Defining Geographic Climate Regions in the Continental US: A Spatial Clustering Approach

Erin M. Ochoa • emo@uchicago.edu • github.com/emochoa • 2017 June

THE UNIVERSITY OF CHICAGO

Why Are Empirically Defined Climate Regions Necessary?

- Climate characterizes long-term average weather conditions and varies regionally.
- Regional climate may affect social phenomena such as crime, health, and economic conditions. Spatial social-science research based in the United States therefore requires the existence of a robust and empirically defined model of climate regions.
- Current models of geographic climate regions are insufficient for research:
  - A 9-region model (undated) by the National Oceanic and Atmospheric Administration (NOAA) is drawn along state lines and does not detail what data or methods were used to create it.
  - The Peel (2007) update to the Köppen-Geiger model is based solely on rainfall and temperature, eschewing conditions such as cloudiness, windiness, wind chill, heat index, dew point (i.e., humidity), barometric pressure, and snowfall.
- Spatial clustering via SKATER (Spatial Kluster Analysis by Tree Edge Removal) is employed to define a climate region model based on 30-year climate data averages.

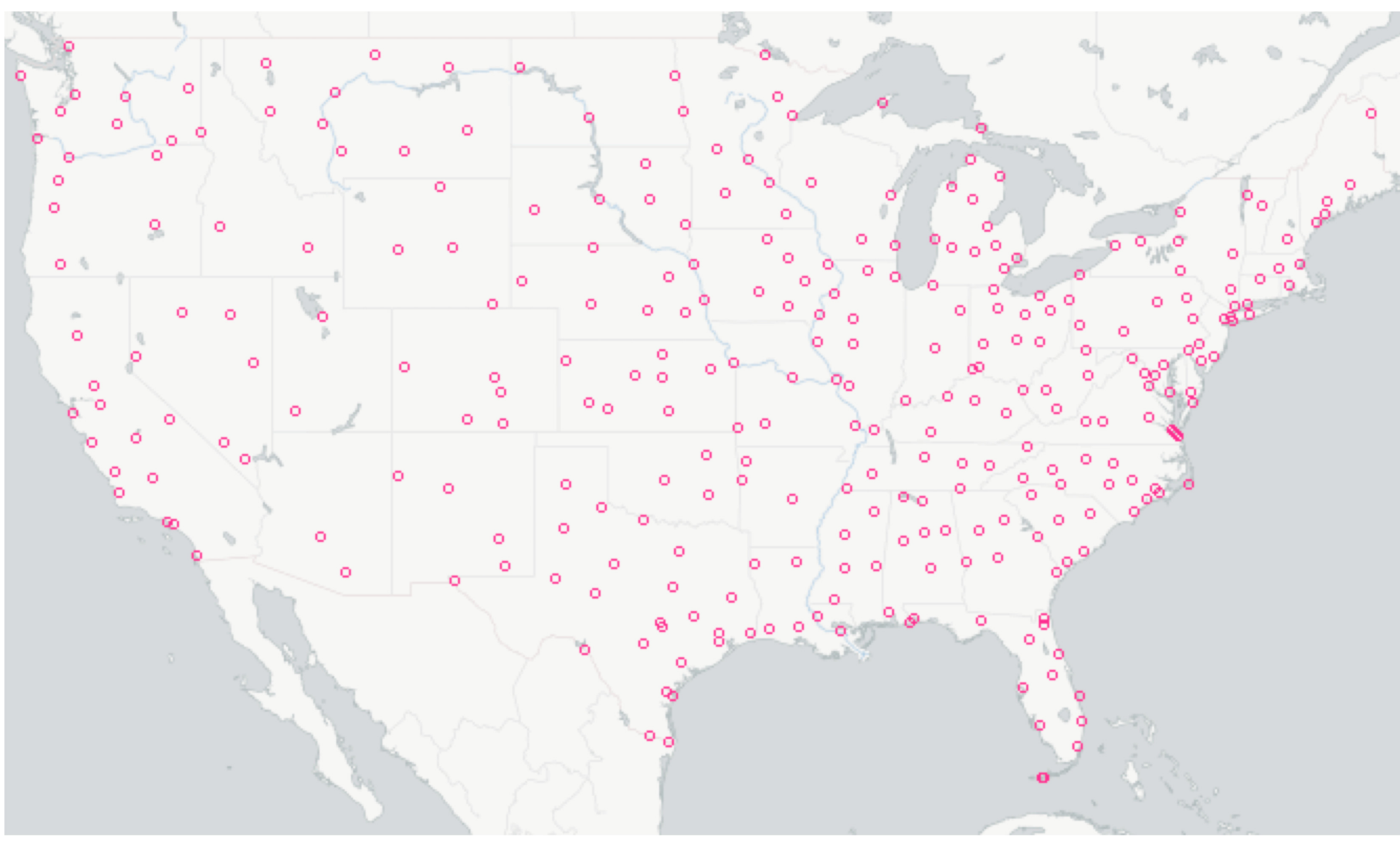

Figure 1: Weather stations (N=302).

Data

- The data set used was the 1981–2010 US Climate Normals from NOAA's National Centers for Environmental Information.
- The data set consists of climate data for over 9,000 weather stations in the United States and its territories.
- Raw climate data were collected hourly or daily, depending on the measure, and 30-year averages were reported as either hourly normals (one value for each hour of each calendar day) or year-to-date normals (one value for each calendar day).
- Station metadata includes latitude and longitude coordinates, as well as elevation.
- Retaining only the weather stations in the continental United States and which reported data for all the climate variables of interest, a final sample of 302 stations was selected. The weather stations are mapped in Figure 1 above.

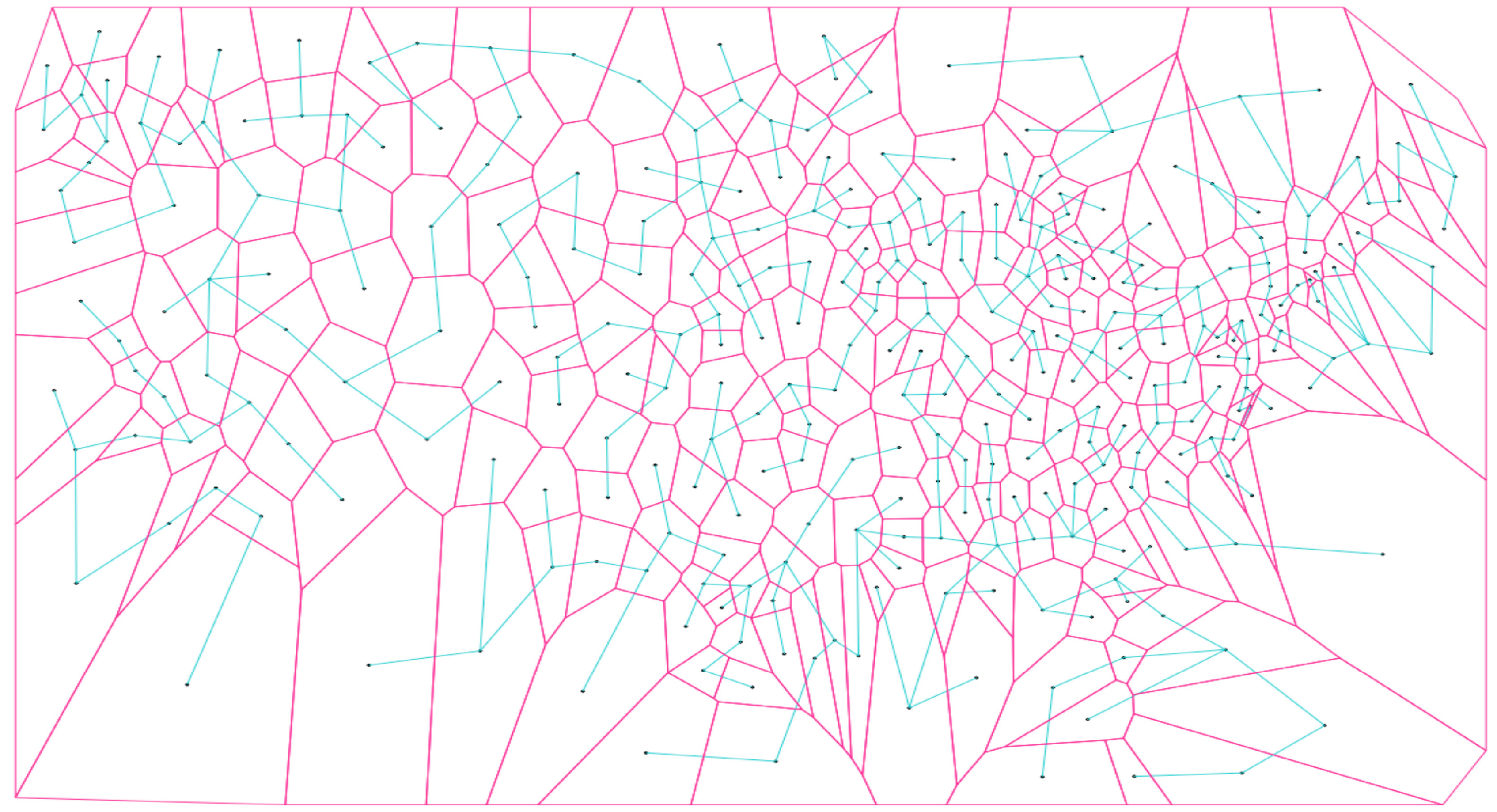

Figure 2: Connectivity diagram. Thiessen polygon boundaries are shown in pink; weather stations are black points; neighbors are connected with blue lines.

Methodology

- The unit of analysis was the weather station and included each station's elevation.
- Averaging each condition by month of the year, 443 climate variables were created.
- Monthly mean minimum and maximum values were recorded for most conditions.
- Daylength data for each station were generated in Python using the Astral package.
- An alternative geography consisting of Thiessen polygons (with weather stations as the centroids) was created and neighbors were assigned using queen contiguity.
- Climate-region models were computed based on scaled and centered variables with the SKATER function in R's spdep library. The number of regions, K, was varied.

	Constant	Per-Month* Mean	Per-Month* Min. Mean	Per-Month* Max. Mean
Elevation	✓			
Daylength		✓		
Rainfall		✓		
Snowfall		✓		
Temperature		✓	✓	✓
Dew point		✓	✓	✓
Heat index		✓		✓
Wind chill		✓	✓	
Wind speed		✓	✓	✓
Winds: % calm		✓	✓	✓
Barometric pressure		✓	✓	✓
Clouds: % clear		✓	✓	✓
Clouds: % scattered		✓	✓	✓
Clouds: % broken		✓	✓	✓
Clouds: % overcast		✓	✓	✓

Figure 3: Conditions measured for each weather station (N=302), resulting in a total of 445 variables. *There are 12 monthly variables, January through December.

Results & Conclusions

- The climate-region model with K=10 strikes the best balance between parsimony and granularity, separating the map into regions that are sized reasonably and fit well with common knowledge of climate conditions across the country.
- With K=9, the majority of the Plains and Rocky Mountain areas are classified into a single vast climate region that dominates the map. This region is undesirable because the granularity is too low, grouping regions known to have distinct climates.
- With K=11, a very small region emerges along the California coast. This region is suboptimal because when used in social-science research, it might contain too few observations and thereby lead to a decrease in statistical power.

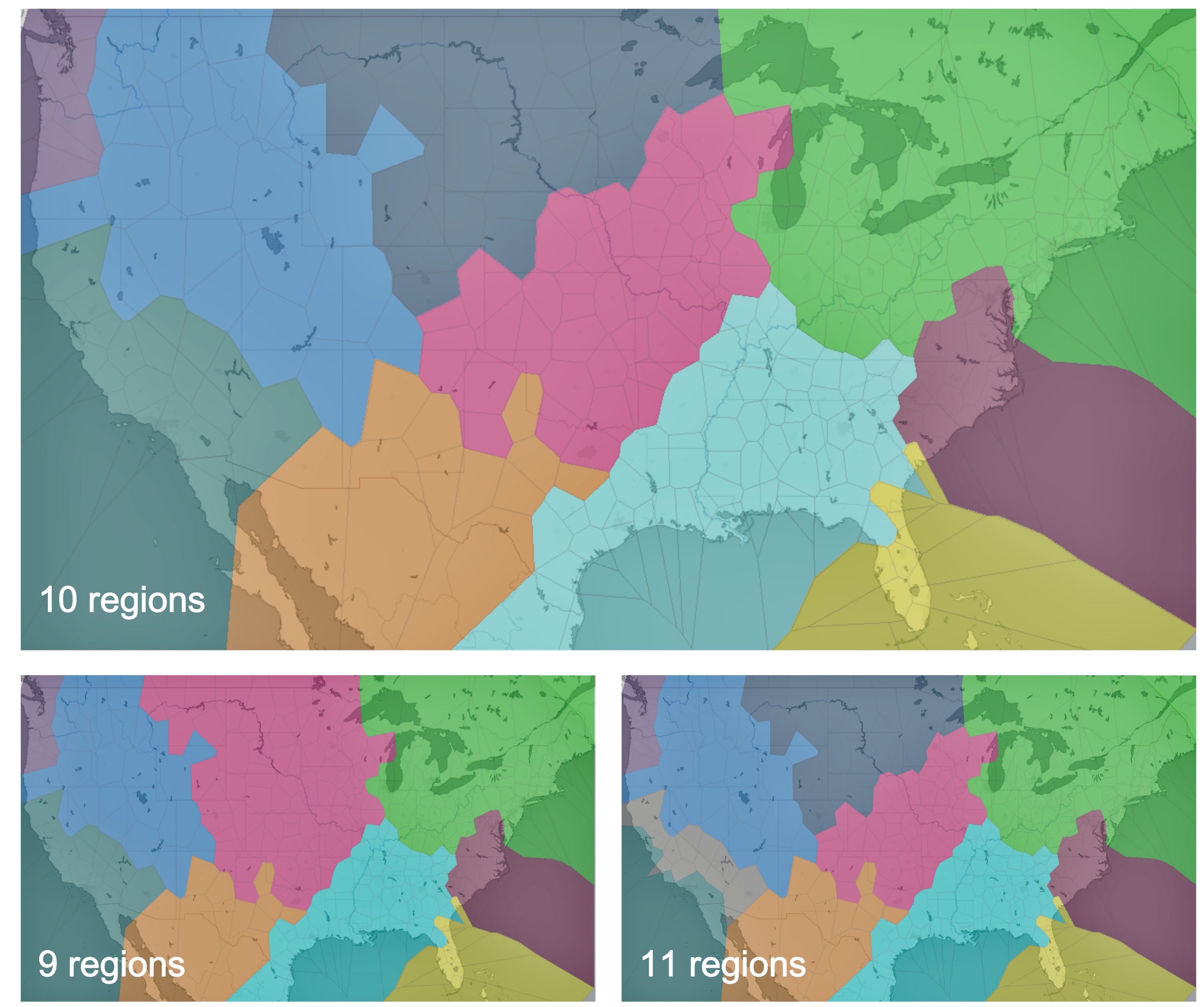

Figure 4: Thiessen polygons (N=302) clustered into climate regions with SKATER.

Limitations & Future Directions

- The results of SKATER do not indicate which climate conditions are the most influential. Determining which conditions contribute the most to the classification of clusters could lead to the development of a more parsimonious model.
- The number of weather stations used (N=302) was limited because of the inclusion of a wide variety of climate conditions; the resulting Thiessen polygons are large in some areas. Smaller polygons could be achieved by using fewer conditions such that there is an increase in the number of weather stations available.
- SKATER does not offer a method for choosing the ideal value for K. Clustering methods that are both spatial and do offer such methods should be considered.