Indexing Environmental Data Quality

A Multiple Criteria Decision-Making Approach

Ranjan Maitra

Department of Mathematics and Statistics

University of Maryland Baltimore County

www.math.umbc.edu/∼maitra

Joint with Bimal Sinha, UMBC & Phil Ross, USEPA

Outline

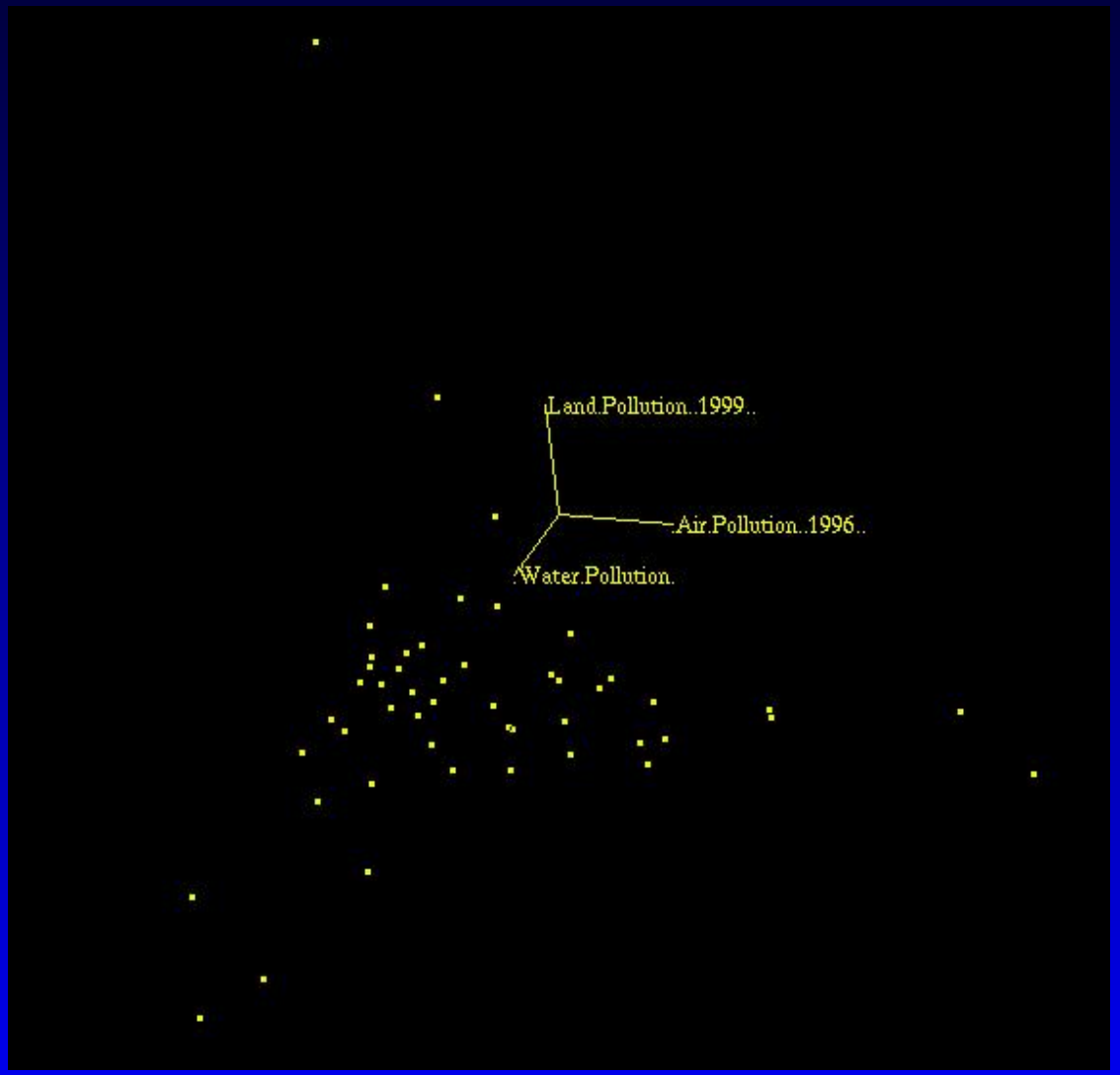
- Introduction and Basic Goals
- Multiple-Criterion Decision-Making
  - Philosophy, Formulation and Extensions
- Application to Environmental Datasets
  - Toxic Release Inventory (TRI) Database
  - United Nations HEI Data
  - US States Data on Air, Water, Land Quality
- Conclusions and Future Directions

Introduction to the Problem

- Have several indicators on data quality
  - Air, water and land environmental quality indicators
  - Measurements on releases of effluents
- Want index of overall environmental quality
  - How do we combine several measures objectively?
  - Can indicators improve program planning, implementation and assessment?
- Do *similar* states perform similarly?

An Illustration

- Environmental indicators on state air, land, water quality



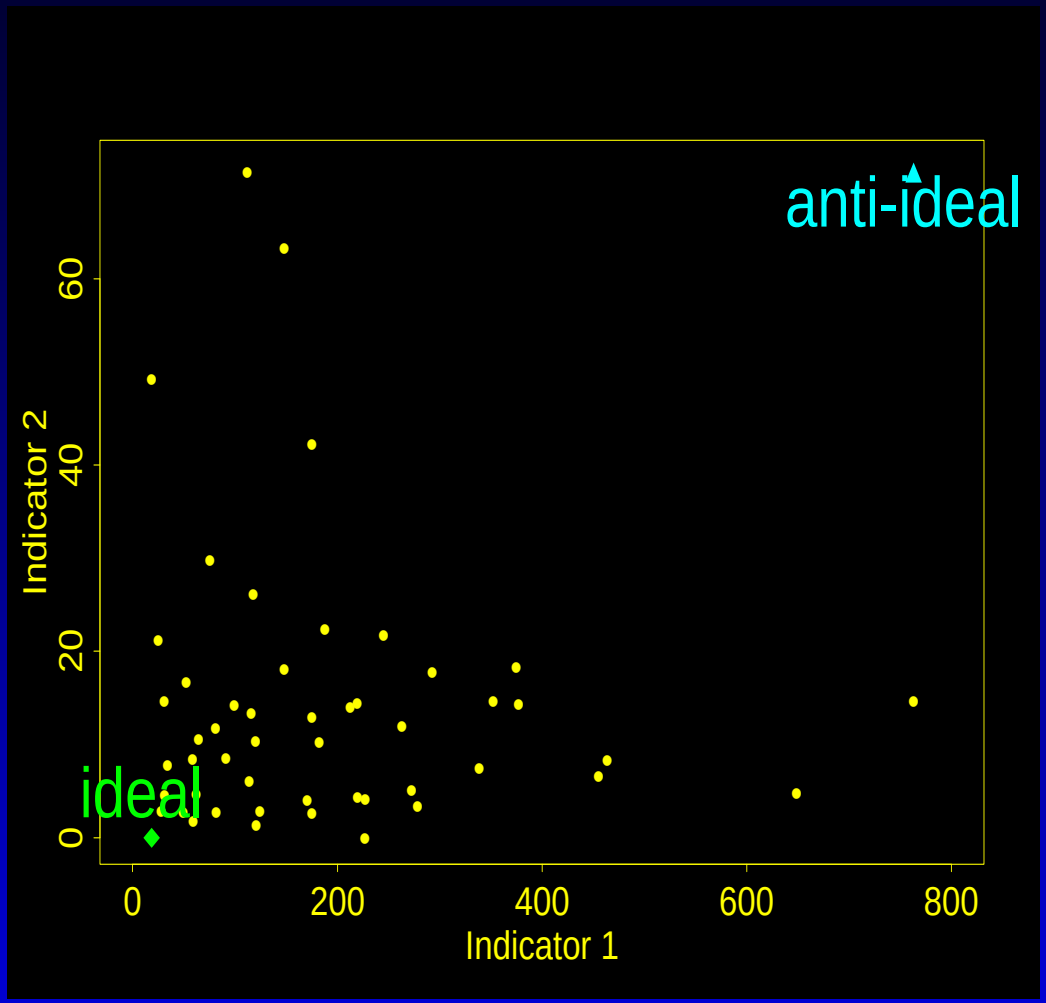
Combining Different Indicators: Options

- Average the different indicators
  - *e.g.* UN Human Environmental Index (HEI) data
  - very simple and straight-forward
  - different indicators have different units.
- Scaling and Averaging
- Use data to objectively derive common index
  - based on actual observed behavior of data
  - can be used for many kinds of data

Multiple Criteria Decision-Making: Basic Goals

- Low values $\rightarrow$ good performance
  - no entity outperforms others for *all* indicators
  - no entity under-performs others for *all* indicators
- Quality indicators are of varying quality
  - *e.g.* TRI database
  - Want to weight indicators appropriately

MCDM – An Illustration



Multiple Criteria Decision-Making: Philosophy

- Want to combine environmental indicators
  - use data to define *ideal*/*anti-ideal* state
    - data-driven, but does not need to be
  - measure deviation of each entity from ideal/anti-ideal
  - derive final index for each entity
    - each factor is weighted by importance
  - can rank sites according to final derived indices

Multiple Criteria Decision-Making Technique – I

- Defining the ideal/anti-ideal:
  - Data $\mathbf{X} = ((X_{ij}))$: $X_{ij} = j$th indicator value for ith entity
  - Ideal: $\wedge = \{\wedge_1, \wedge_2, \ldots, \wedge_K\}$ where $\wedge_i = \bigwedge_i X_{ij}$
  - Anti-ideal: $\vee = \{\vee_1, \vee_2, \ldots, \vee_K\}$ where $\vee_j = \bigvee_i X_{ij}$
- Calculate entity distance from ideal/anti-ideal:
  - $L_2(i, \vee) = \sum_{j=1}^{K} \omega_j \frac{(X_{ij} - \vee_j)^2}{\sum_{j=1}^{K} X_{ij}^2}$; $\quad L_2(i, \wedge) = \sum_{j=1}^{K} \omega_j \frac{(X_{ij} - \wedge_j)^2}{\sum_{j=1}^{K} X_{ij}^2}$
    - weights ω_js lie between 0 and 1; can be chosen in many ways
    - can use norms other than L_2

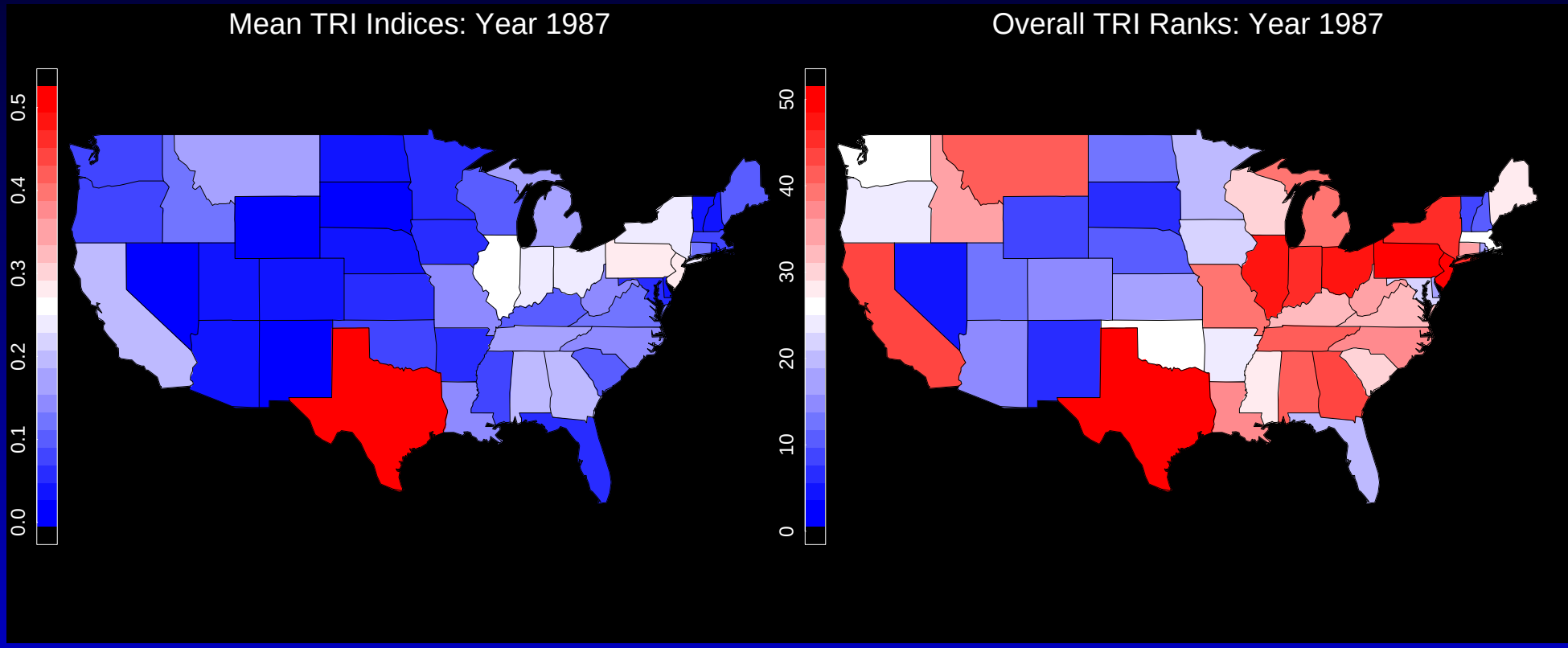
Multiple Criteria Decision-Making Technique – II

- Choosing weights
  - Several options: *eg.* Shannon's entropy
    - Obtain $\phi_j = -\sum_{i=1}^{N} p_{ij} \log p_{ij} - \log N; \qquad p_{ij} = X_{ij} / \sum_{i=1}^{N} X_{ij}$
    - ϕs measure closeness among proportions for each indicator
      - ϕ_j large $\longrightarrow$ large variation among entities: *useful* indicator
      - $\phi_j \equiv 1 \leftrightarrow$ no variation in indicator values: *not useful indicator*
  - Choose $\omega_j = (1 - \phi_j) / \sum_{j=1}^{N} (1 - \phi_j)$
- Derive overall entity indices $I_i = \frac{L_2(i,\wedge)}{L_2(i,\vee)+L_2(i,\wedge)}$
  - use overall indices to rank facilities/entities

Application to Environmental Datasets

- Releases of 17 toxic chemicals (TRI, 1987-94)
- UN HEI data on air, land and water quality
  - *Air*: sum % renewable energy use to total consumption
  - *Land*: % undomesticated land (USGS)
  - *Water*: sum % annual withdrawals of water resources
- US States data on air, land and water quality
  - *Air*: Annual Pollutant Emissions in lbs. (EPA AIRDATA)
  - *Land*: TRI Land releases, 1999.
  - *Water*: surface water quality data, 1999 (*Env. Def.*)

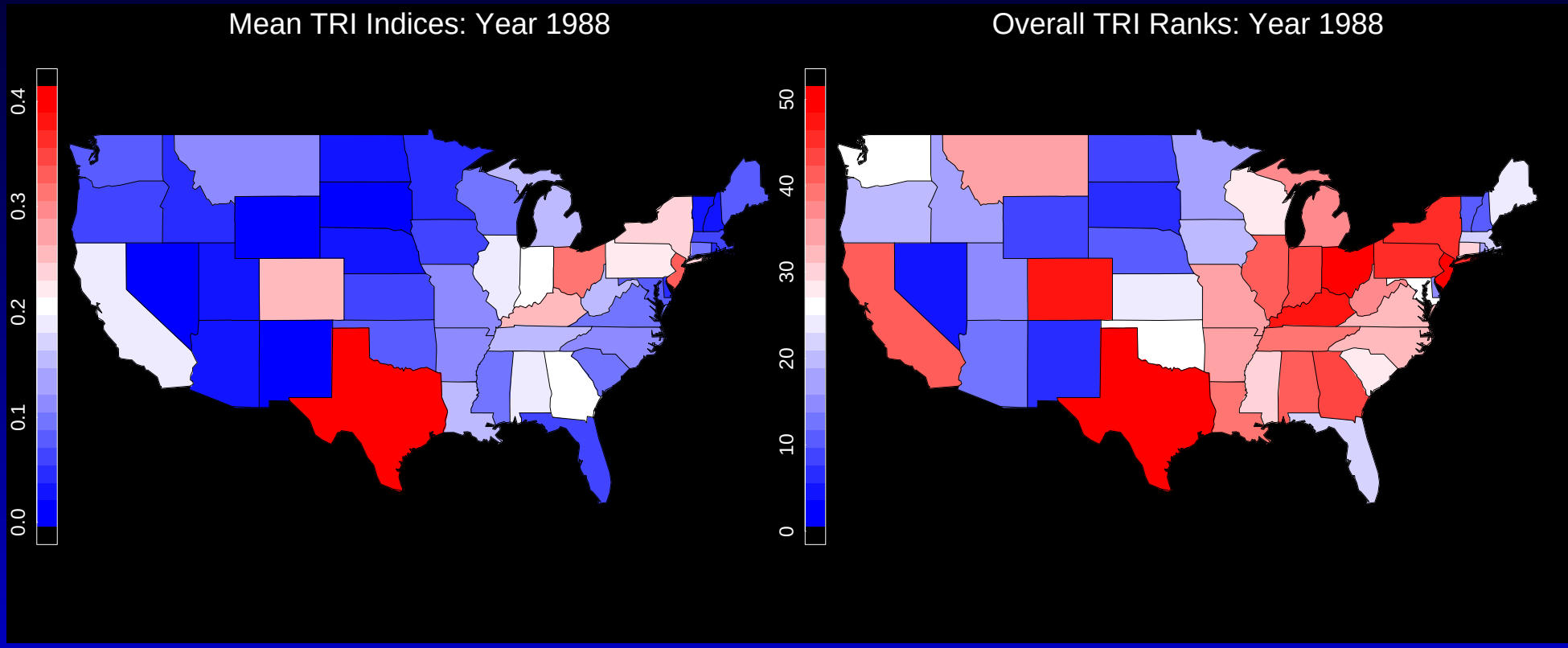
Application to TRI Data



Highest: TX (0.52), PA (0.29), OH (0.24), NJ (0.28), IL (0.27)

Lowest: DC (0.00), HI (0.00), NV (0.00), NM (0.00), SD (0.00)

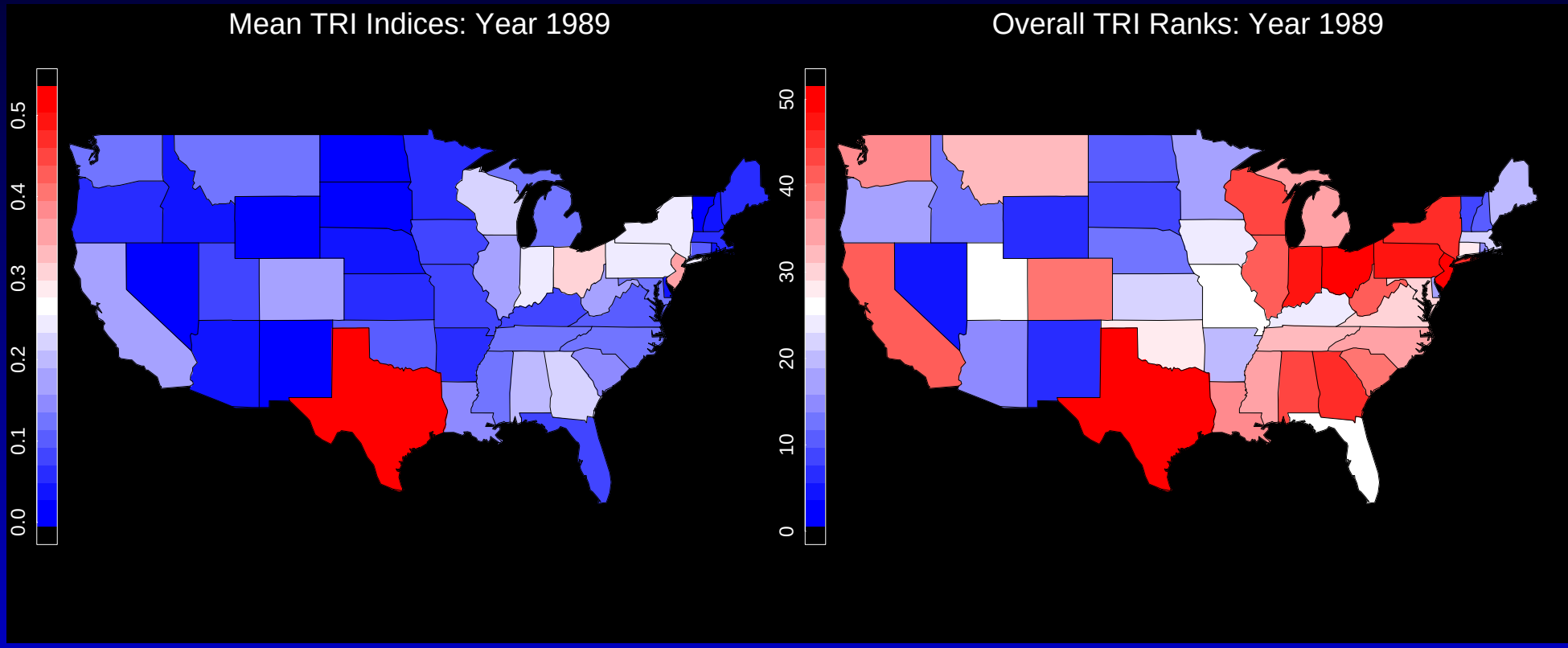
Application to TRI Data



Highest: TX (0.41), NJ (0.33), OH (0.31), KY (0.27), CO (0.27)

Lowest: DC (0.00), NV (0.00), HI (0.00), NM (0.00), SD (0.00)

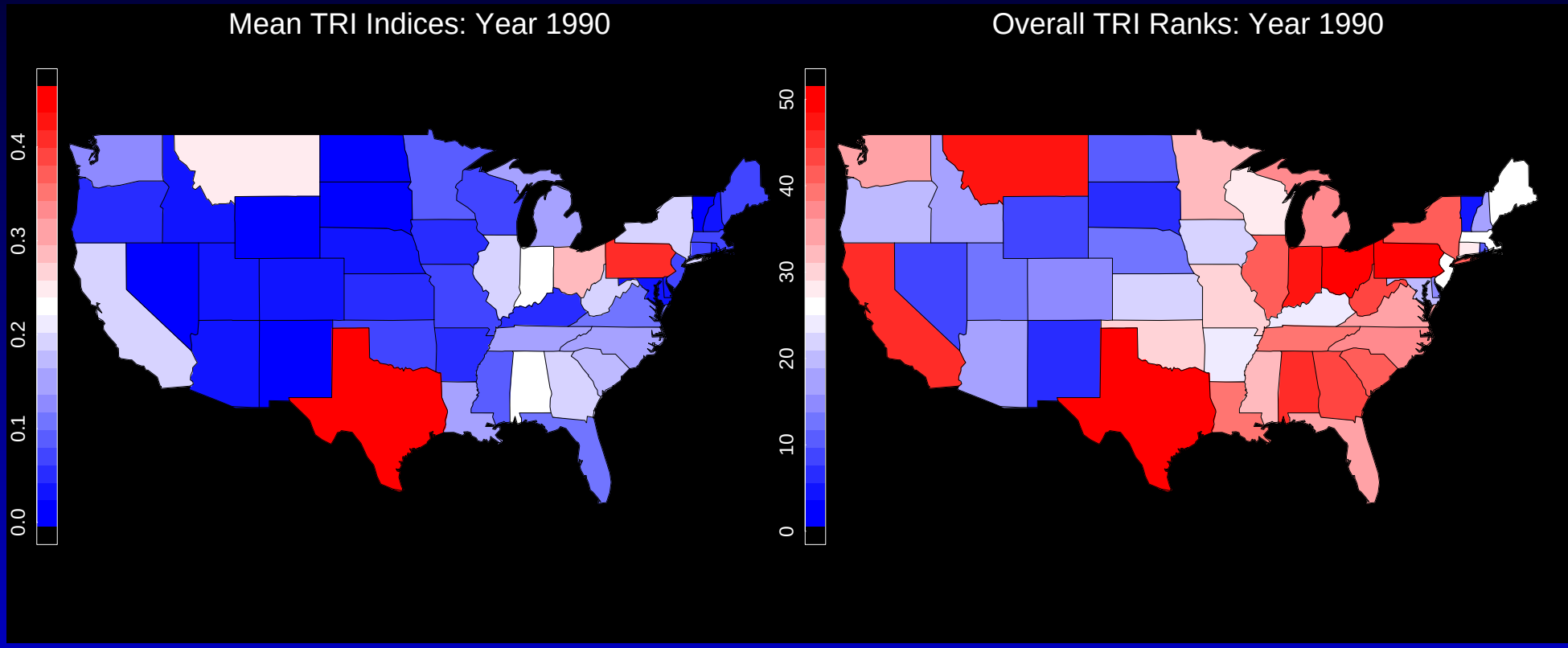
Application to TRI Data



Highest: TX (0.53), NJ (0.36), OH (0.32), PA (0.24), IN (0.24)

Lowest: DC (0.00), HI (0.00), NV (0.00), WY (0.00), NM (0.00)

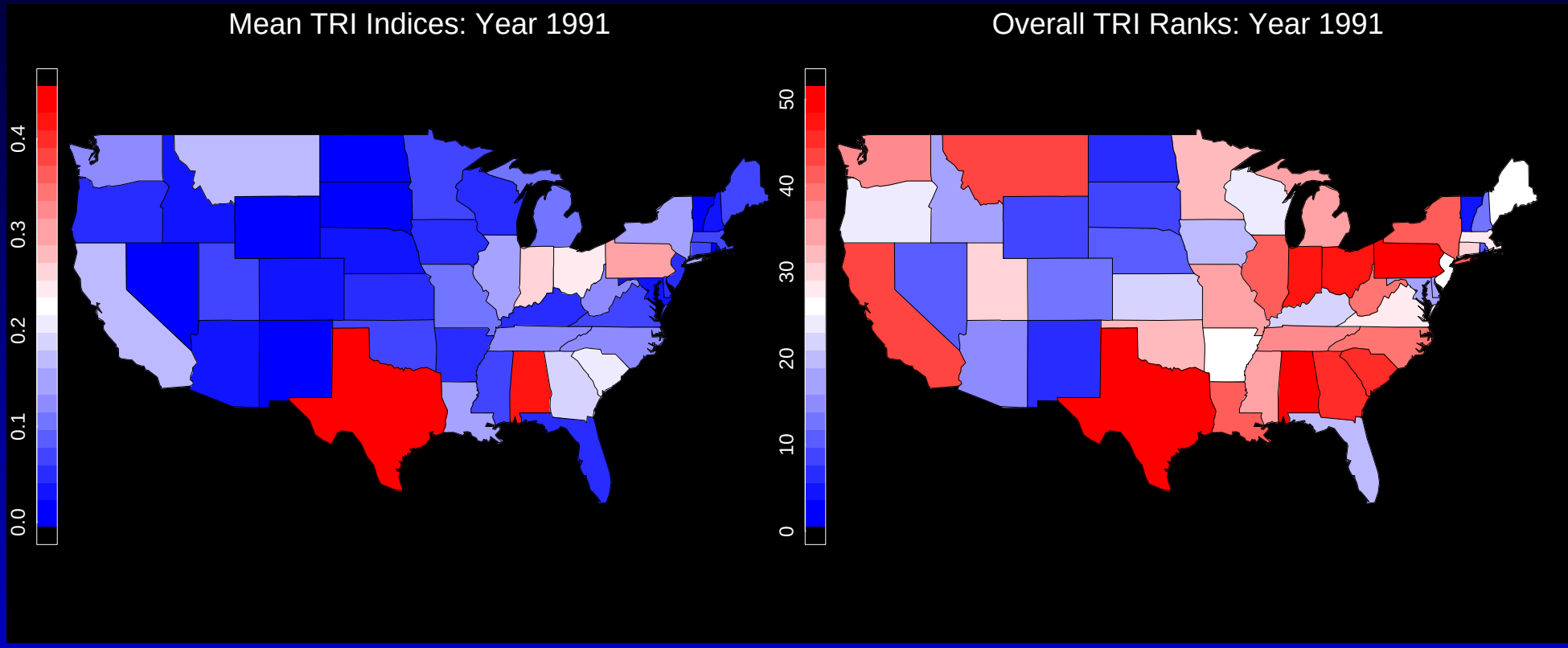
Application to TRI Data



Highest: TX (0.46), PA (0.42), OH (0.29), MT (0.25), IN (0.24)

Lowest: DC (0.00), VT (0.00), HI (0.00), NM (0.00), SD (0.01)

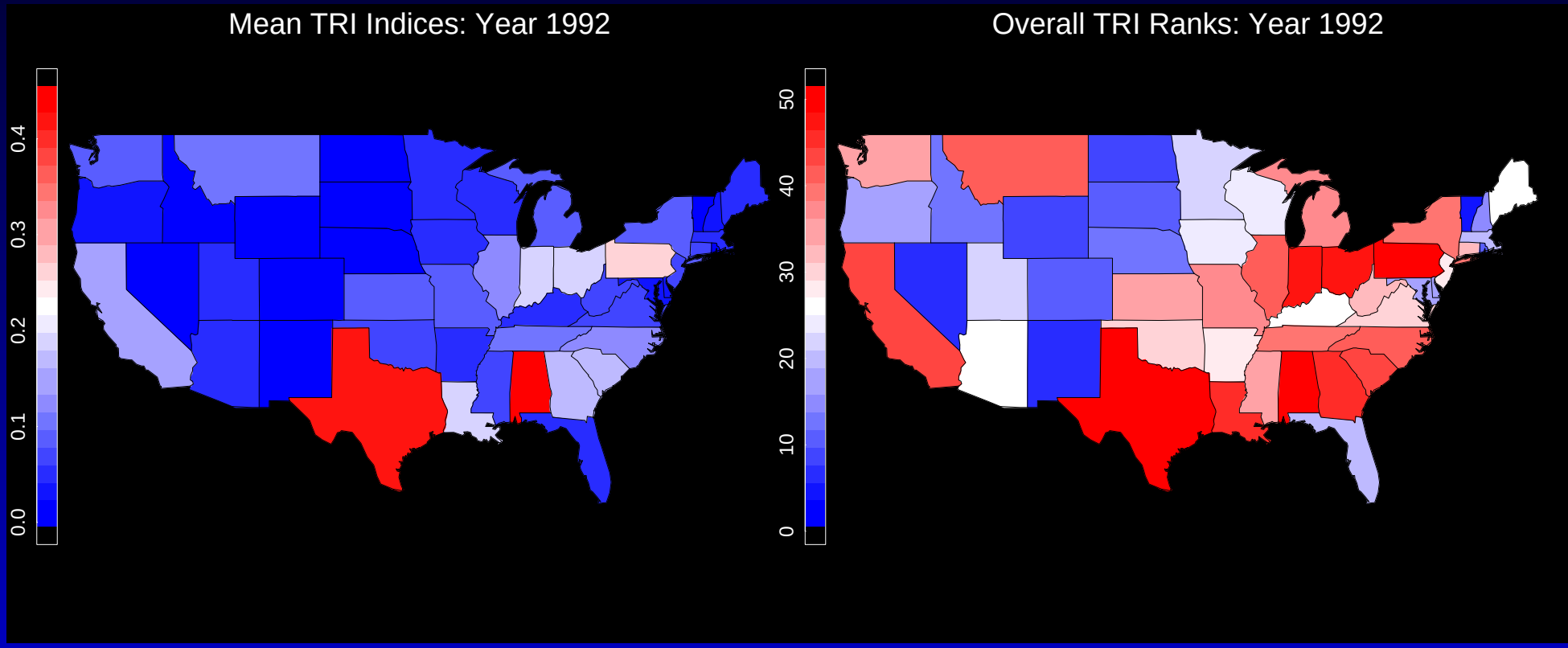
Application to TRI Data



Highest: TX (0.45), AL (0.43), PA (0.30), IN (0.27), OH (0.24)

Lowest: DC (0.00), VT (0.00), HI (0.00), NM (0.00), ND (0.00)

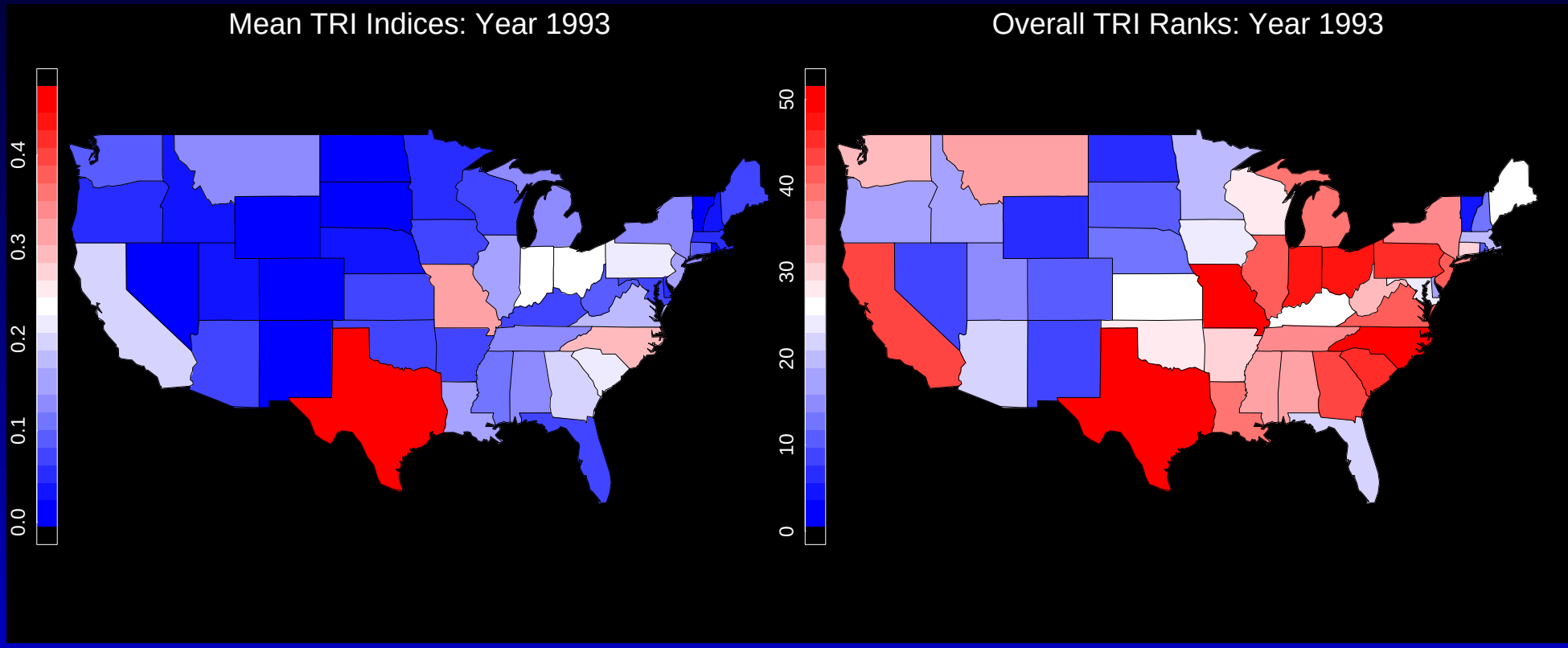
Application to TRI Data



Highest: AL (0.45), TX (0.42), PA (0.27), OH (0.18), IN (0.18)

Lowest: DC (0.00), HI (0.00), VT (0.00), NM (0.00), NV (0.00)

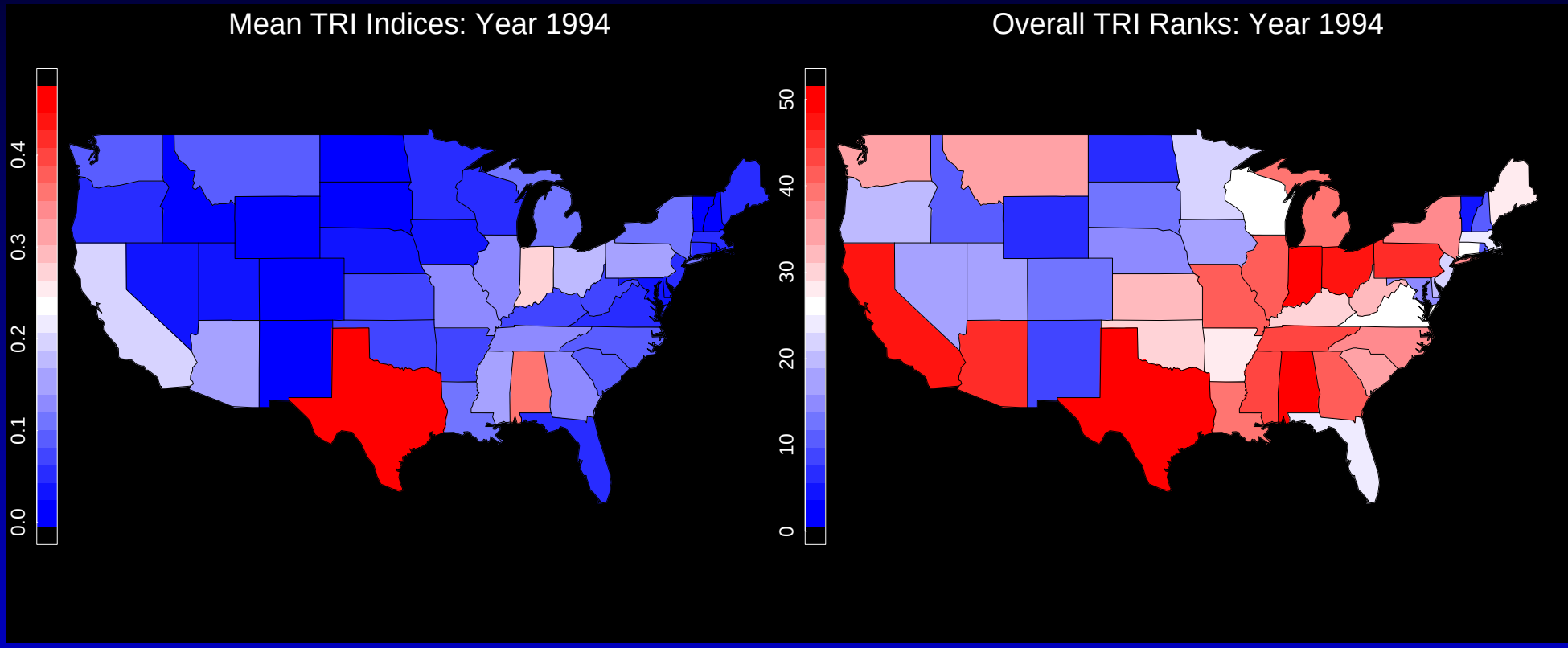
Application to TRI Data



Highest: TX (0.47), MO (0.33), NC (0.30), IN (0.24), OH (0.23)

Lowest: DC (0.00), HI (0.00), VT (0.00), WY (0.00), ND (0.00)

Application to TRI Data



Highest: TX (0.47), AL (0.36), IN (0.28), CA (0.19), OH (0.17)

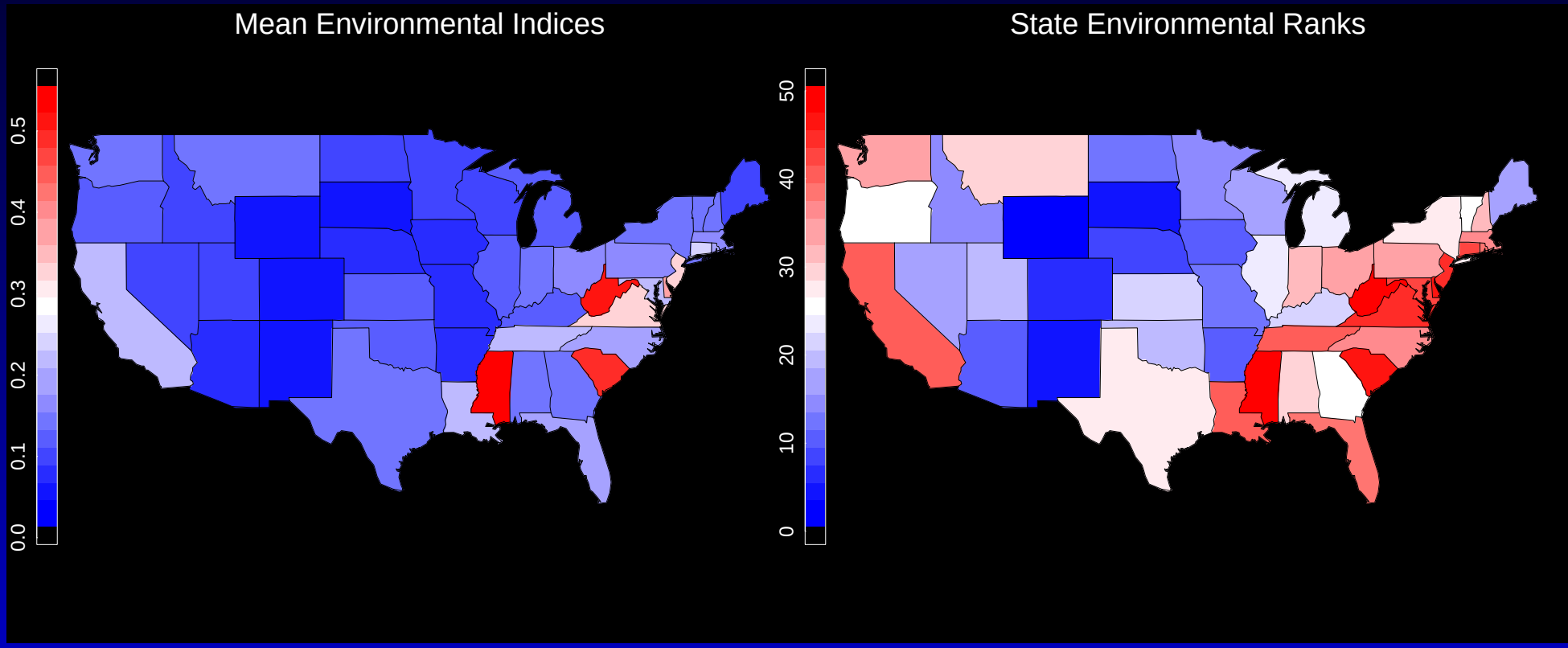
Lowest: DC (0.00), VT (0.00), HI (0.00), ND (0.00), e WY (0.00)

Countries' HEI Data

Country	MCDM Mean	MCDM Rank	HEI Rank	Country	MCDM Mean	MCDM Rank	HEI Rank
Moldova	0.14	1	1	Costa Rica	0.89	106	106
Trinidad & Tobago	0.17	2	2	Ghana	0.81	105	102
Ukraine	0.19	3	6	Bangladesh	0.80	104	76
Azerbaijan	0.21	4	7	Guatemala	0.80	103	103
Bulgaria	0.22	5	9	Sri Lanka	0.80	102	75
Hungary	0.24	6	3	El Salvador	0.80	101	87
Uzbekistan	0.26	7	10	Cameroon	0.79	100	101
Belgium	0.26	8	5	Peru	0.77	99	100
The Netherlands	0.27	9	4	Uruguay	0.77	98	72
Turkmenistan	0.29	10	11	Albania	0.77	97	57
				The Congo	0.76	96	95

Highest- and Lowest-ranked countries of the world

States Air, Water, Land Quality Data



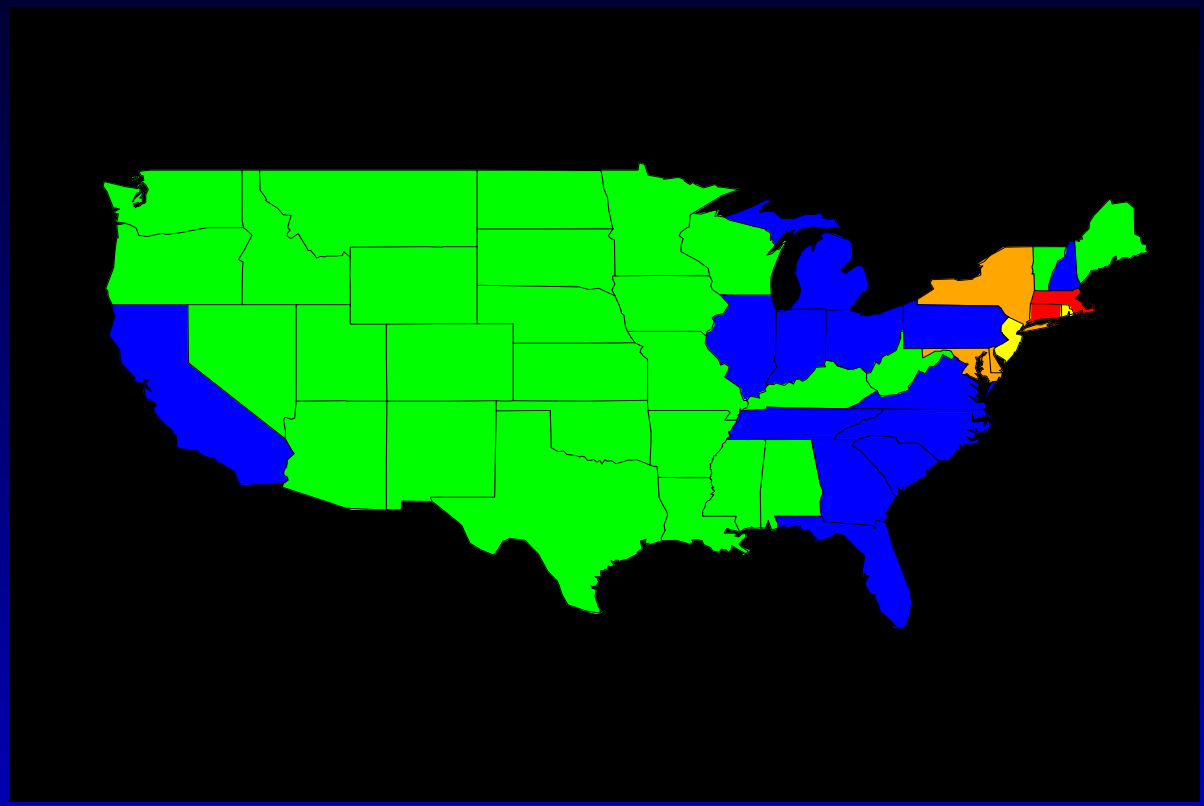
Highest: MS (0.55), WV (0.51), SC (0.49), DE (0.38), NJ (0.34)

Lowest: WY (0.02), SD (0.02), NM (0.03), HI (0.04), CO (0.04)

Analysis of States

- Does ranking all states give us the most information?
  - different states have different characteristics
    - population densities, land use, economy vary
- Cluster states into homogeneous groups based on population density
  - could also cluster on characteristics such as economic measure, sales of gasoline, etc.
  - compare states in each subset

States Grouped by Population Density



- Agglomerative hierarchical clustering using average-linking

MCDM on State Groups

- Group I:

  WY (.02), SD (.03), NM (.04), CO (.04), AL (.04), NE (.05), AZ (.05), NV (.07), IA (.07), MO (.07), AK (.07), MN (.08), ND (.08), ID (.08), UT (.08), ME (.09), WI (.09), OK (.11), KY (.11), KS (.12), WA (.12), MT (.13), OR (.13), VT (.14), TX (.15), AL (.15), LA (.22), WV (.44), MS (.64)

- Group II:

  GA (.01), CA (.01), IL (.05), NC (.07), HI (.08), IN (.08), TN (.09), FL (.10), MI (.12), PA (.13), OH (.18), NH (.25), SC (.26)

MCDM: Conclusions and Further Work

- MCDM is scientific and flexible
- Can be used for decision-making
  - can analyze data clustered at regional level, etc.
    - Do groups behave differently because of enforcement, implementation, policy?
- Can provide indirect assessment of data quality
  - why does one entity behave differently than others very much *like* it?
  - can identify facilities with aberrant data