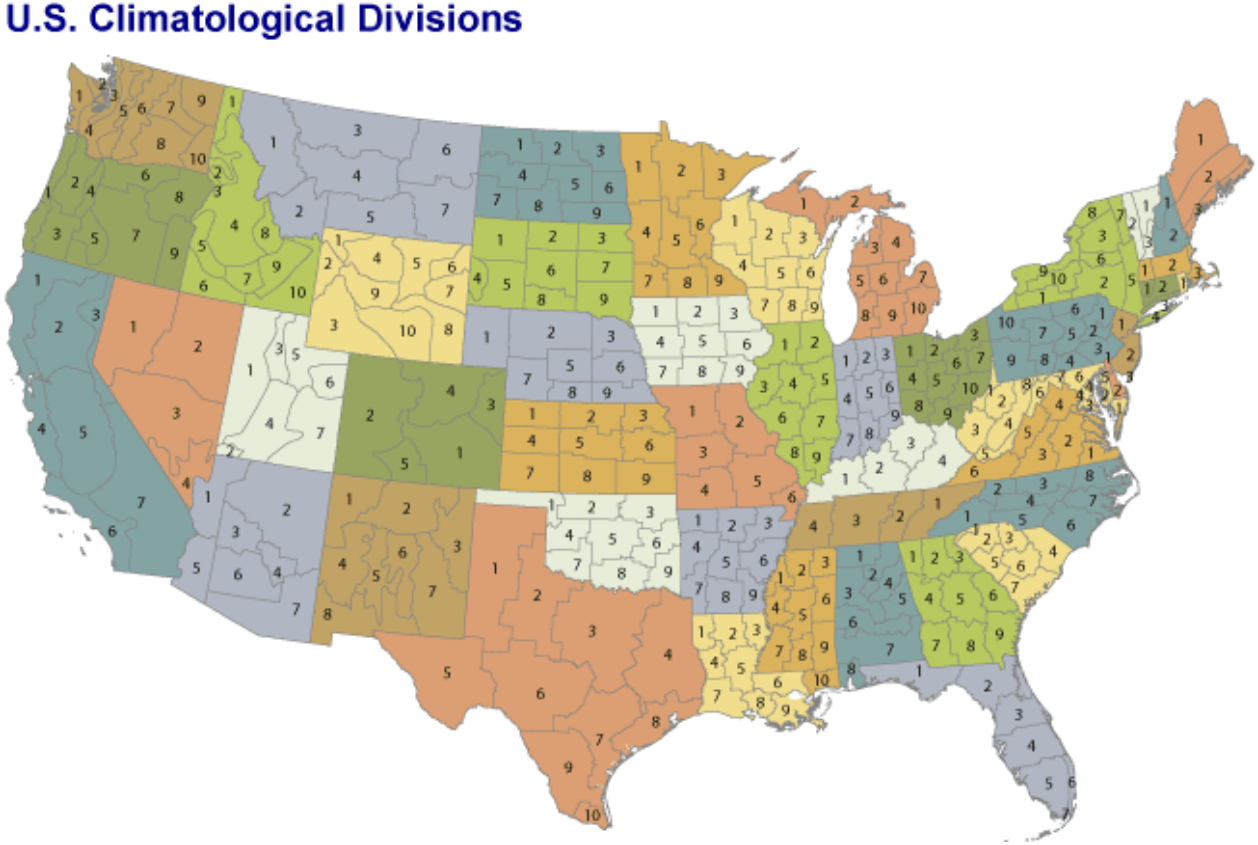
A Case Study

Climate change is a hotly debated topic. Climatologist claim the world is at a teetering point, and the damages will soon be irreversible if we don't make swift changes. Opponents claim there is no solid proof that demonstrates climate change is real, and that the data behind it is *cooked*, or falsified. Is there any truth to their claim?

The **NOAA**, National Oceanic and Atmospheric Administration, has long produced the only spatially complete, long-term (1895-2013) dataset for climate analyses within the US. In the Climate Divisional Dataset, each state is represented by 6-10 climate divisions, as shown on the map below. The **monthly** temperature readings per climate division are then calculated by averaging the **daily** observations from their ground stations. By using climate divisions and not individual weather stations, and by using monthly divisional averages for data-samples, the overall data collection process was eased considerably. After all, we are talking about government agency that formed over 200 years ago! Another added benefit is that less concern need be placed on the accuracy of specific station measurements, in case of local inconsistencies. Everything should just get *smoothed over* by averaging out.

Dive Deeper

- 5. Data Modeling



In 2012, the NOAA must have come into better funding because they exerted an open effort to recalculate their historical climate dataset. They added in thousands of past observations that had only recently been digitized (remember when we used to use paper and pencil?), and adjusted their interpolation formula to remove some *known biases*, such as jumps introduced by replacing instruments, observation practices such as differences in readings based on when in the day temperatures are recorded, inaccuracies caused by station moves, urbanization around the station, etc. The NOAA further went on to retrospectively adjust their historic dataset as well, to be more in line with the current observations at each station, and to make use of the new interpolation method.

Skeptics were quick to criticize the NOAA's alterations to the historic dataset, claiming they've manipulated temperature records to create a warming trend by literally *cooling* the past and *warming* the present. It seemed that early decades of temperature records were consistently adjusted downward, and the current, century-long temperature trend was higher than in the original dataset. Even some of the historic 'hottest days ever' lost their titles.

Disproving or proving climate change is outside the scope of and intent of this section. Rather it is to demonstrate that your analysis will only ver be as good as your data. Care, vigilance, and explicit domain expertise is often required to catch and gauge potential inaccuracies introduced into your data, particularly when big data is amassed over time from a multitude of sources.

Part of the data science umbrella of topics include the use of the scientific method. When experimenting, a control is usually introduced to reduce variations in the data due to anything except the features being measured. Knowing how to apply and what type of control to use typically requires human intervention, as we haven't found a way to get machines to learn to automatically account for that just yet. There are different types of controls, for instance negative control or randomization, which are used depending on the type of experiment being conducted and the data being gathered. Unfortunately, we only have one planet to test with and can't go backwards in time. Retrospective efforts like the NOAA's to polish up their data, will likely continue to be met with some level skepticism —at least until it gets a lot warmer.