Optimal Data Analysis
Vol. 4 (December 20, 2015), 182-183

2155-0182/10/$3.00

Globally-Optimal CTA Model of World War II Recruit Training Camp Location Preference

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

A globally-optimal classification tree analysis (GO-CTA) model yields strong predictive accuracy in modeling the training camp location preference of N = 8,036 WWII recruits.

Data available for every recruit included *race* (Caucasian, African American), *region* of origin (north, south), *location* of present camp (north, south), and where the recruit *prefers* to be located (north, south).[1] The minimum-denominator selection algorithm (MDSA) was used to identify the descendant family of all possible enumerated-optimal CTA (EO-CTA) models that exist in this application (all four family models had sufficient statistical power).[2] The unrestricted initial (most granular) model in the family was identified using the following CTA software[3] syntax:

```
OPEN recruit.dat;
OUTPUT recruit.out;
VARS race region location prefers;
CLASS prefers;
ATTRIBUTE race region location;
CATEGORICAL race region location;
MC ITER 5000 CUTOFF .05 STOP 99.9;
PRUNE .05;
ENUMERATE;
GO;
```

Table 1 is a summary of the descendant family of four EO-CTA models: model number indicates order of discovery by MDSA; N_{MIN} is the size of the smallest (least populated) strata (endpoint) in the model; *ESS* is a chance- and maximum-corrected measure of predictive accuracy; and *D* is the number of additional equivalent effects needed to obtain a theoretically ideal statistical model in this application.[2]

Table 1: The Descendant Family

Model	N_{MIN}	Strata	*ESS*	*D*
1	280	4	55.0	3.28
2	933	3	53.1	2.65
3	2,473	3	53.1	2.65
4	3,986	2	52.3	1.82

Model 4 (Figure 1) has the lowest *D* statistic and is therefore the globally-optimal classification tree analysis (GO-CTA) model. *ESS* yielded by the model was statistically significant ($p < 0.0001$) and reflected a relatively strong effect: the exact discrete 95% confidence interval (CI) for *ESS* is 50.1 - 54.4 for the model, and is 0.07 - 21.2 for chance.

2155-0182/10/$3.00

Figure 1: GO-CTA Model Predicting Training Camp Preference

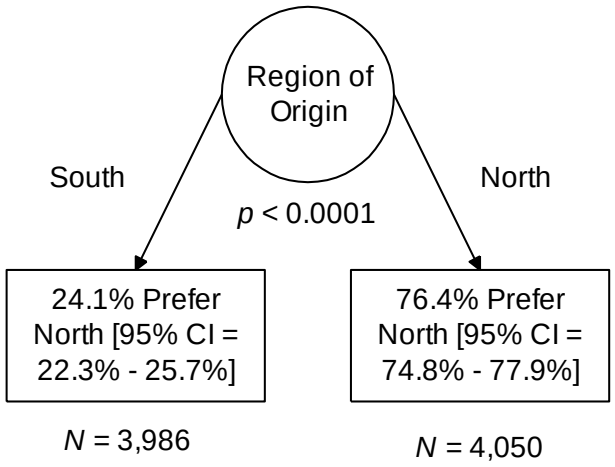

Table 2 presents the confusion table summarizing the predictive accuracy of the GO-CTA model: one in four recruits originally from the south preferred a training camp in the north, compared with three in four recruits originally from the north.

Table 2: GO-CTA Model Predictive Accuracy

		Predicted Preference		
		North	South	Sensitivity
Actual	North	3,027	958	76.8%
Preference	South	959	3,092	76.3%
Predictive Value		75.9%	76.4%	

References

[1]Gilbert N (1993). Analyzing tabular data: Log-linear and logistic models for social researchers. London, England: UCL Press (pp. 108-109).

[2]Yarnold PR, Soltysik RC (In Review). *Maximizing predictive accuracy*. Chicago, IL: ODA Books.

[3]Soltysik RC, Yarnold PR (2010). Automated CTA software: Fundamental concepts and control commands. *Optimal Data Analysis, 1*, 144-160. URL: http://odajournal.com/2013/09/19/62/

Author Notes

The study analyzed de-individuated data and was exempt from Institutional Review Board review. No conflict of interest was reported.

Mail: Optimal Data Analysis, LLC
6348 N. Milwaukee Ave., #163
Chicago, IL 60646
USA