

Application of machine learning on determination of gem provenance – Peridot as example

Y. ZHONG¹, A. H. SHEN^{1*}, Y. HAN²

¹ Gemmological Institute, China University of Geoscience, Wuhan 430074, China (*correspondence: ahshen1@live.com; zhongy1024@qq.com)

² Geological Museum of Zhangjiakou, Zhangjiakou 075000, China

Geographic origin determination is an important aspect in gem identification and evaluation. It is commonly done by geochemical methods such as trace element chemistry. Several ‘fingerprint’ elements are plotted and those from the same provenance tend to cluster. However, with more newly discovered gem deposits, these ‘fingerprint’ plots often showed overlap among different localities, making determination difficult.

Machine learning (ML) can be used to solve this problem. ML can process more variables than plotting. All chemical components can be used in ML to reduce the overlap among localities. Furthermore, with more data, ML model can improve accuracy. Among all ML algorithms, linear discriminant analysis (LDA) is most commonly used. However, LDA might not be the best model. When the boundaries among different categories of data points in high-dimensional space are not hyperplanes, but curve surfaces, a non-linear classification algorithm would perform better. For data with different distribution features and boundary shapes, choosing a suitable ML algorithm can improve the accuracy of discrimination.

In this study, we used the LA-ICP-MS data of 262 peridot samples (62 from Hebei, China; 100 from Jilin, China and 100 from Changyon District, N. Korea) and six ML algorithms - LDA, decision tree, random forest, k nearest neighbour, support vector machine (SVM) and logistic regression. The samples are randomly divided into a training set and a testing set according to a ratio of 7:3. Training set is used to build ML models, while testing set is used to simulate samples of unknown localities. Different models are established by using Python and packages of Python like Numpy and Scikit-learn. Any sample can be predicted by the models, and the result can be compared with its actual locality to verify models’ accuracy. The accuracy of LDA on training set is 0.825, while its score of testing set is 0.797. Other five algorithms perform better than LDA. Among these, SVM gains the highest scores on testing set, and shows the best generalization, which means the difference of accuracy is narrow between testing set and training set.