Calculating Representativeness of Geographic Sites Across the World

Ashwinkumar Ganesan, Tim Oates, Matt Schmill & Erle Ellis

cognition | robotics | learning
CORAL

UMBC
AN HONORS UNIVERSITY IN MARYLAND

Motivation

- Studies of *land change* are conducted across the world
- The number of studies conducted is limited due to expense
- Global information about regions includes tree cover, temperature, precipitation, etc
- Our goal is to determine the extent to which study results can be generalized geographically

Problem Statement

- GLOBE is a global collaboration engine, a project to study global effects of Land Change.
- To find how a study or a set of studies of specific geographic areas can be generalized to other areas of the world based on a set of parameters
- Suggest regions where a study can be conducted

Defining Representativeness

Given a distribution where

- D is a data (e.g., temperature and tree cover at locations around the globe)
- *S* is a sample such that $S \subseteq D$ (e.g. locations in Baltimore)
- *H* is a histogram based on *D*
- *Bin(H, x)* is the bin where data value x falls in *H*
- *P(H, i)* is the height / probability of bin *i* in histogram *H*
- All unique bins are defined in a set

$B = \{b | \forall_{x \in Sb} = bin\ (H, x)\}$

Representativeness R is

$$R(S|D) = \sum_{b \in B} p(H, b)$$

Objectives

- Calculate representativeness of given Sample locations/ regions in real time
- Perform the computation for a total of 1.6 million regions and render on Google Maps

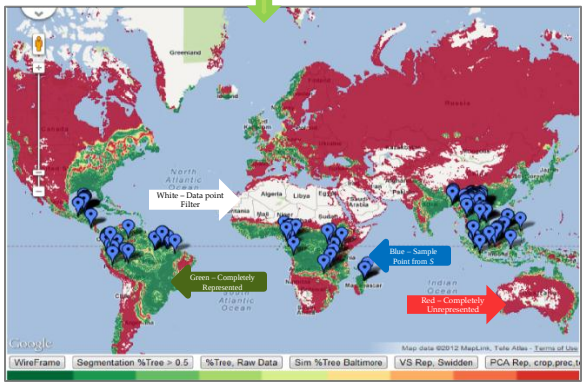
Challenges

- Standard K-Means clustering is slow for large data set provided
- Multivariate Dataset with over 50 dimensions

Measuring R for User Defined Samples

- User selects a set of regions & the variable set to be analyzed
- Perform dimensionality reduction using principal components analysis (PCA)
- Let single dimension PCA projection be D_p
- Final Distance between data point *x* and sample *S*

$$FD(x) = min(Dp(x) - Dp(s)_{s \in S})$$

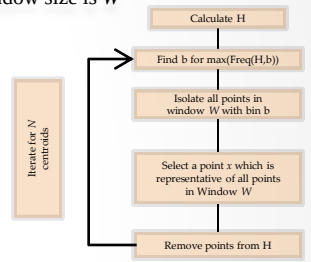


Future Work

- Compare the method against *Monte Carlo* sampling
- Measure accuracy of centroids selected by defining region preservation and specificity
- Applying spatial autocorrelation methods
- Consider a weighted variable set where region sizes differ

Selecting "Ideal" Rep. Points

- User would want a set of regions to be suggested for given variable set based on equal probability or equal area distribution
- *H* is the histogram of D_p
- *Freq(H,b)* is frequency of bin b
- Window size is *W*



- Calculate *R*

$$R(S|D) = \frac{\sum_{x \in C} \sum_{b \in W} Freq(H,b)}{D}$$

Where C is the set of centroids

Measure Effectiveness of Method

- Consider a random Dataset *D* of 50,000 points
- Check the cluster created against distance of variables (with the origin) in original data space.