlier use of nondirectional legacy analysis[1], the first novometric analysis conducted exploratory enumerated-optimal classification tree analysis (EO-CTA[9-14]) and a single optimal model emerged: if 1956 social class=Middle, predict 1958 social class=Middle; otherwise predict 1958 social class=Working. The confusion matrix for this model in training (and in jackknife) analysis is given in Table 2: as seen, the model correctly predicted the actual class status of 7 of 9 of the observations self-identified as Working class, and of 7 of 9 of the observations self-identified as Middle class.

Table 2: Confusion Table for EO-CTA Model

Actual	Predicted Class		
Class	Working	Middle	Sensitivity
Working	161	45	78.2
Middle	33	121	78.6

Because this was the only optimal model identified, it therefore is the globally-optimal (GO) model in this application.[5,9] The model was statistically reliable (exact $p<0.0001$); it yielded moderate-relatively strong ESS=56.7 (for 10,000 bootstrap iterations, exact discrete 95% CI for *model* ESS=46.5-66.7; for 10,000 Monte Carlo experiments, exact discrete 95% CI for *chance* ESS=0.01-10.2); and it had stable (identical) classification accuracy in training and leave-one-out (one-sample jackknife[17-19]) analysis. For this model D=1.53 (exact discrete 95% CI=2.30-1.00).[5,9] In contrast to the finding obtained on the basis of visual examination and omnibus chi-square—that perceived change in financial status predicted 1958 social class self-classifications (with self-identified social class in 1956 held constant), novometry found that only the self-classifications recorded in 1956 predicted self-classifications made in 1958.

The second novometric analysis used EO-CTA to test the *a priori* hypothesis[1] that perceived worsening of financial status occurring between 1956 and 1958 (the attribute) predicted comparatively greater identification as a member of the Worker class in 1958 (the class variable). A single optimal model emerged: if perceived finances are worse predict class= Worker; otherwise predict class=Middle. The confusion matrix for this model in training (and jackknife) analysis is given in Table 3: as seen, the model correctly predicted the actual class status of 2 of 5 of observations self-identified as Working class (a sensitivity of 50 is expected by chance), and of 7 of 9 of the observations self-identified as Middle class.

Table 3: Confusion Table for EO-CTA Model

Actual	Predicted Class		
Class	Working	Middle	Sensitivity
Working	88	118	42.7
Middle	35	119	77.3

This was the only optimal model that emerged and so is the GO model in this application. Although the model was statistically reliable (exact $p<0.0001$), it yielded relatively weak-moderate ESS=20.0: for 10,000 bootstrap iterations, exact discrete 95% CI for *model* ESS=8.5-31.2. And, for 10,000 Monte Carlo experiments, the exact discrete 95% CI for *chance* ESS=0.44-9.8. Because the lower bound of the CI for the model (ESS=8.5) overlaps the upper bound of the CI for chance (ESS=9.8), the predictive accuracy achieved by the model versus by chance are coincident with prevalence that exceeds the criterion Type I error rate. For this model D= 8.00 (exact discrete 95% CI= 21.4-4.40). Thus, based on point estimates, the model for financial status is [(8.0/1.53)x 100]-100, or 422.9% further from its theoretically ideal counterpart than is the model based on social class identification in 1956.

References

[1]Markus GB (1979). *Analyzing panel data*. Beverly Hills, CA: Sage (pp. 15-16).

2155-0182/10/$3.00

[2]Yarnold PR (2016). Using novometric analysis of transition matrices to ascertain Markovian order. *Optimal Data Analysis, 6*, 4-7.

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

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

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

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

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

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

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

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

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

[12]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.

[13]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.

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

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

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

[17]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.

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

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

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[5]

```
data real;
infile datalines;
input finance 1956 1958;
cards;
1 1 1
;
Data example;
Do n=1 to 93;
put '1 2 2';
end;
Do n=1 to 23;
put '1 2 1';
end;
Do n=1 to 26;
put '1 1 2';
end;
Run;
Do n=1 to 95;
put '1 1 1';
end;
Do n=1 to 28;
put '0 2 2';
end;
Do n=1 to 22;
put '0 2 1';
end;
Do n=1 to 7;
put '0 1 2';
end;
Do n=1 to 66;
put '0 1 1';
end;
Output;
```