Four-coloring the US counties

Mark McClure and Stan Wagon

Mark McClure is an Associate Professor of Mathematics at The University of North Carolina at Asheville.
Stan Wagon is a Professor of Mathematics at Macalester College in St. Paul, Minnesota.

Introduction

The celebrated four-color theorem states that at most four colors are required to color any map in such a way that any two adjacent regions receive different colors. Since the easiest way to clarify any theorem is with an example, we present a complicated but interesting one. Figure 1 shows a map of the contiguous United States. The entire collection of 3109 counties is colored according to the theorem. The state boundaries have been added to provide perspective but have no bearing on the coloring.

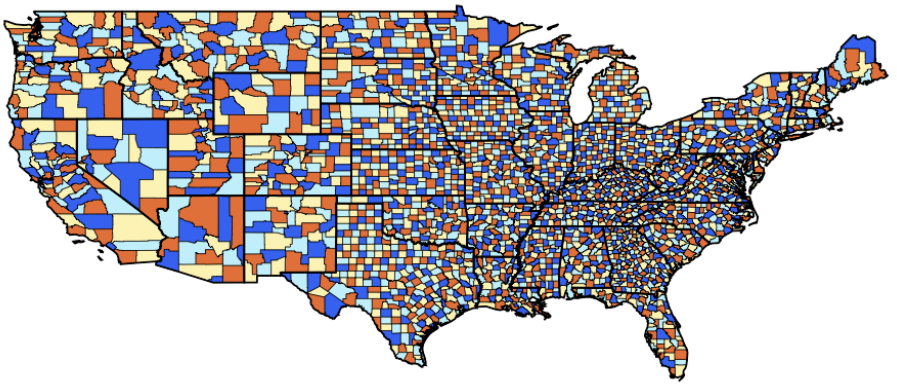

Figure 1: A four-colored map of the US counties

Of course, the four-color theorem has a precise, mathematical statement. For our purposes, it suffices to suppose that each region is a simple polygon. Two regions are considered adjacent if their intersection contains a line segment; a meeting at a single point doesn't count. Since many real world maps have disconnected regions, even if the regions are approximated by polygons, the four-color theorem is frequently not applicable. Furthermore, a glance through a typical atlas reveals that cartographers rarely use the minimal number of colors required to create a map. Nonetheless, it seems an interesting challenge to write a program that automates the procedure of coloring a map and apply it to a large, real world example. Figure 1 is the result of our efforts.

Generating and coloring the map

The four-color theorem was conjectured in 1852 by Francis Guthrie while he was trying to color a map of the English counties. The conjecture was published by Cayley in 1879 and an attempted proof by Kempe appeared later that year. That "proof" was shown to be incorrect by Heawood 11 years later and the subsequent history, culminating in the 1976 computer aided proof of the theorem, has been well documented. An outstanding reference on this history is Robin Wilson's *Four Colors Suffice* [1], but it is the early history that concerns us here.

As it turns out, the ideas in Kempe's 1879 paper provide the elements of an excellent four-color heuristic. The resulting algorithm has been implemented in *Mathematica* by the second author and the program accompanies his book *Mathematica in Action* [2]. Wagon knew that McClure had worked with complicated map data before and asked for an interesting, real world challenge. This was found courtesy of the US Census Bureau. The bureau maintains detailed records on political boundaries and their information is made freely available in the form of shapefiles ??? a proprietary file format designed to store complicated geographic information. Many such files, including the boundaries of US counties, are available via their website: http://www.census.gov/geo/www/cob/.

After we obtained a shapefile describing the US county boundaries, we read the data into *Mathematica* and computed the adjacency information for the counties. This adjacency information is most easily stored in a graph; each vertex represents a county and two vertices are connected by an edge precisely when the corresponding counties are adjacent. A graph that (with one simplification) represents this adjacency information is shown in figure 2.

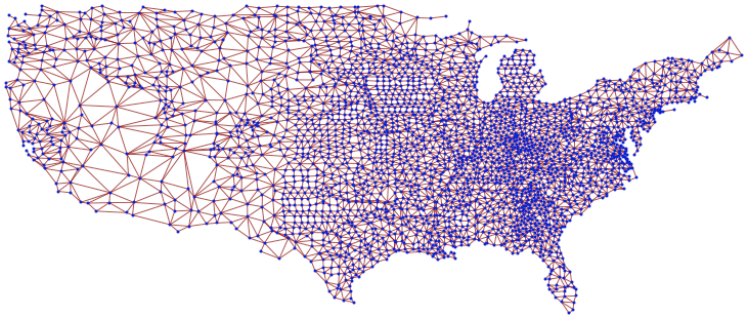

Figure 2: The planar adjacency graph for the us counties

The graph in figure 2 is not only (almost) an adjacency graph for the US counties but it is planar, i.e. there are no intersecting edges. This graph was created by using the centroids of the polygons as vertices, connecting adjacent vertices with straight lines, and then making adjustments to eliminate crossings. When the regions are simple polygons, it is possible to automate this process by using curved edges but that is more complicated [2, chap. 17]. Our regions, however, are not all connected. In fact, one county had to be represented as two distinct vertices. In particular, the two

dark blue regions with heavy outline and yellow vertices in figure 3 are actually two parts of the same county: St. Martin's Parish in Louisiana. Since the two parts are the same color, we ultimately end up with a four-coloring of the entire map. This is the one essential simplification in the adjacency graph. There are also two islands that are parts of disconnected counties with the other portions on the mainland in this picture.

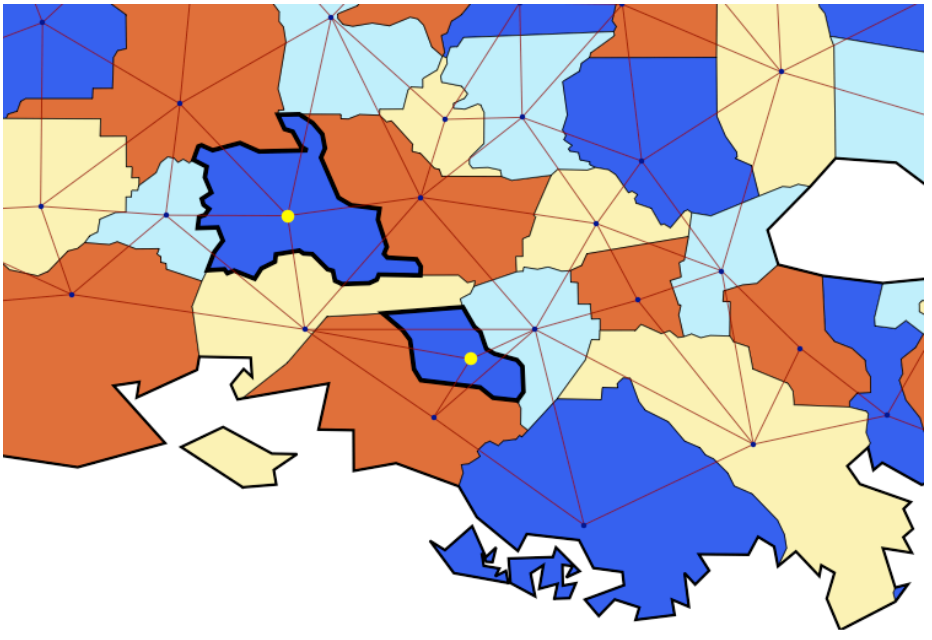

Figure 3: A zoom into St. Martin's parish

An interesting observation (pointed out to us by Bernard Lidicki and Robin Thomas) is that this graph has the property that vertices can be pulled off so that each has degree 4 or less in the graph that remains. This is the first step of Kempe's proof, where he allows for vertices of degree 5 or less. The fallacy in his proof arises in the degree-5 case, however, so the fact that this case does not arise means that the method of Kempe chains is guaranteed to work in this situation but these don't affect the adjacency information in any essential way.

Fun facts

We made a number of cute observations while playing with this data. For example, our graph separates into six connected components; the largest has 3102 vertices corresponding to 3101 counties, since St. Martin's parish is represented by two vertices. Long Island is a separate strand of four counties and there are four more island counties with no adjacent counties at all. These are, San Juan county in the Puget sound of Washington State, Richmond county off the coast of New Jersey, and Dukes county and Nantucket county both just south of Cape Code in Massachusetts.

Washoe county in Nevada has 13 neighbors, more than any other; this it might be called the friendliest county in the country. Next is another San Juan county in Utah with 11 neighbors. There are five counties with 10 neighbors.

There are 13 connected counties that have holes; they are not simply connected in

mathematical parlance. Some of the counties have more than one hole and there are 17 counties that actually form the holes. These are mostly city/counties in Virginia. Charlottesville, for example, is a county unto itself and is surrounded by Albemarle county.

Conclusion

Kempe's algorithm, though ultimately flawed, is quite practical. It can work with very large, real world data sets, *provided* that a genuinely planar adjacency graph can be constructed for the map. Because of the discrepancies between the idealized maps of the four-color theorem and real world maps, there can still be quite a lot of work to do. The entire procedure of going from ESRI shapefile to a planar graph represented in *Mathematica,* for example, was quite involved and somewhat tedious. A *Mathematica* notebook and all the required data is provided on our webpage: http://facstaff.unca.edu/mcmcclur/County4Color/.

References

1. Robin Wilson, *Four Colors Suffice: How the Map Problem Was Solved*. Princeton University Press, 2002.
2. Stan Wagon, *Mathematica in Action*, third edition, Springer, New York, 2009.