From “Topaz from Mason County, Texas”
Roy Bassoo, Diane Eames, Matthew F. Hardman, Kenneth Befus, and Ziyin Sun
Gems & Gemology, Vol. 59, No. 4, pp. 414–431

APPENDIX 1: STATISTICAL METHODS

All statistical analysis in this study is conducted using R version 4.3.1.

DATASET PREPARATION

Trace-element compositions are provided for all samples, as ppmw, for the elements ^{45}Sc, ^{47}Ti, ^{72}Ge, ^{53}Cr, ^{31}P, ^{51}V, ^{71}Ga, ^{182}W, ^{93}Nb, ^{57}Fe, ^{181}Ta, and ^{118}Sn. For analytes where the measured concentration is at or below the lower limit of detection (LLD), the LLD value is imputed, and used in all calculations.

All topaz are assigned to one of two classes based on their provenance (“Texas” and “Not from Texas”).

RANDOM FOREST MODEL CALIBRATION

The random forest (RF) machine learning model is a robust statistical approach to the determination of topaz provenance. It is a supervised technique that generates large ensembles of decisions trees, with the majority vote, aggregated from all trees, taken as the final outcome for a tested sample (Breiman, 2001). In this study we apply the randomForest() function from the *randomForest* library in R. For this study we define the classification problem as a binary “Texas” vs. “Not from Texas” problem.

Due to the fact that topaz in the “Not from Texas” class are more abundant than in the “Texas” class, during model calibration we have defined the *sampsize* parameter to be equal to the abundance of the minority class (here, “Texas”). This reduces bias in the RF model, associated with imbalanced calibration data. The model is calibrated by randomly selecting a set of samples from the minority and majority classes in equal abundance to the minority class (*sampsize*), and then generating a decision tree. This is equivalent to a “balanced” random forest approach for model calibration using imbalanced datasets (Chen et al., 2004). Additional parameters for model calibration include *ntree*, the number of decision trees to be generated (here, *ntree* = 1500), and *mtry*, the number of variables to be randomly attempted for calibration at each branch in the decision tree (here, *mtry* = 4).

MODEL VALIDATION

The final RF model can be validated using held-out data that are not used to calibrate the model, by assessing if the held-out test data are accurately classified by the model. Due to the fact that for each topaz we have acquired replicate analyses, during model validation we have ensured that replicate analyses only occur in the validation datasets but not in the calibration datasets, to avoid overestimating the success of the final model.

For each topaz and its replicate analyses, we have withheld these data and calibrated a RF model using the remaining topaz data. We then validated the strength of the model using the withheld analyses. We repeated this procedure for every topaz in the dataset and then summed the total errors, equivalent to a *k*-fold cross-validation approach with $k = n$ (with n corresponding to the total number of samples).

Using this approach, we find the weighted classification error rate for the final random forest model to be 9.5%. See Table S1 for a detailed summary of the model error rate.

Table S1. Summary of calibration errors for all topaz in this study, determined via leave-one-out cross-validation.

Actual Provenance	**Predicted**		**Error rate (%)**
	Texas	**Other**	
Texas (n = 178)	156	22	12.4
Not from Texas (n = 735)	48	687	6.5

REFERENCES

Breiman L. (2001) Random forests. *Machine Learning*, Vol. 45, pp. 5–32.

Chen C., Liaw A., Breiman L. (2004) Using random forest to learn imbalanced data. University of California, Berkeley.