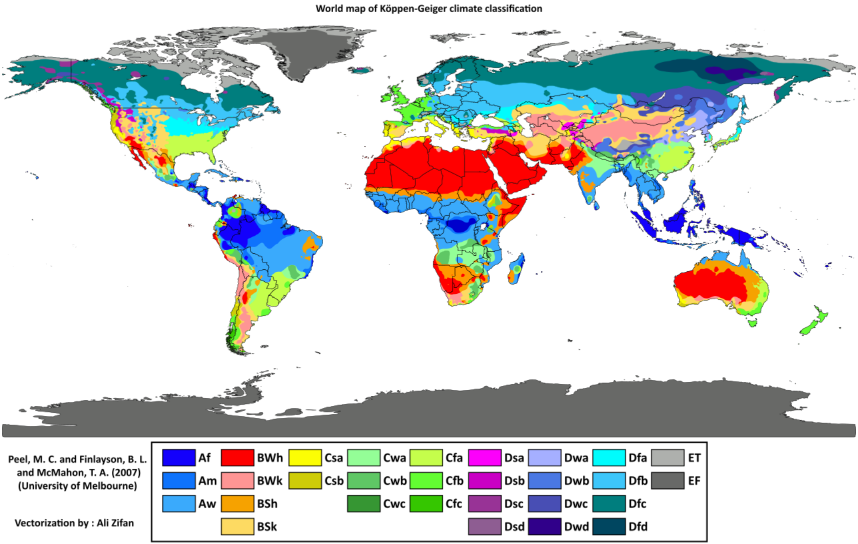
A wavelet based approach to climate biome clustering

Derek Desantis

University of Nebraska - Lincoln

August 7, 2018

Type	Description	Criterion
A	**Equatorial climates**	$T_{min} \geq +18\ °C$
Af	Equatorial rainforest, fully humid	$P_{min} \geq 60$ mm
Am	Equatorial monsoon	$P_{ann} \geq 25\,(100 - P_{min})$
As	Equatorial savannah with dry summer	$P_{min} < 60$ mm in summer
Aw	Equatorial savannah with dry winter	$P_{min} < 60$ mm in winter
B	**Arid climates**	$P_{ann} < 10\,P_{th}$
BS	Steppe climate	$P_{ann} > 5\,P_{th}$
BW	Desert climate	$P_{ann} \leq 5\,P_{th}$
C	**Warm temperate climates**	$-3\ °C < T_{min} < +18\ °C$
Cs	Warm temperate climate with dry summer	$P_{smin} < P_{wmin}$, $P_{wmax} > 3\,P_{smin}$ and $P_{smin} < 40$ mm
Cw	Warm temperate climate with dry winter	$P_{wmin} < P_{smin}$ and $P_{smax} > 10\,P_{wmin}$
Cf	Warm temperate climate, fully humid	neither Cs nor Cw
D	**Snow climates**	$T_{min} \leq -3\ °C$
Ds	Snow climate with dry summer	$P_{smin} < P_{wmin}$, $P_{wmax} > 3\,P_{smin}$ and $P_{smin} < 40$ mm
Dw	Snow climate with dry winter	$P_{wmin} < P_{smin}$ and $P_{smax} > 10\,P_{wmin}$
Df	Snow climate, fully humid	neither Ds nor Dw
E	**Polar climates**	$T_{max} < +10\ °C$
ET	Tundra climate	$0\ °C \leq T_{max} < +10\ °C$
EF	Frost climate	$T_{max} < 0\ °C$


World map of Köppen-Geiger climate classification
Peel, M. C. and Finlayson, B. L. and McMahon, T. A. (2007) (University of Melbourne)
Vectorization by : Ali Zifan
Af
Am
Aw
BWh
BWk
BSh
BSk
Csa
Csb
Cwa
Cwb
Cwc
Cfa
Cfb
Cfc
Dsa
Dsb
Dsc
Dsd
Dwa
Dwb
Dwc
Dwd
Dfa
Dfb
Dfc
Dfd
ET
EF

Problem

Problem

- *Only applies to land data*

Problem

- *Only applies to land data*
- *Only uses precipitation and temperature data*

Problem

- *Only applies to land data*
- *Only uses precipitation and temperature data*
- *As the climate changes, so must the biomes*

Problem

- *Only applies to land data*
- *Only uses precipitation and temperature data*
- *As the climate changes, so must the biomes*
- *Köppen-Gieger is just a heuristic*

Problem

- *Only applies to land data*
- *Only uses precipitation and temperature data*
- *As the climate changes, so must the biomes*
- *Köppen-Gieger is just a heuristic*

Goal

- *Cluster on any chosen variables*

Problem

- *Only applies to land data*
- *Only uses precipitation and temperature data*
- *As the climate changes, so must the biomes*
- *Köppen-Gieger is just a heuristic*

Goal

- *Cluster on any chosen variables*
- *Detect where biomes are shifting*

Problem

- *Only applies to land data*
- *Only uses precipitation and temperature data*
- *As the climate changes, so must the biomes*
- *Köppen-Gieger is just a heuristic*

Goal

- *Cluster on any chosen variables*
- *Detect where biomes are shifting*
- *Want a data driven model*

Supervised Learning: Discover salient features of data to separate into predetermined classes - Data comes with labels.

Supervised Learning: Discover salient features of data to separate into predetermined classes - Data comes with labels.

Example

Given an image of a leaf, determine which tree (from a predetermined list) it came from.

Supervised Learning: Discover salient features of data to separate into predetermined classes - Data comes with labels.

Example

Given an image of a leaf, determine which tree (from a predetermined list) it came from.

Unsupervised Learning: Discover classes hidden in the data - Data does **not** come with labels.

Supervised Learning: Discover salient features of data to separate into predetermined classes - Data comes with labels.

Example

Given an image of a leaf, determine which tree (from a predetermined list) it came from.

Unsupervised Learning: Discover classes hidden in the data - Data does **not** come with labels.

Example

Given images of leaves, automatically sort images into bins based of features (not set or necessarily known).

Remark

- *Determining biomes directly from data is unsupervised*

Remark

- *Determining biomes directly from data is unsupervised*
- *Generically speaking, supervised is "easier" than unsupervised*

Remark

- *Determining biomes directly from data is unsupervised*
- *Generically speaking, supervised is "easier" than unsupervised*
- *Large scale unsupervised learning is notoriously difficult (AKA prohibitively expensive):*

$$K\text{-}means \sim \mathcal{O}(K * number\ data * dim)$$

Example

- *Let* $f = [1, 1, 2, 2, .5, 0, 0, 0, 3, 1]$.

Example

- *Let* $f = [1, 1, 2, 2, .5, 0, 0, 0, 3, 1]$.
- *Taking the wavelet transform yields two new signals:*
  1. *Approximation Information - Averages of pairs of points*
  2. *Detail Information - Differences from averages*

Example

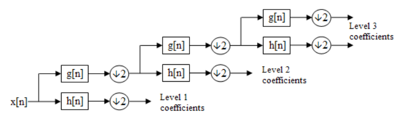
- *Let* $f = [1, 1, 2, 2, .5, 0, 0, 0, 3, 1]$.
- *Taking the wavelet transform yields two new signals:*
  1. *Approximation Information - Averages of pairs of points*
  2. *Detail Information - Differences from averages*
- *Taking DWT:*

$$
\begin{aligned}
g_1 &= [1, 2, 0.25, 0, 2] \\
h_1 &= [0, 0, 0.25, 0, 1]
\end{aligned}
$$

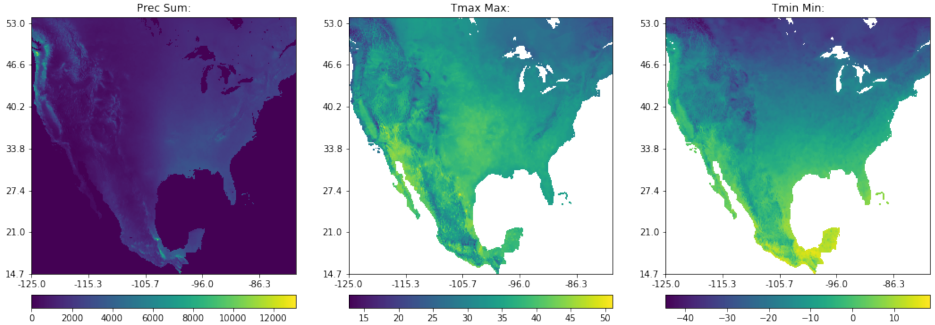
Example

- *Let* $f = [1, 1, 2, 2, .5, 0, 0, 0, 3, 1]$.
- *Taking the wavelet transform yields two new signals:*
  1. *Approximation Information - Averages of pairs of points*
  2. *Detail Information - Differences from averages*
- *Taking DWT:*

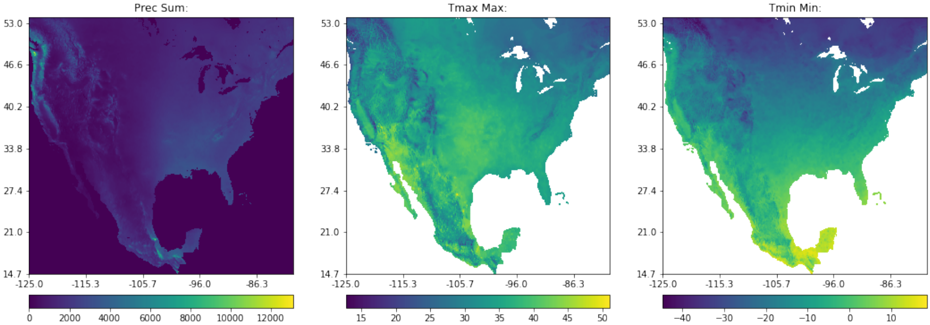
$$\begin{aligned} g_1 &= [1, 2, 0.25, 0, 2] \\ h_1 &= [0, 0, 0.25, 0, 1] \end{aligned}$$



LOCA Data: 1950-1970



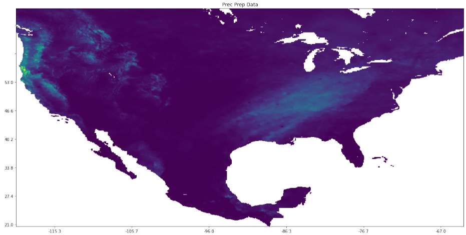
LOCA Data: 1950-1970



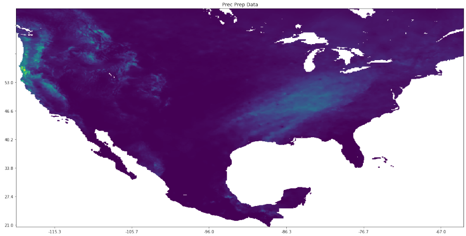
Choose wavelets:

- Space: Haar
- Time: db2

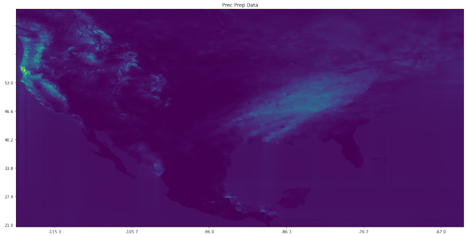
Prec Data: t=0


Prec Prep Data
53.0
46.6
40.2
33.8
27.4
21.0
-115.3
-105.7
-96.0
-86.3
-76.7
-67.0

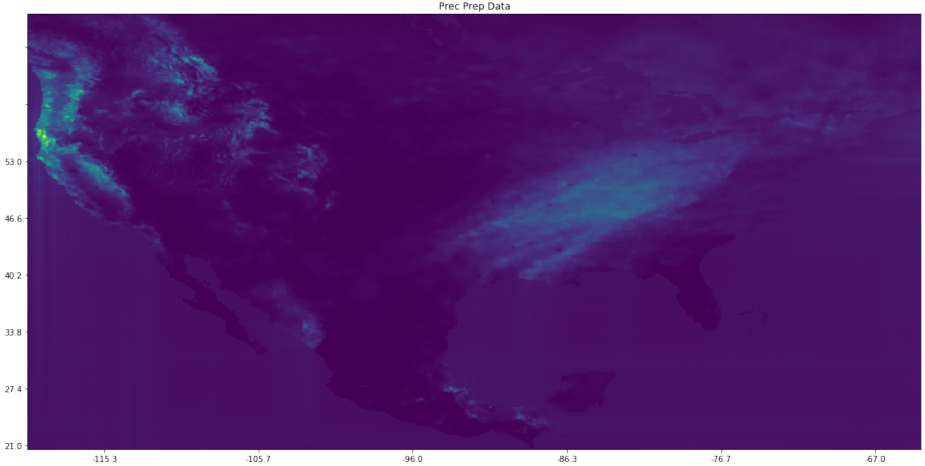
Prec Data: t=0



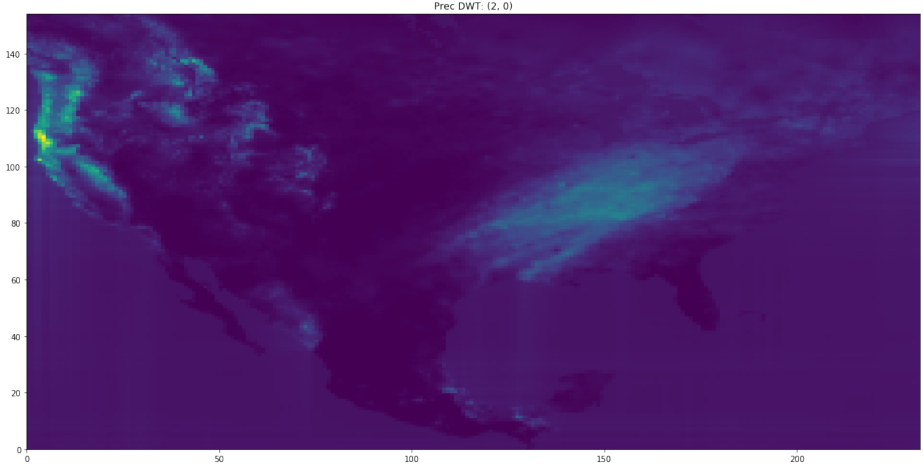
Interpolate Nan:



Interpolate Nan:



DWT: 2 space, 0 time

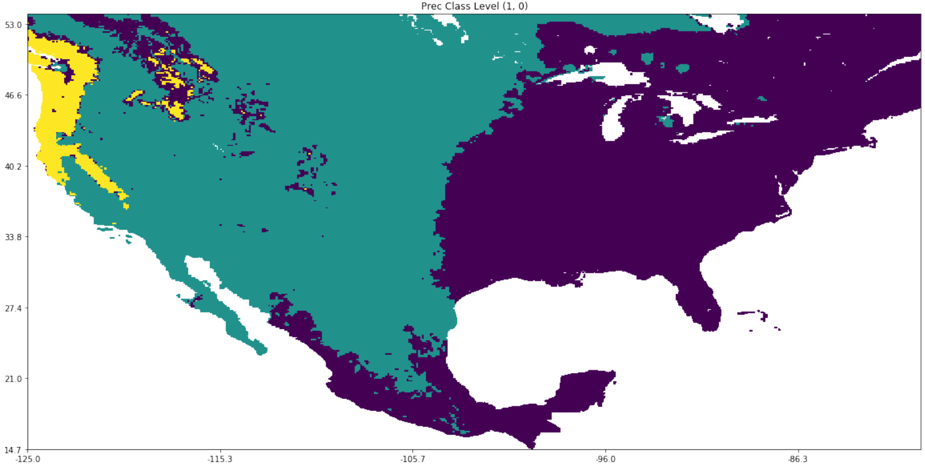


- Locate data values corresponding to non-NAN values (with ϵ boundary)

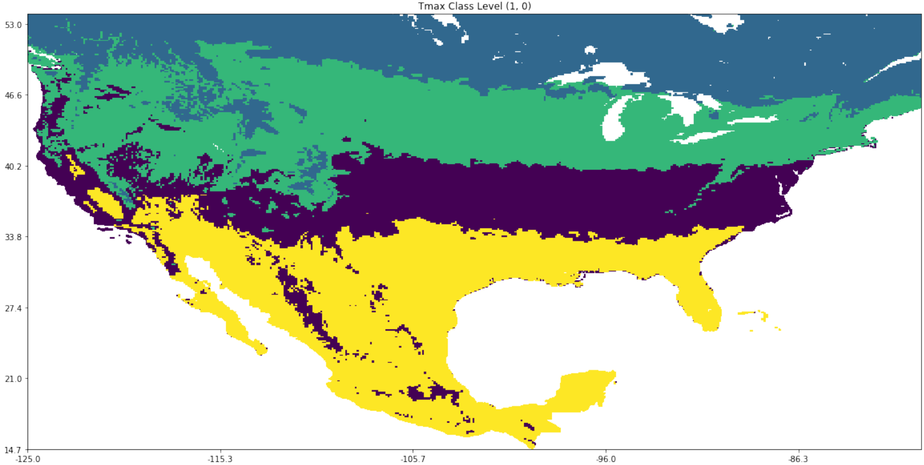
- Locate data values corresponding to non-NAN values (with ϵ boundary)
- Cluster the approximation coefficients for each variable

- Locate data values corresponding to non-NAN values (with ϵ boundary)
- Cluster the approximation coefficients for each variable
    - Settled on K-means
    - Determined number of clusters using silhouette and Calinski Harabaz scores
    - Used 3 clusters for Prec, 4 clusters for Tmin and Tmax

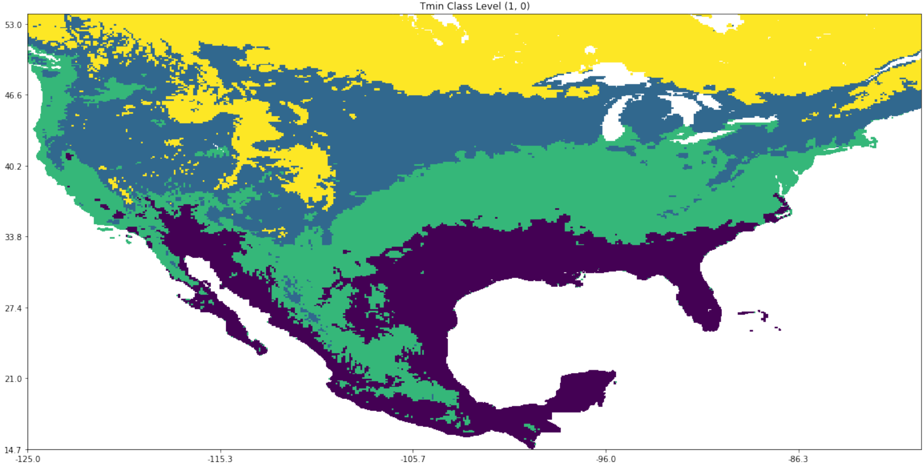
Data Clusters (1,0): 1950-1970



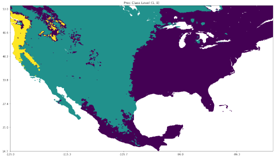
Data Clusters (1,0): 1950-1970

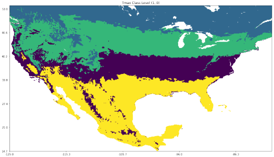


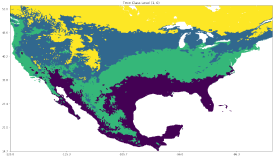
Data Clusters (1,0): 1950-1970



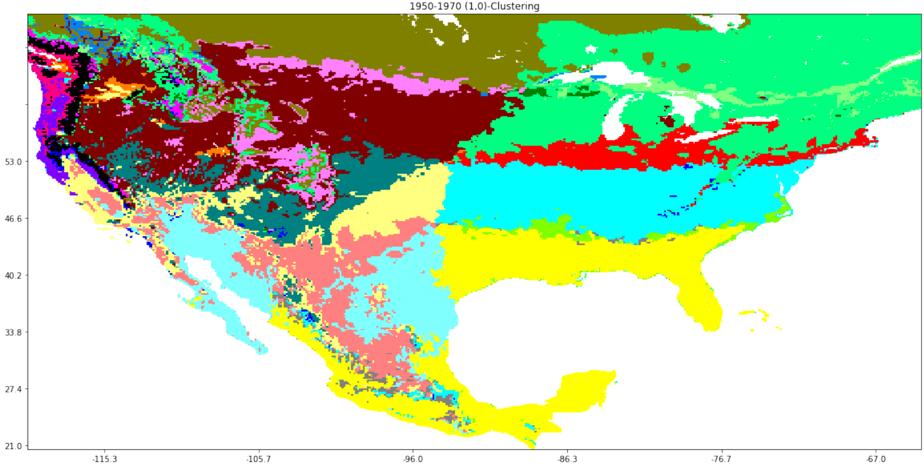
Data Clusters (1,0): 1950-1970



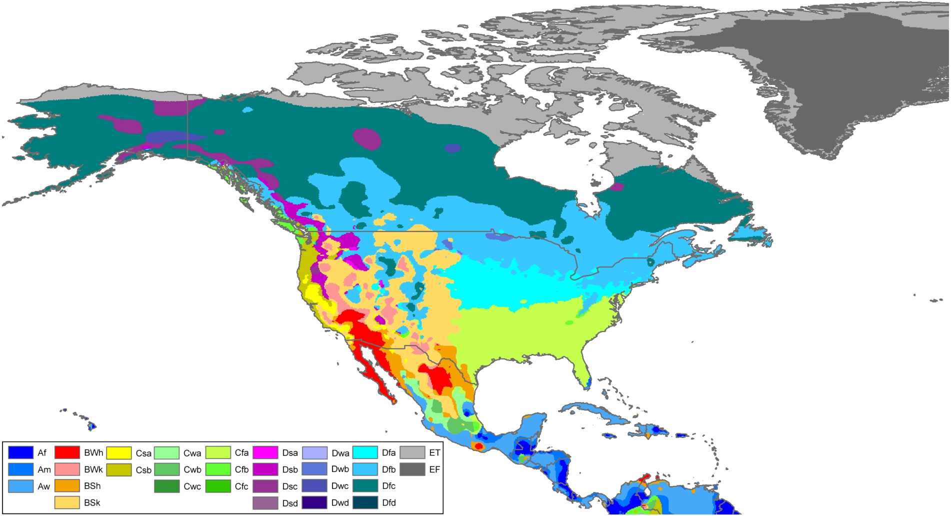




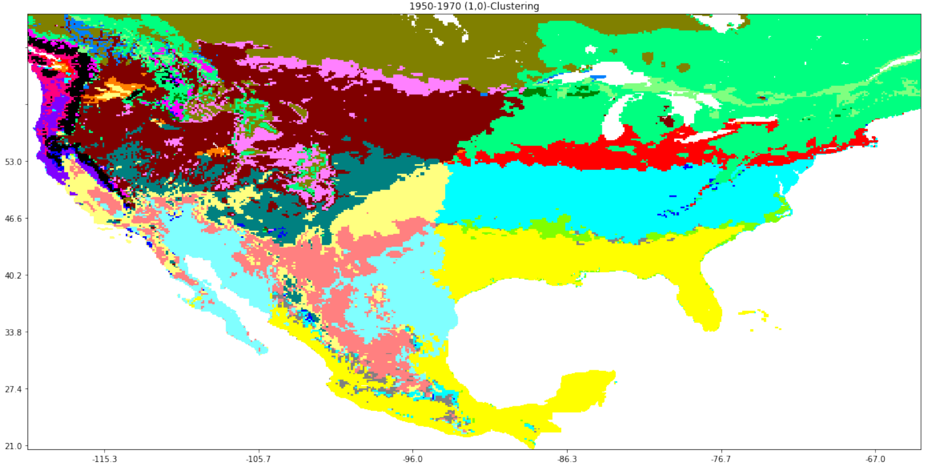
Combined Data Clusters (1,0): 1950-1970



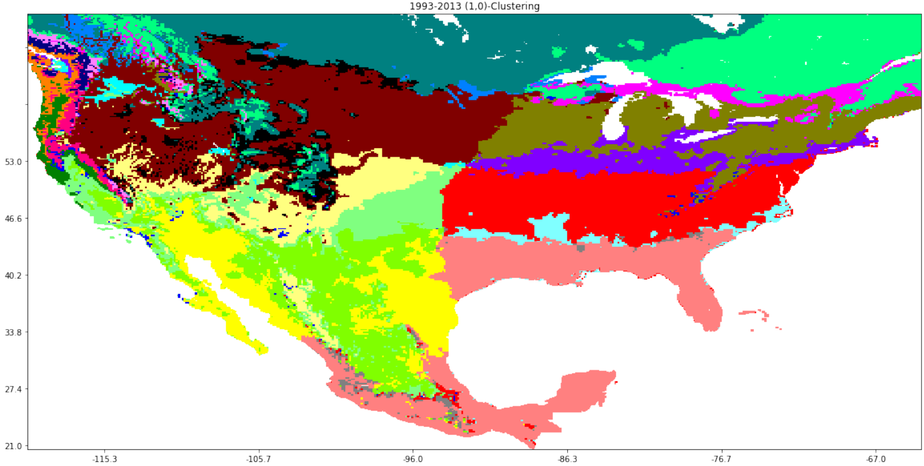
North America Köppen-Gieger Model



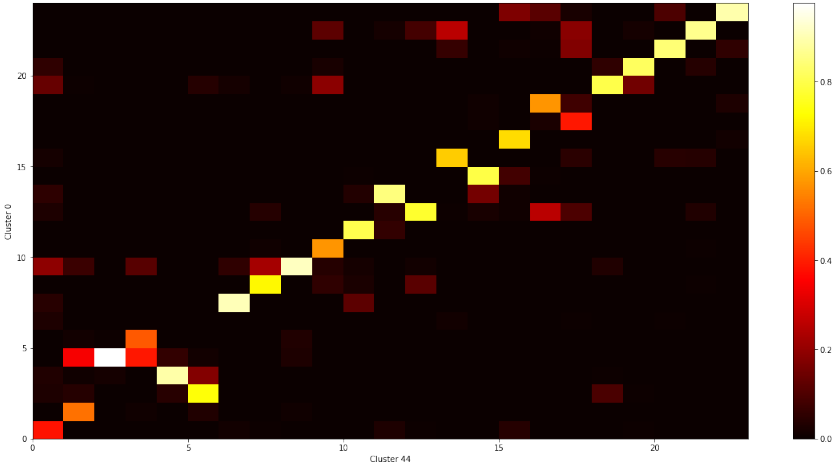
Combined Data Clusters (1,0): 1950-1970



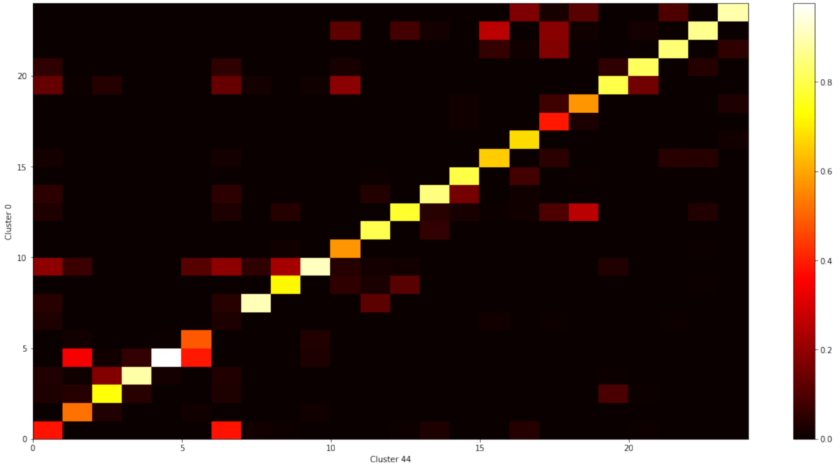
Combined Data Clusters (1,0): 1993-2013



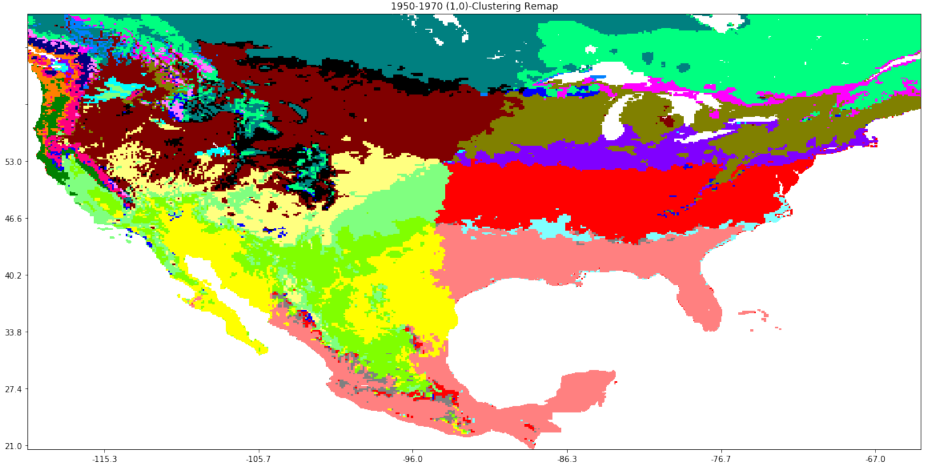
Correlation Between 1950-1970 Clusters and 1993-2013 Clusters



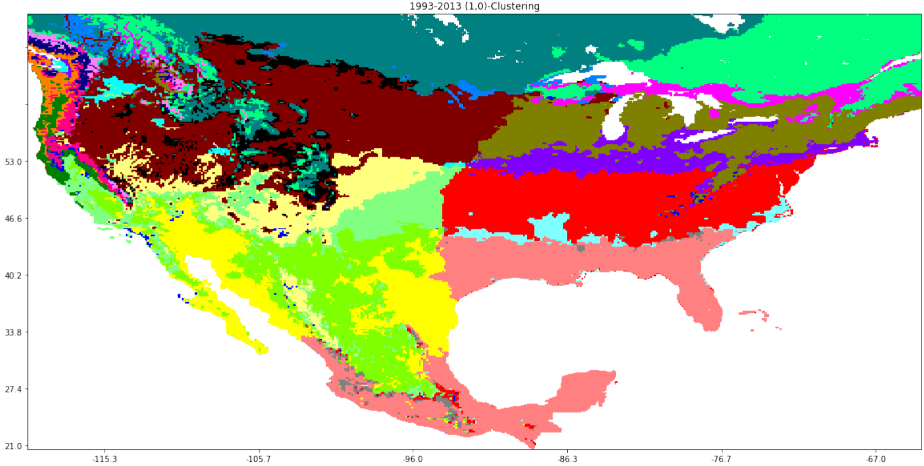
Sorted Correlation Between 1950-1970 Clusters and 1993-2013 Clusters



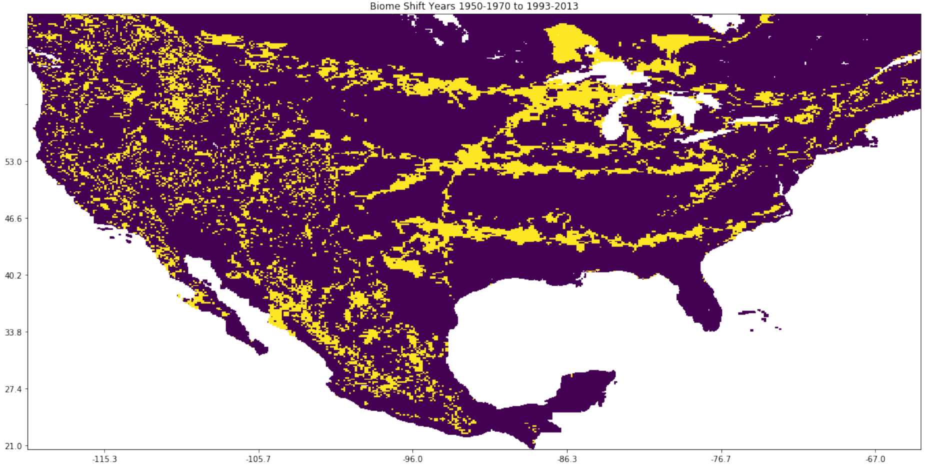
Reindex Combined Data Clusters (1,0): 1950-1970



Combined Data Clusters (1,0): 1993-2013



Difference Between 1950-1970 Clusters and 1993-2013 Clusters



1. Add other variables (e.g. wind)

1. Add other variables (e.g. wind)
2. Parallelism and optimization

1. Add other variables (e.g. wind)
2. Parallelism and optimization
3. Perform an analytical comparison to the Köppen-Gieger Model

1. Add other variables (e.g. wind)
2. Parallelism and optimization
3. Perform an analytical comparison to the Köppen-Gieger Model
4. Apply this clustering method to the ocean data