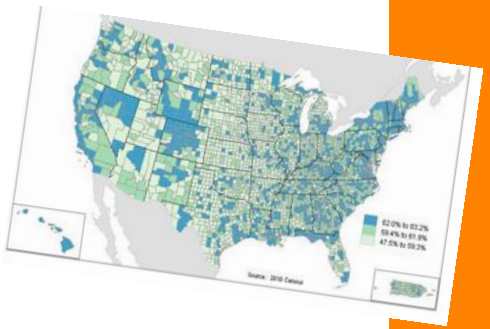
Big Data and the Old Scale Issues

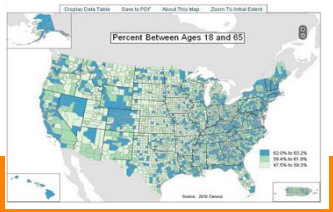
By Liem Tran, Ph.D.
Department of Geography



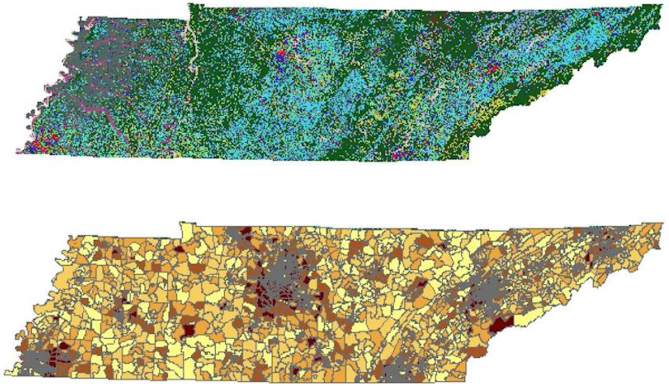
THE UNIVERSITY OF TENNESSEE KNOXVILLE

Roadmap

- Setting: issues & reasons of the study
- Scale & spatial heterogeneity
- Scale in spatial data mining
- Scale-based spatial clustering
- Discussion



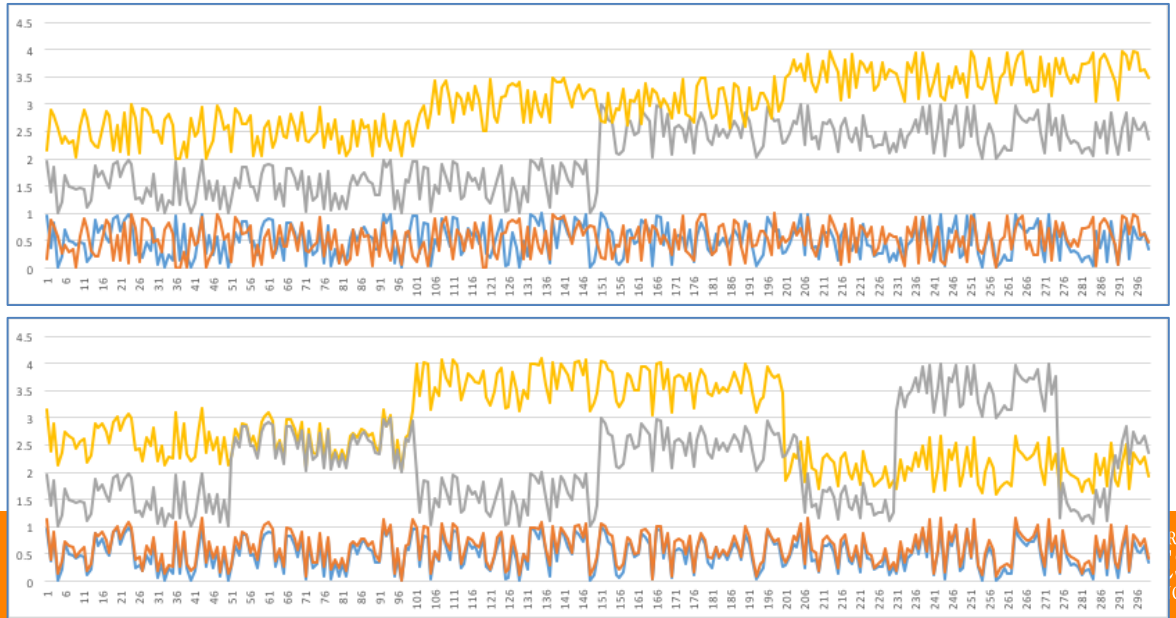
Big Data



- Large size,
- Incongruency,
- Incompleteness,
- Complexity,
- **Multiplicity of scales**, and
- Heterogeneity of information-generating sources.

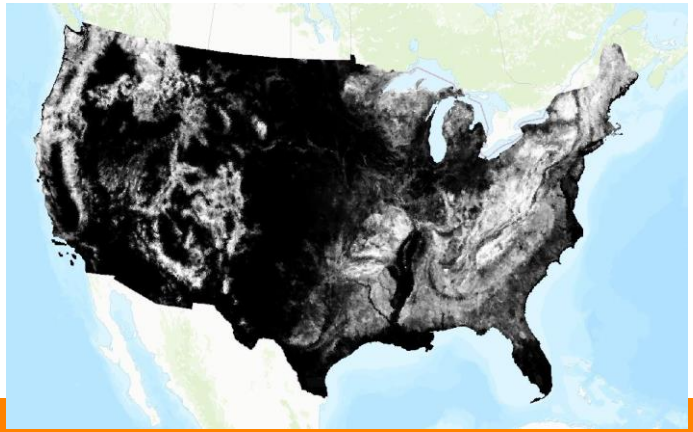
See “Scale” via Spatial Heterogeneity

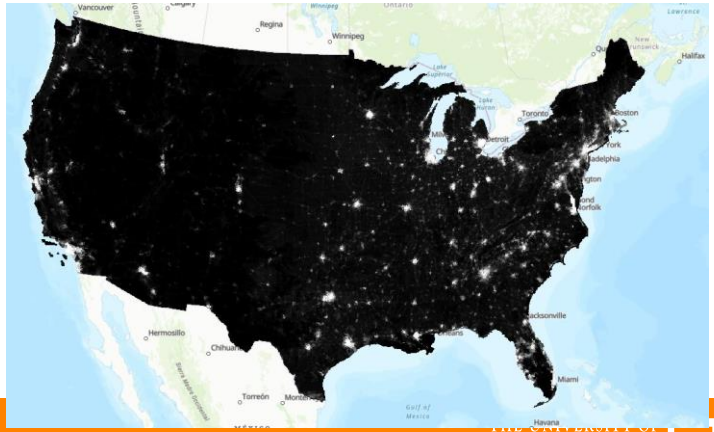
- Are heterogeneity and variability the same?
- Shifts in scale may produce more than averages or constants; they may make homogeneity out of heterogeneity and vice versa.



See “Scale” via Spatial Heterogeneity

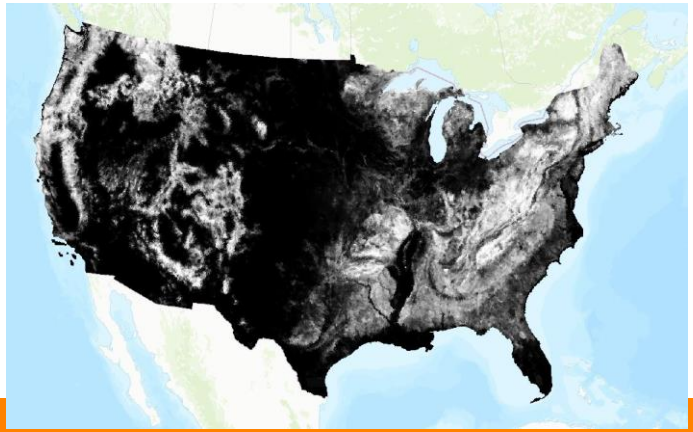
- Functional heterogeneity vs. observed heterogeneity.
- Heterogeneity may involve deterministic, random, and chaotic variations.
- Continuous heterogeneity vs. patchy heterogeneity.

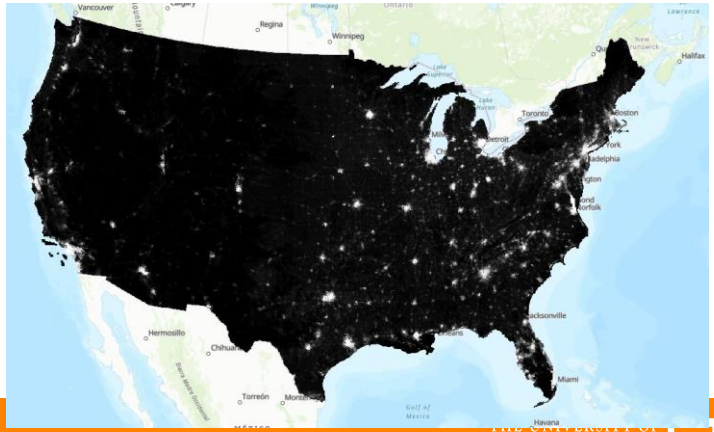




See “Scale” via Spatial Heterogeneity

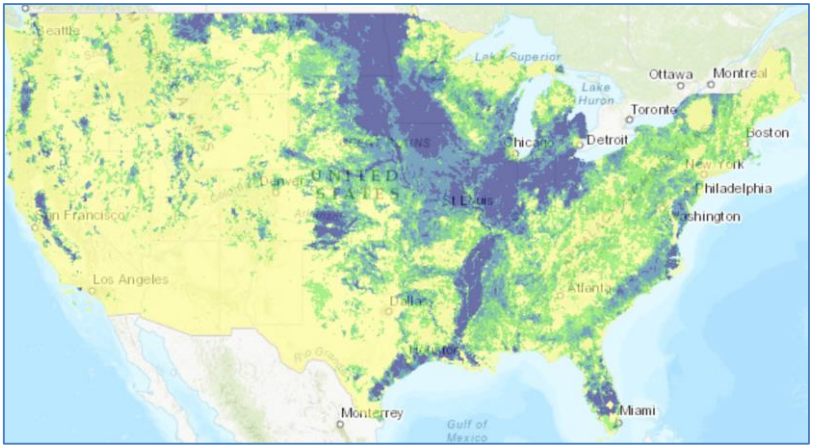
- Arbitrary measures of heterogeneity are tempting and popular, but their ability to reflect the relevant properties of the system of interest is unclear and questionable.





Case study: Mining ecosystem service data

- EnviroAtlas: more than 200 metrics of ecosystem services for ~83,000 HUC-12 units.
- Scale and analyze all metrics (with suitable multivariate methods).
- Case closed.
- Wait a minute!



- Are data across space compatible?

Approaches to multiscale analysis

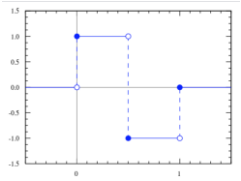
- Indirect approach
  - Designed for single-scale analysis
  - Statistical measures, spatial indicators
  - Realized by sampling data at different scales
- Direct approach
  - Semivariance analysis, wavelet analysis, spectral analysis, fractal analysis, lacunarity analysis, blocking quadrat analysis

Multi-level wavelet analysis

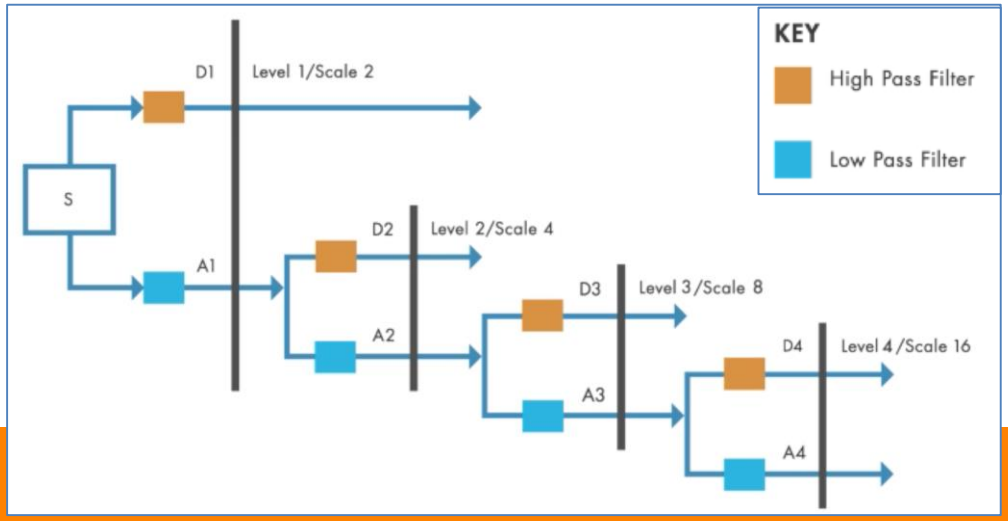
- Wavelet analysis (in a nutshell)
  - Time series data example: to decompose signals (i.e., amplitudes) into different temporal resolutions (i.e., frequency; diurnal, daily, monthly, seasonal, decadal, etc.), and how those patterns change over time.
  - Spatial data: to decompose signals (i.e., magnitudes) into different spatial resolutions (i.e., local, sub-regional, regional, etc.), and how those patterns change over space.

Multi-level wavelet analysis

- Haar wavelet: a sequence of rescaled "square-shaped" functions

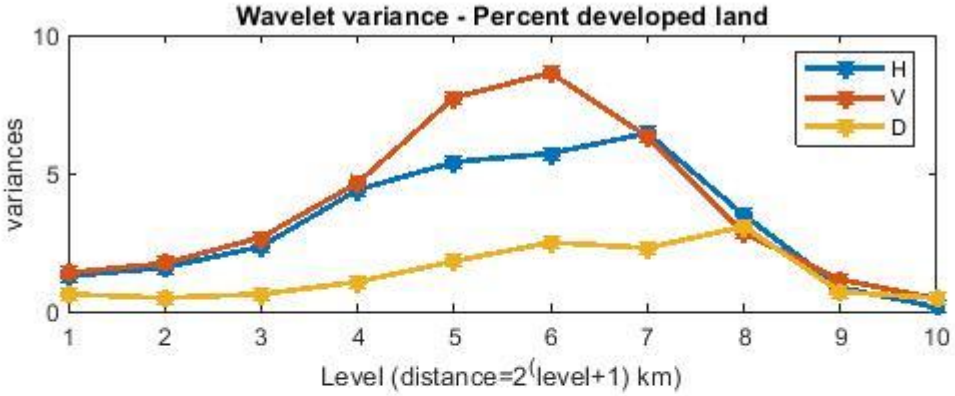


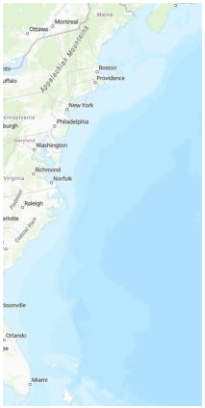
- Multi-level discrete wavelet analysis

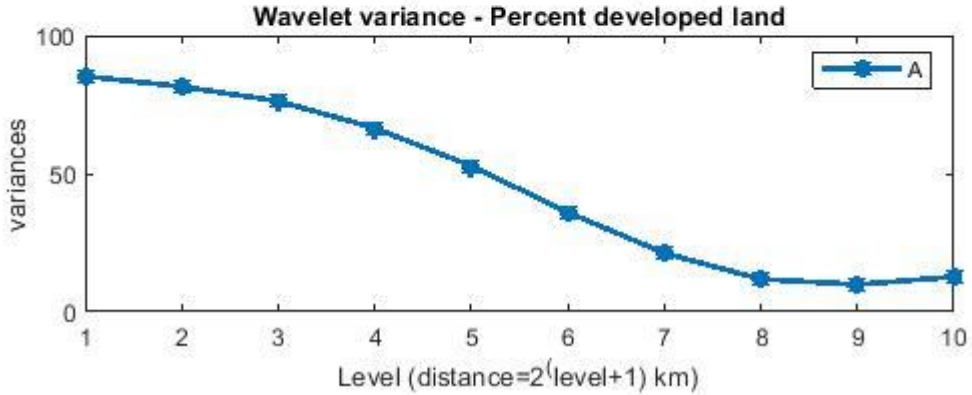


Multi-level wavelet analysis on ecosystem services data

- Example: percent developed land

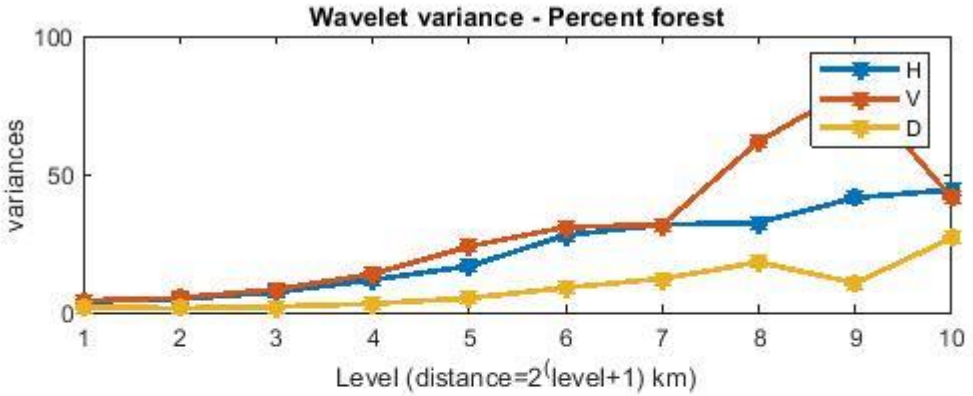


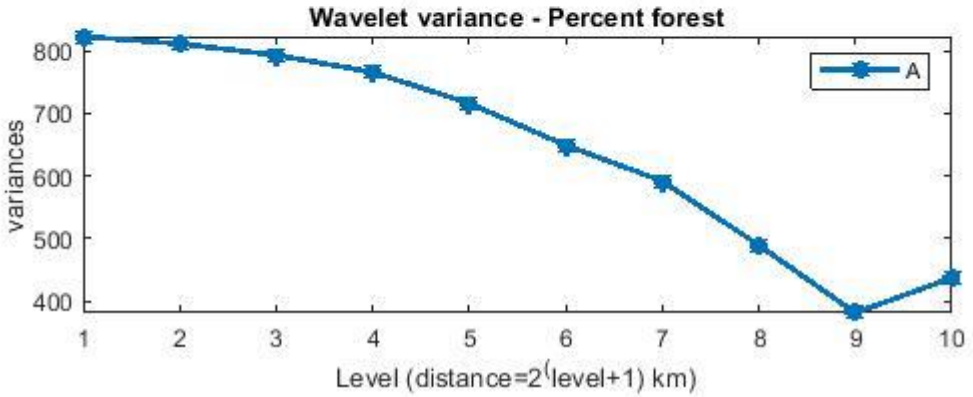




Multi-level wavelet analysis on ecosystem services data

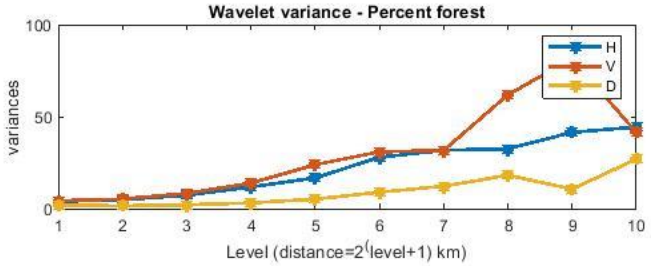
- Example: Percent forest

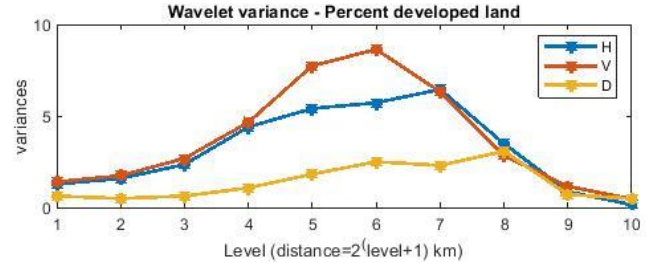




Multi-level wavelet analysis on ecosystem services data

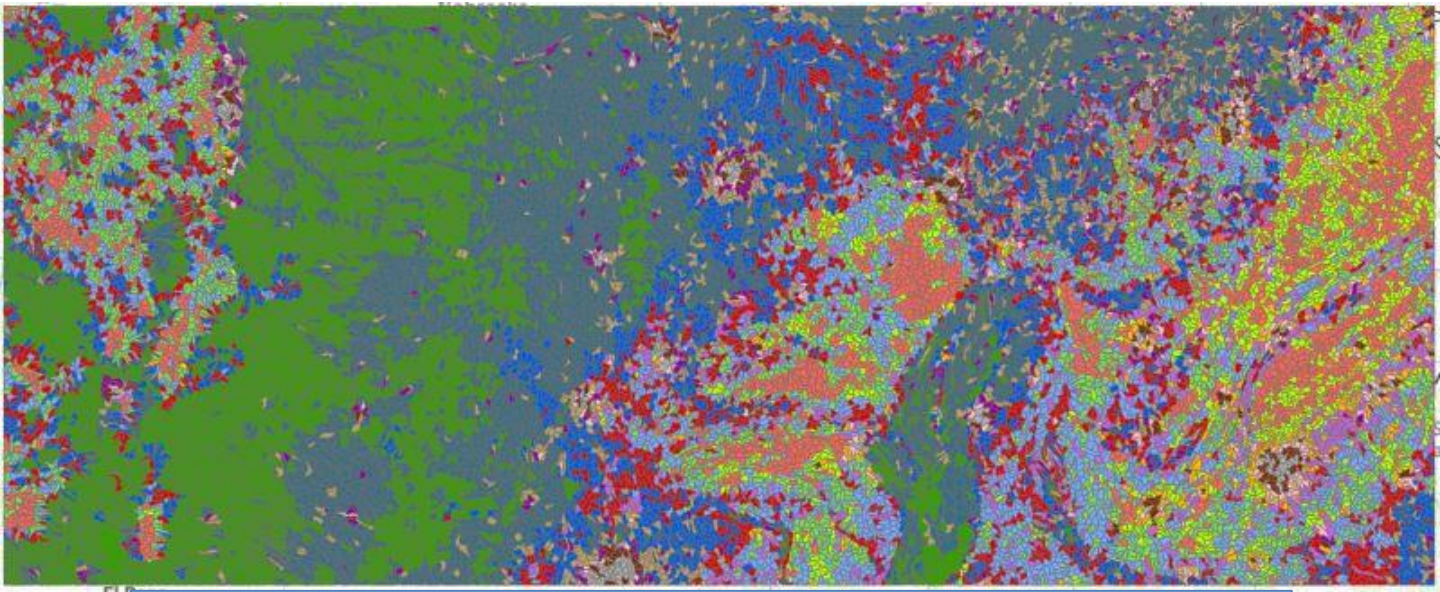
- Putting things together

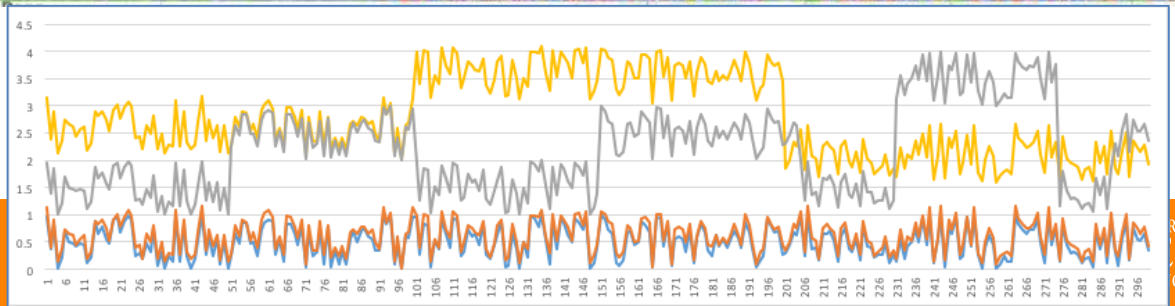




Data mining: looking for pattern

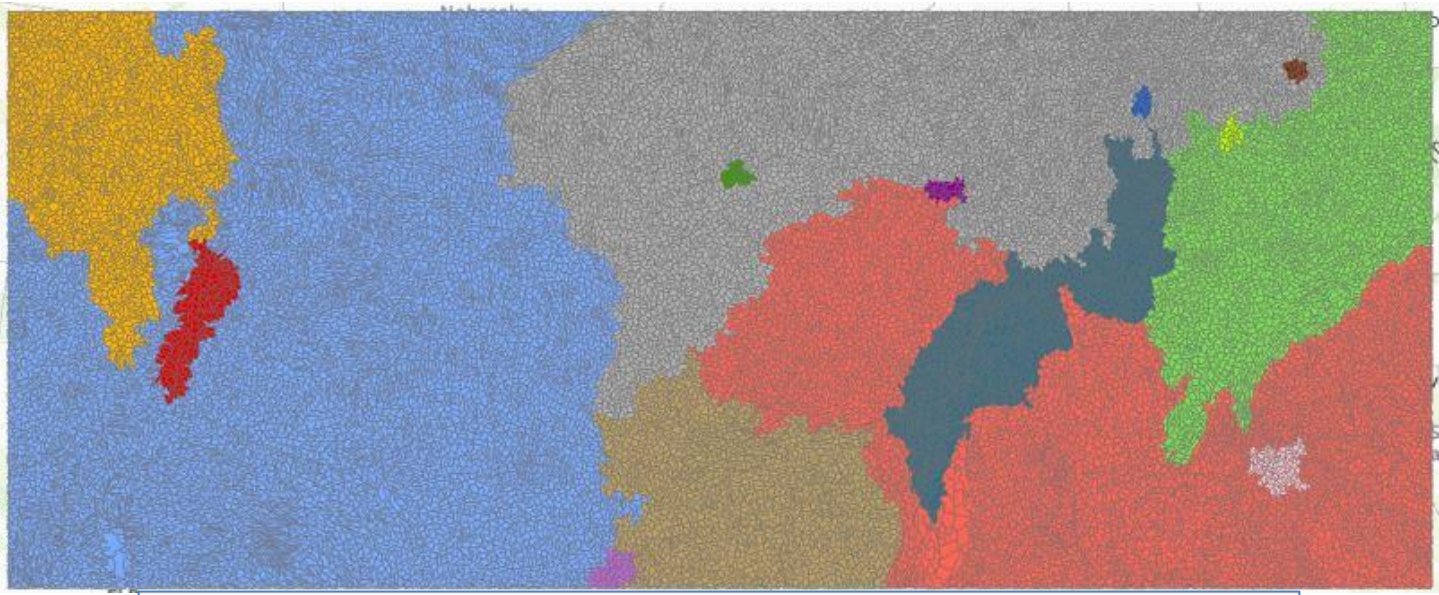
- K-mean clustering: percent forest and percent developed

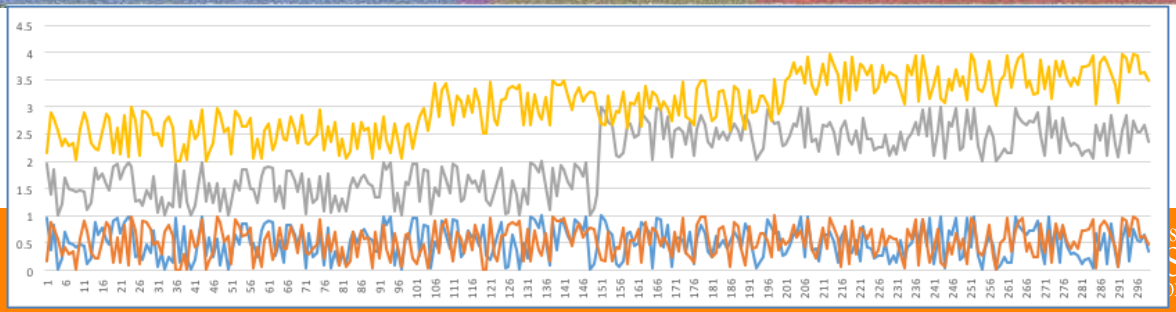




Data mining: looking for pattern

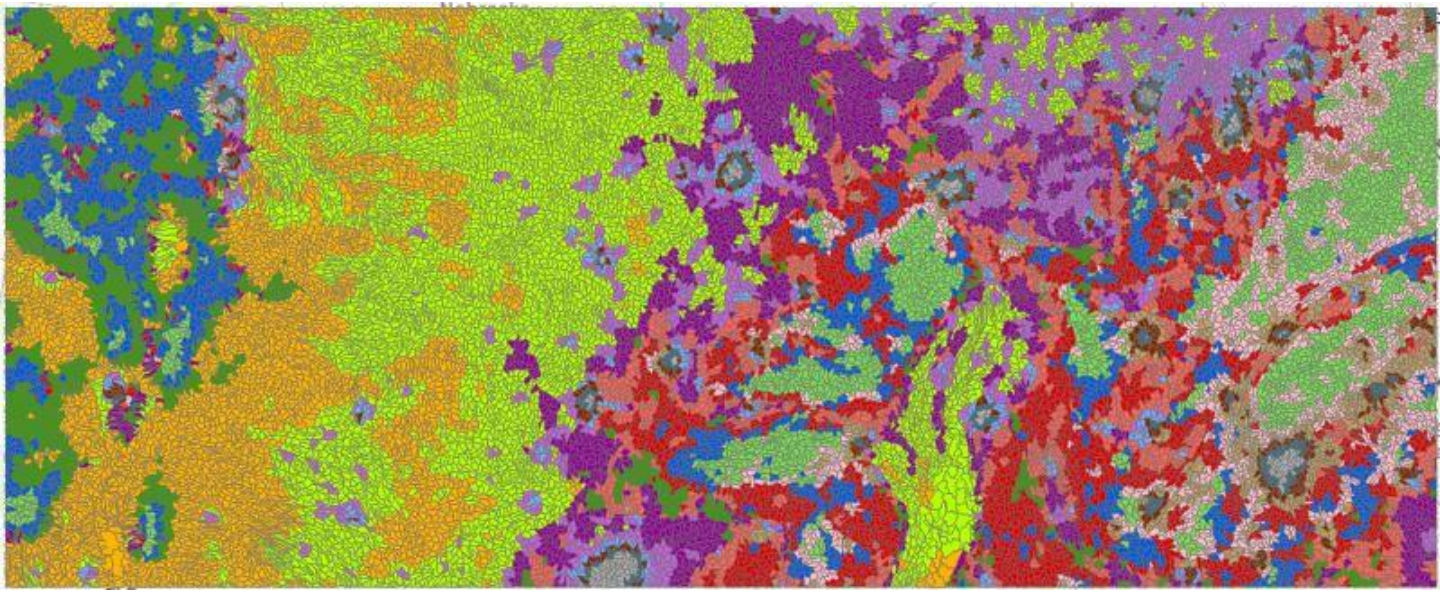
- Spatial-constraint clustering: percent forest and percent developed





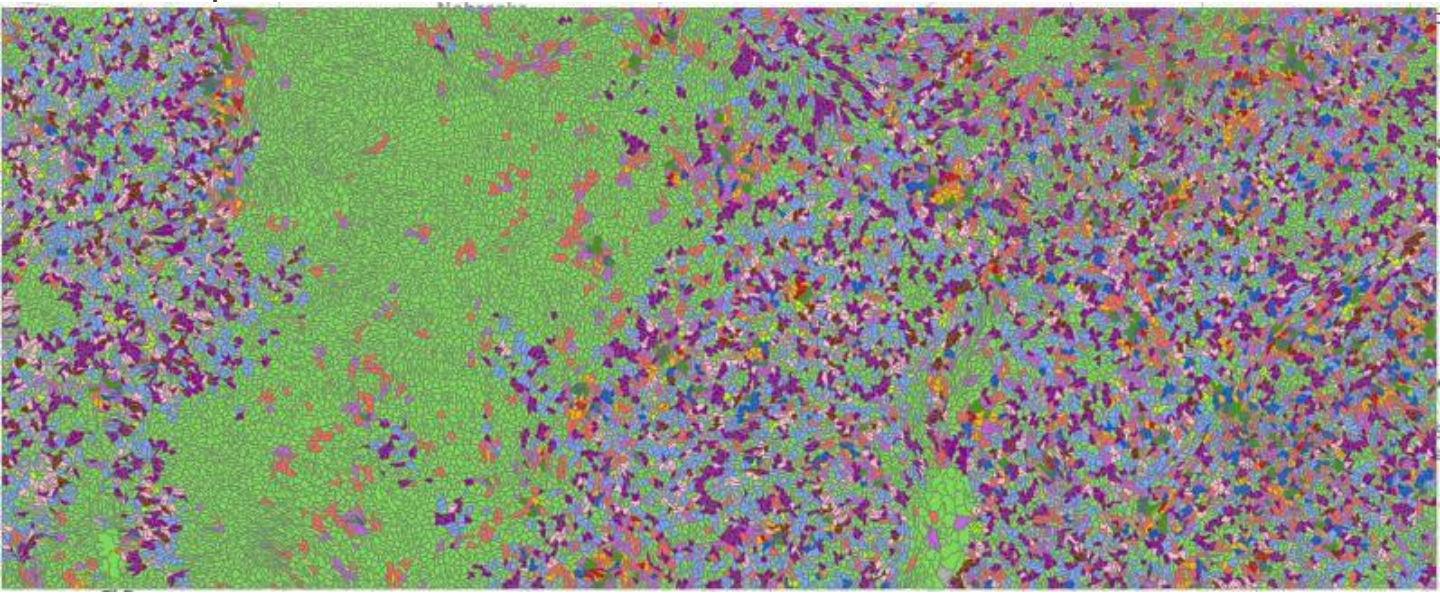
Data mining: looking for pattern

- Scale-based clustering (with A4): percent forest and percent developed



Data mining: looking for pattern

- Scale-based clustering (with D1): percent forest and percent developed



Discussion

- Scale multiplicity of spatial data exists.
- Scale multiplicity of natural landscapes/ phenomena are different from those of socioeconomic landscapes/phenomena.
- Care is needed in putting data which are different from in each other in term of scale multiplicity into the same (data mining) analysis.

Discussion

- Implications:
  - Scale of observation significantly influences what is to be observed and vice versa.
  - Big data need to be treated in a multiple-scaled or hierarchically structured fashion.
  - Meaningful spatial data mining on big data requires a multiple-scale characterization of spatial pattern and processes.
  - Data mining methods developed for big data need to scale in consideration, otherwise …..

Thank you for your time!

Q & A