# 112 Steps on a path: An application of machine learning using a random forest algorithm to predict visitor use levels on trails in Rocky Mountain National Park, USA.

**Evan Bredeweg**[1], **Ashley D'Antonio**[1], **Scott Esser**[2], **Andrea Jacobs**[1], [1]Oregon State University, USA. [2]National Park Service, USA

Understanding the location and level of recreation use in park and protected areas (PPA) can be useful for effective visitor use management. While there is a wealth of geospatial data available online and in the manager databases of many PPA, the development and format of these datasets may be shaped more by the nature of GIS software than the way visitors explore and use a PPA system. Moreover, aspects important for visitor management such as quantification of visitor use levels on trails may be more difficult to source for each trail segment than physical trail characteristics (length, location, elevation profile, etc.). It would be expected that trail characteristics would influence the traffic of visitors, but there are many other factors such as accessibility, parking, or nearby attractions that can influence visitor behavior in complex ways. While we can obtain the physical characteristics, available amenities, and relative locations of trails within the entire PPA, we often do not have visitor use levels on the same extent. In order to examine visitor use levels on the scale of the entire PPA, we need to be able to model the relationship between physical location, trail characteristics, and amenities that ultimately shape visitor use.

The goal of this study was to predict visitor use levels on less intensely monitored trails in Rocky Mountain National Parks, USA using a small subset of well-monitored trails and trail characteristics extracted from existing geospatial datasets. To develop this park-wide trails geospatial database with estimated levels of visitor use, we employed a random forest algorithm trained on a subset of trails within the park assigned with expert-derived use levels. The training dataset consisted of 53 different trails which consisted of 94 individual trail segments with had one of five usage levels (Low - 1, Moderate - 2, Fairly Heavy - 3, Heavy - 4, Very Heavy - 5). The geospatial dataset consisted of the trails database from the park (468 trail segments in total) which included associated geolocations and specific attributes such as park management unit, establishment dates, and classification for trail surface and development. In addition to the variables within this dataset, we computed a variety of variables for each trail segment that we predicted could shape visitor use levels. These included values such as the elevation profile, trail grade profile, park entrance proximity, distance from nearest trailhead, parking availability at nearest trailhead, and visitor-count/entrance weighted accessibility. We fit a classification random forest (RF) algorithm using a set of 24 covariates with our 1 categorized response variable of trail visitor-use. We fit the RF model using 1000 trees (ntree) and optimized the number of candidate variables per split (mtry) by exploring possible values from 1-24. This model had an out-of-bag error rate of 25.52% for the training dataset. The out-of-bag error rate is used in Random Forest algorithms to estimate prediction error in a form of 'leave-one-out' cross-validation. However, with trail usage being an ordered categorization, we also wanted to assess the errors of misclassification that were greater than one classification level. These gross misclassifications of trail visitor-use only occurred on 8.5% of the trail segments in the training dataset. Using the optimized Random Forest model, we predicted the trail visitor use levels for the remaining 374 segments with unknown levels.

Importantly, the covariates used in this model are often values that can be derived from publicly accessible geodata sourced from OpenStreetMap or other online geodatabases. Even the geospatial data layer for hiking trails can be accessed through such sources. This means that an approach such as this one can be used in small PPA areas with limited in-house geospatial resources or across multiple PPA jurisdictions. To improve model performance, this approach can include additional covariates that capture more detailed aspects of trail attractions such as water access, wildflower viewing, wildlife attractions, etc.

A variety of additional avenues of research and monitoring are possible with an informed estimate of visitor use levels on trails. This kind of modeling effort can inform the best locations for more quantitative measurements of trail use (i.e., where to install trail counters) and aid in a more systematic measurement of visitor behavior. It can highlight areas of differential use for future surveys of visitor perception of crowds while hiking. Poor model performance can highlight the need for additional monitoring of trails or the possible departures of visitor behavior from trail-centric recreation decisions. Variation in the visitor use of trails can inform the further analysis of social trail formation or negative impacts on wildlife species.

While this model is not a perfect representation of human behavior within the Rocky Mountain National Park trail network, it provides an informed estimate of visitor use levels on unmonitored or minimally monitored trails by leveraging a subset of known trails. Having a geospatial network of informed visitor use levels at the scale of an entire PPA can provide information for new research projects and the context to support PPA managers in their management and conservation goals.