12-10-2020

Autocart: Spatially-Aware Regression Trees for Ecological and Spatial Modeling

Ethan Ancell
Utah State University, ethan.ancell@aggiemail.usu.edu

Recommended Citation

Ancell, Ethan, "Autocart: Spatially-Aware Regression Trees for Ecological and Spatial Modeling" (2020). *Fall Student Research Symposium 2020*. 1.
https://digitalcommons.usu.edu/fsrs2020/1

Autocart: spatially-aware regression trees for ecological and spatial modeling

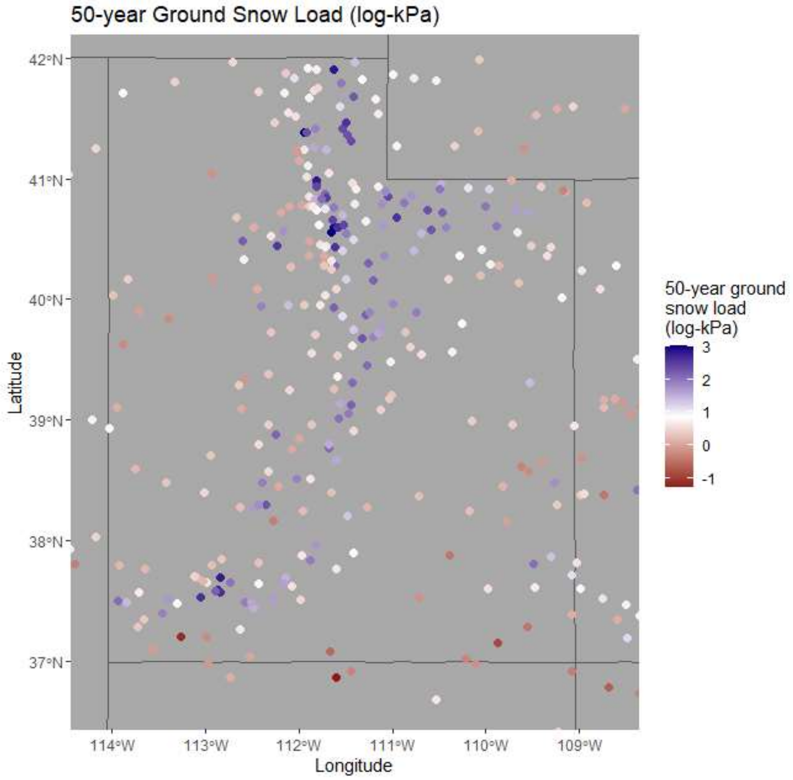
Ethan Ancell

Utah State University

Introduction and motivation

We want a model for predictions of ground snow load at any given longitude and latitude in Utah. How do we do this?

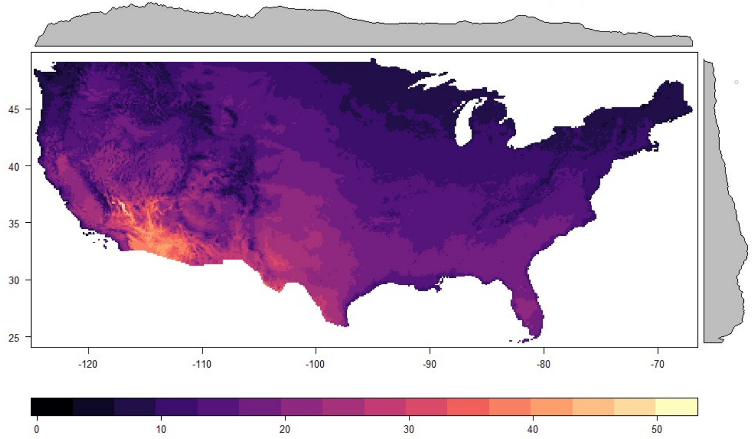
1. Gather snow data at a variety of locations in Utah (shown in the figure on the right) – we call this the labelled training data.
2. Gather gridded climate information across the state of Utah.
3. Use the gridded climate information and the training data to "learn" what climate factors contribute to different levels of ground snow load.



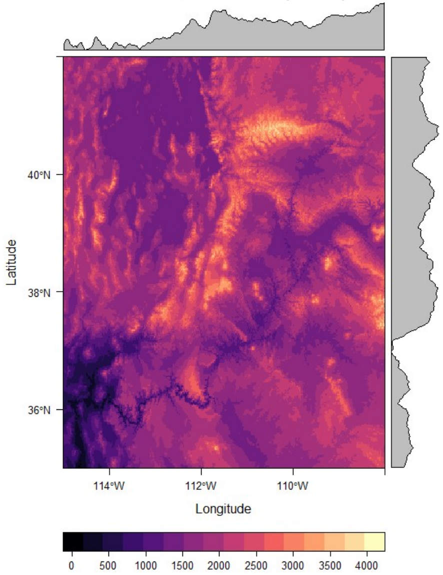
Predictor variables

We have various gridded climate predictor variables to help us out:

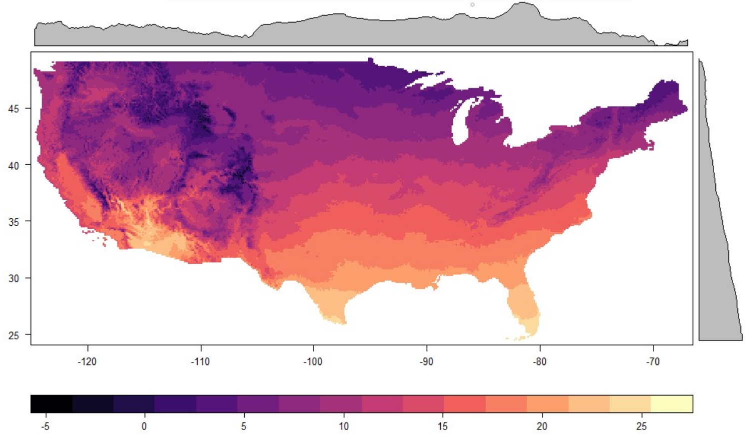
Max vapor pressure deficit in the United States (hPa)



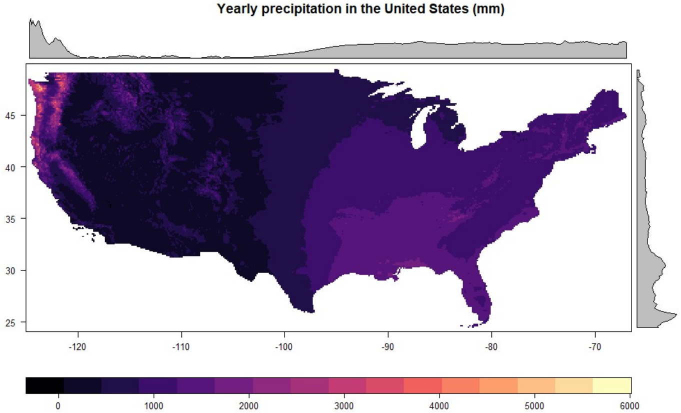
Elevation in Utah (meters)



Yearly mean temperature in the United States (degrees Fahrenheit)



Yearly precipitation in the United States (mm)



What model to use?

- Linear statistical model such as linear regression

  Pros: Interpretable and easy to implement

  Cons: Climate information is probably too complex for this data

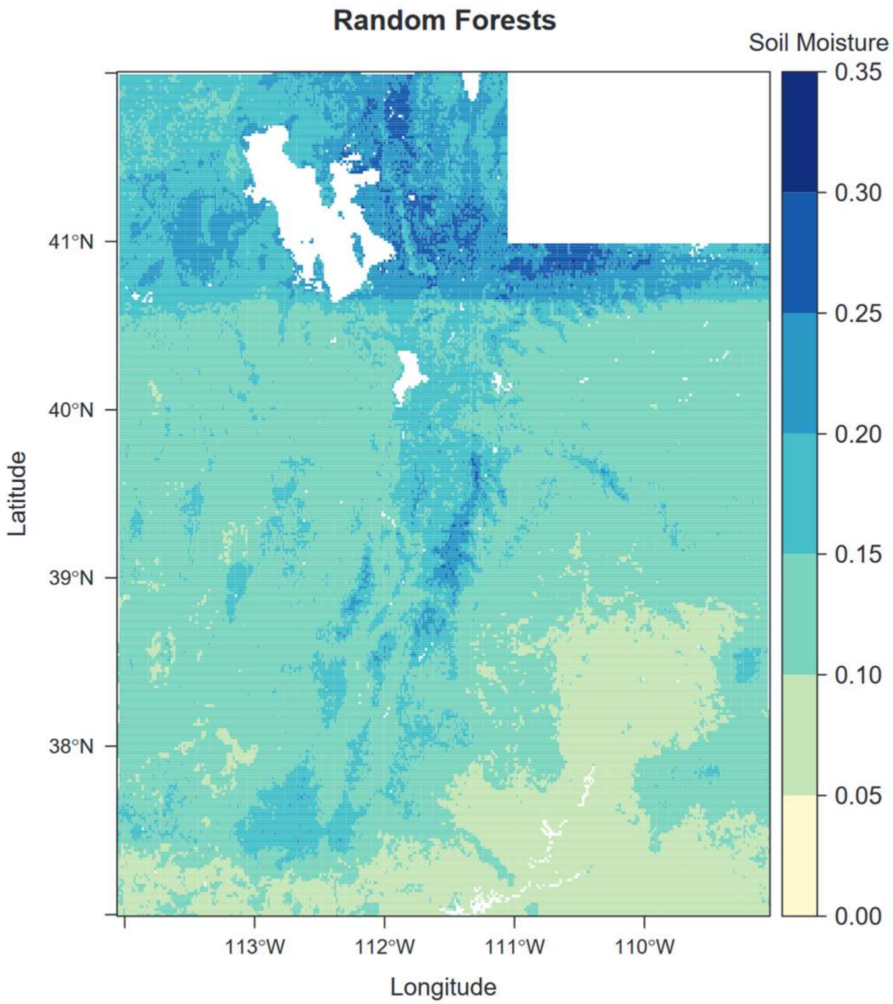
- A machine-learning model

  Pros: Ability to learn complex patterns, particularly in a complex climate setting

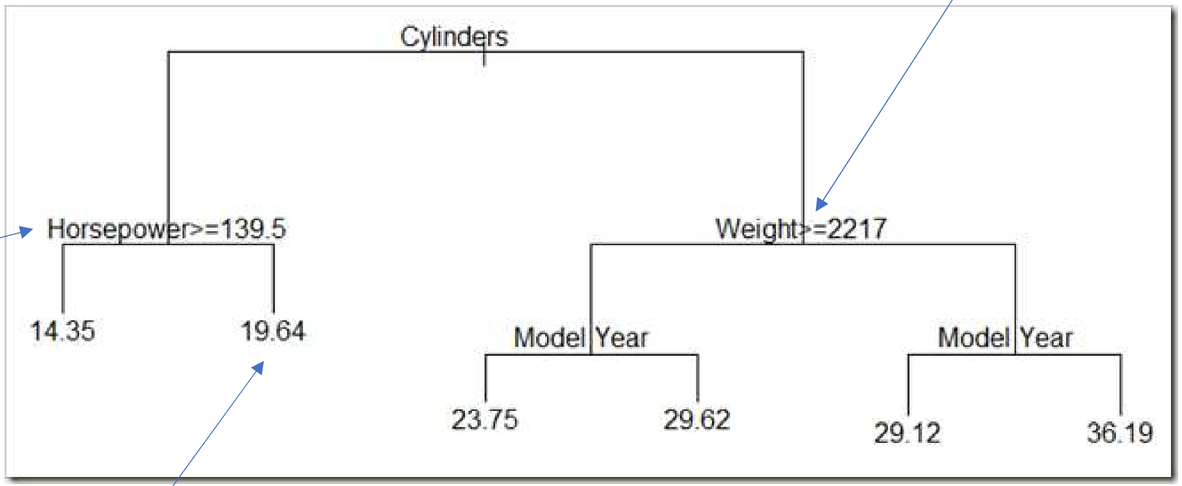
  Cons: Machine-learning models do not properly account for unique qualities of spatial data: primarily the spatial dependence issue.

Solution: Build a variant of an existing machine-learning model so that it properly accounts for spatial data. In this project, I modify the regression tree algorithm.

Example of poor result from machine-learning model



Regression trees



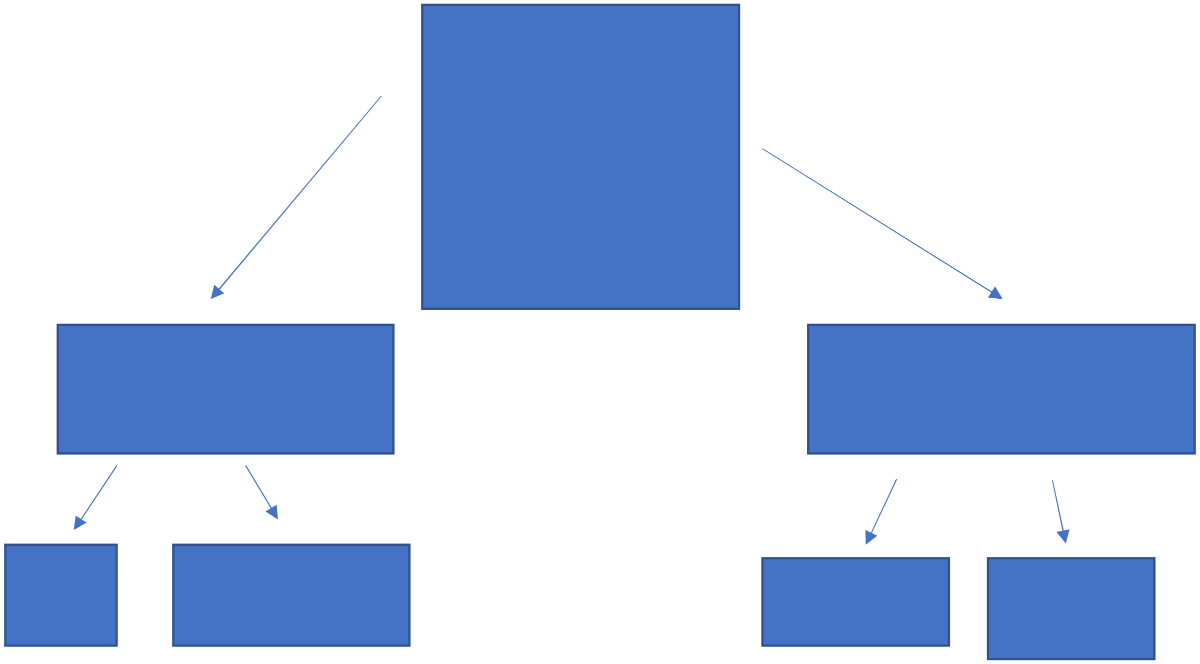
A node

An example of a splitting rule. We would move right if we had a car with a horsepower over 139.5 and left otherwise.

A terminal node

How are regression trees formed?

1. Gather a collection of data labeled with a known response variable (e.g., our training snow data points)
2. Evaluate every single possible splitting rule on every climate variable with an objective function called g_{rss}. Tells you how good the splits are (clustering similar with similar)
3. With the best split in hand, start at step 1 again with the newly partitioned data.



Overview of the autocart model

1. Train a regression tree with a modified objective function

$$g = (1 - \alpha - \beta) g_{rss} + \alpha g_{ac} + \beta g_{sc}$$

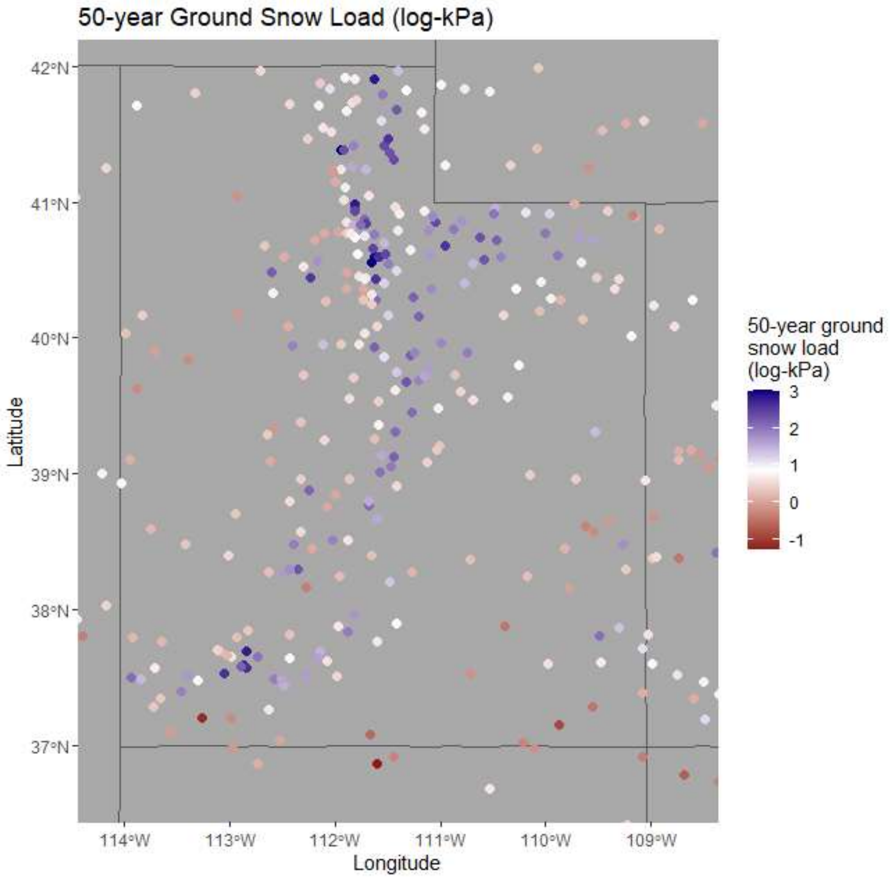
$$\alpha, \beta \in [0,1] \text{ and } \alpha + \beta \leq 1$$

2. Make a prediction and supplement it with an interpolation of the residuals of the training observations that reside in the terminal node of the tree where the prediction occurs.

$$\hat{Y} = \bar{y}_T + u(\boldsymbol{s})$$

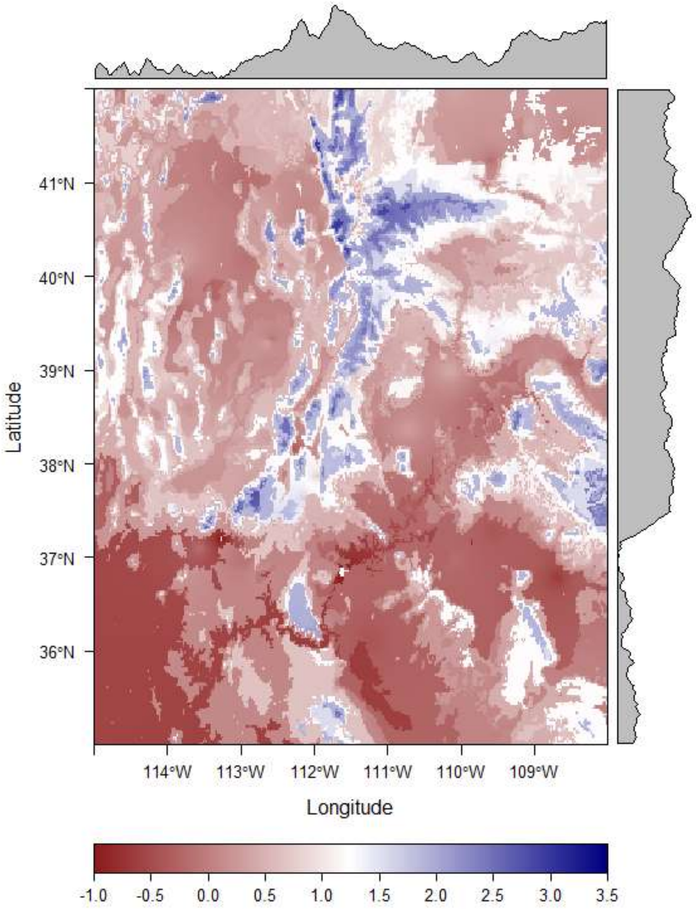
$\bar{y}_T$ is the original prediction made with the average response variable in the terminal node, and $u(\boldsymbol{s})$ is the interpolated residual.

Examples of mapping results – 50-year Utah ground snow load (prediction of log-kPa)



Original raw data

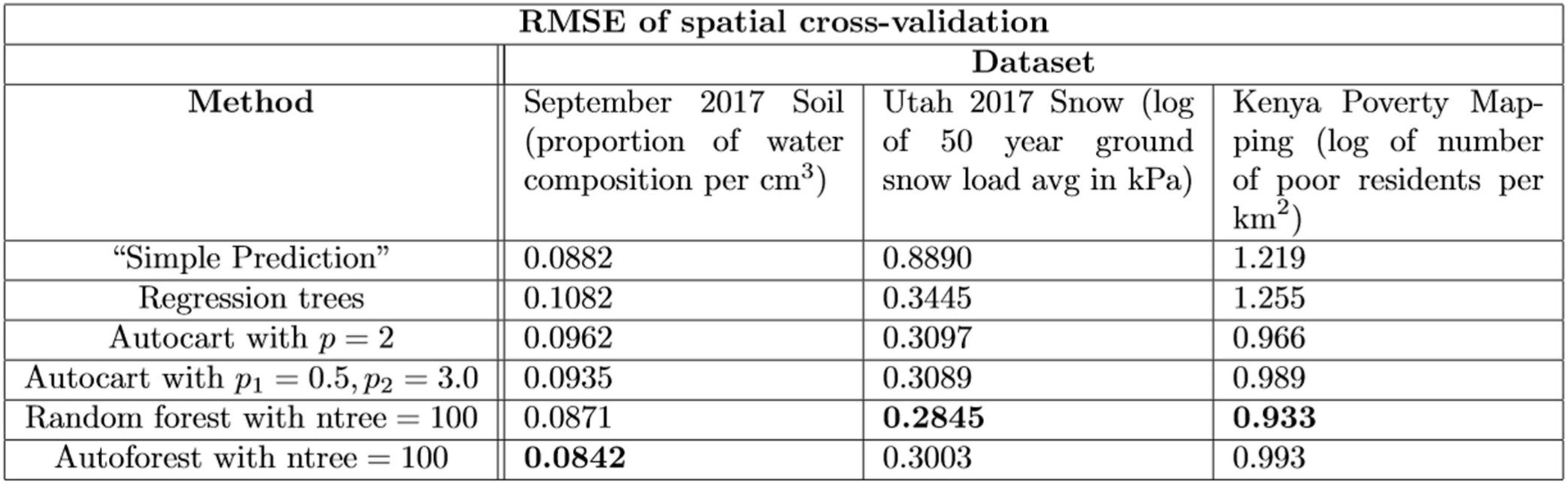
Autocart +
Climate data



Predicted 50-year ground snow load (log-kPa)

Some cross-validated results

Autocart/autoforest doesn't just produce nicer maps– there is some evidence it produces more accurate predictions in terms of cross-validated RMSE on three tested datasets.



RMSE of spatial cross-validation			
	Dataset		
Method	September 2017 Soil (proportion of water composition per cm^3)	Utah 2017 Snow (log of 50 year ground snow load avg in kPa)	Kenya Poverty Mapping (log of number of poor residents per km^2)
"Simple Prediction"	0.0882	0.8890	1.219
Regression trees	0.1082	0.3445	1.255
Autocart with $p = 2$	0.0962	0.3097	0.966
Autocart with $p_1 = 0.5, p_2 = 3.0$	0.0935	0.3089	0.989
Random forest with ntree = 100	0.0871	**0.2845**	**0.933**
Autoforest with ntree = 100	**0.0842**	0.3003	0.993

(Both autocart and regression trees are pruned to the same level)

Developed software

Available as an R package at www.github.com/ethanancell/autocart.

The package is primarily written in C++ for fast computations. It is also parallelized so you can take advantage of multiple cores on your processor.

Also contains “autoforest”, a Random Forest extension to the autocart model.

Future research directions

1) Exploring more ensemble methods of autocart. A random forest extension called “autoforest” was explored and implemented, however the gains in predictive accuracy over traditional random forests are very minor compared to the gain in predictive accuracy that autocart has over regression trees. Is this also true for other ensemble methods such as boosted trees?

2) An automatic selection of the power parameter in the interpolation step. Although slide 6 mentions a way autocart can pick the power parameter from a range of values using Moran’s I, it would be helpful to have a process to pick $[p_1, p_2]$ automatically.

3) An interaction function for the interaction between g_{ac} and g_{sc}. Does the effect these objective functions have on the power of autocart depend on each other?

4) A formal evaluation of smoothness over the region. The results in slides 7/8 are only visually assessed. Is there study on formally defining a measure of smoothness?

Contact Information

ethan.ancell@aggiemail.usu.edu

References

H. Meyer, C. Reudenbach, S. Wollauer, and T. Nauss, “Importance of spatial predictor variable selection in machine learning applications–moving from data reproduction to spatial prediction,” *Ecological Modelling*, vol. 411, p. 108815, 2019.

D. Stojanova, M. Ceci, A. Appice, D. Malerba, and S. Dzeroski, “Dealing with spatial autocorrelation when learning predictive clustering trees,” *Ecological Informatics*, vol. 13, pp. 22–39, 2013.