Goldschmidt 2022 Abstract
https://doi.org/10.46427/gold2022.11096

Unsupervised and Supervised machine learning methods to identify Geographical Origin using LA-ICP-MS analysis of emeralds

RAQUEL ALONSO-PEREZ[1], PROF. JAMES M.D. DAY[2], D. GRAHAM PEARSON[3], YAN LUO[3], MANUEL PALACIOS[4], SUDHAKAR RAJU SATYANARAYAN[5] AND AARON PALKE[6]

[1]Harvard University
[2]Scripps Institution of Oceanography
[3]University of Alberta
[4]Eigenvector Research Inc.
[5]Rockhurst University
[6]Gemological Institute of America

Presenting Author: araquel69@gmail.com

Gemstone sourcing has a significant impact on the economies and development of the countries of their origin. In the 1980s in Colombia, for example, significant conflict occurred between the emerald mining industry and drug cartels to ensure the entities remained separate. Despite “digital” efforts to trace gemstones from extraction to the consumer, such methods have critical vulnerabilities, especially at the point of origin. Alternative, robust sample-based methods for geographic origin determination are needed to compliment these approaches. Three factors make emeralds particularly amenable to geographic origin studies. First is the diverse element chemistry of emeralds in minor (ppm) to trace (sub-ppm) quantities, with diagnostic and unique inter-element fractionations between deposits. Second, is the ever-increasing ability to exploit the geochemistry of emeralds to provide a 'genetic stamp' by measuring these diagnostic trace-element signatures using minimally destructive techniques, such as laser ablation inductively coupled plasma mass spectrometry (LA-ICP-MS). Third, the application of machine learning methods to big data sets can create powerful statistical discrimination approaches and prediction models provided that deposits have been sufficiently well characterized. We utilize a curated dataset of >800 high-precision LA-ICP-MS analyses from gem-quality emeralds (including >500 new analyses from this study) from sixteen of the most productive world-wide deposits (from a total of ~50 known global emerald deposits) in combination with unsupervised (clustering and Principal components analysis [PCA]) and supervised machine learning methods (Partial Least Square Discriminant Analysis [PLSDA] and Logistic Regression [LR]). While cluster analysis visualizes two major genetic classifications, sedimentary versus magmatic emerald deposits, PCA analysis highlights eight elements as statistically important (Li, V, Cr, Fe, Sc, Ga, Rb, and Cs) among 39 studied trace elements. PLSDA algorithms are a robust method to differentiate between samples from two different deposits within the Mananjary region in Madagascar, while LR algorithms provide a less than 7.2% misclassification error for the entire dataset. Unsupervised and Supervised machine learning methods, coupled with high quality LA-ICP-MS data have unrivalled potential for accurate identification of geographical origin in emeralds.