A Re-evaluation of Paleosol Elemental Proxies for Climate through Cross-validation and Machine Learning

WILLIAM E. LUKENS[1], GARY E. STINCHCOMB[2], LEE C. NORDT[3], STEVEN G. DRIESE[4], DAVID J. KAHLE[5], JACK D. TUBBS[6]

[1]Geosciences Department, Baylor University, Waco, TX 76798, bill_lukens@baylor.edu
[2]Watershed Studies Institute & Department of Geosciences, Murray State University, Murray, KY 42071, gstinchcomb@murraystate.edu
[3]Geosciences Department, Baylor University, Waco, TX 76798, lee_nordt@baylor.edu
[4]Geosciences Department, Baylor University, Waco, TX 76798, steven_driese@baylor.edu
[5]Department of Statistical Science, Baylor University, Waco, TX 76798, david_kahle@baylor.edu
[6] Department of Statistical Science, Baylor University, Waco, TX 76798, jack_tubbs@baylor.edu

The bulk elemental composition of soil subsurface (B) horizons is influenced by environmental, biological, geological, and climatic factors. Because fossil soils (paleosols) are common in the geologic record, quantitative models that link climate to paleosol geochemistry are highly desirable in the paleoclimate community. Over the last ca. 15 years, numerous transfer functions for climate have been developed using B horizon bulk elemental composition. Error for these models is typically reported as the root mean square error (RMSE) associated with regression analysis, and is understood to be the variance imparted by non-climatic influences on soil formation. However, prediction error must be estimated through cross-validation—a method frequently overlooked in simple regression analysis.

Here we re-evaluate the chemical index of alteration minus potassium (CIA-K) proxy for mean annual precipitation, perhaps the most widely applied pedotransfer function for paleosols, by performing cross-validation on two continental-scale soil data sets. Cross-validation of exponential regression models and data set inter-comparisons indicate that previously reported RMSE values underrepresent realistic prediction errors. We introduce promising results from recursive partitioning regression and random forest machine learning, which offer a new direction for automated variable selection and error estimation. These new models offer an optimized balance between wide applicability and relatively low prediction error, and are the necessary next steps in paleosol proxy development.