Exercise 8 Report: Geospatial and Temporal Visualization of West Nile Virus in California, Years 2006 - 2015

Motivation

The West Nile virus (WNV) is a mosquito-borne disease that was originally found in Africa. It was first detected in the eastern United States in1999; since then the virus has spread across the continental United States and is well established in most states, including California. In 2000, CDPH and other agencies expanded the statewide mosquito- borne virus surveillance program to enhance the state's ability to detect WNV. The California West Nile virus surveillance system includes human case detection, mosquito testing, dead bird testing, and monitoring of sentinel chickens. CDPH works with local vector control agencies and health departments to detect and monitor WNV activity and respond with enhanced mosquito control and public outreach to reduce WNV transmission risk. The visualization provides a set of geospatial and temporal idioms that illustrate how the virus spread across counties in the state of California. It is important to notice that the virus did not spread from areas where the number of cases were highest although that is what someone might think from looking at cumulative cases by county.

Problem Statement / Task Description

The dataset provided includes number of positive virus cases by county, year and week for the state of California. The original dataset has been augmented to include county code, and month of the year, both extracted directly from existing features. Additionally, county population data has been used to join with the original dataset. The county population figures have been obtained from the US census website.

The visualization is a story composed by five idioms. The story's main task is to show that the positive cases were reported each year according to a spatial pattern. Each year, the virus originated in the warmer months with the first cases being frequently observed in Kern county (in central California). In most years, the virus seems to then spread to nearby central counties with some periods of higher incidence in Southern California (LA and Orange Counties) and in interior counties in the Northern part of the state. The virus reported cases would then decrease to zero each year in the months of October, November and December, to then resume in the Spring of the following year.

Visualization

The visualization consists of five elements. The first is an interactive map of California depicting state county borders and reporting total cases (in absolute terms) for the entire ten year period between 2006 and 2015. This idiom uses an orange-red palette to color counties to reflect the amount of positive cases. Salience is defined by intensity of the color red. Counties are labeled by name and positive cases, with counties where cases were not reported being excluded from the visualization. The idiom clearly shows that Los Angeles county and Orange county were the two geographic areas with the highest cumulative number of cases; in general, Southern California seemed to have a higher number of positive cases.

A second map also depicting counties but interactively showing a monthly progression timeline is added as the second element in the story. The color palette is green-blue and is used to define the incidence (calculated as the number of cases as a percentage of the total population). Counties with green and less intense color saw a lower incidence, while counties color in darker shades of blue saw higher incidence. The viewer can move across time by means of the toolbar and will notice that while certain counties may have had a higher cumulative count of cases in the 10 year time period, the highest incidence actually occurred in the Northern counties of the state, with Glenn county reaching 0.004% in the August of 2015.

The third idiom is a stacked bar chart by year. Months are shown on the x-axis while count of positive cases on the y-axis. Each state's positive cases are stacked one upon the other, with each state's rectangle colored in unique color as shown in the legend accompanying the figure. The viewer can highlight and select each rectangle in order to learn more about incidence and cases for a particular county. While as many as twenty or more rectangles (each in a different color) are stacked on each bar, in most cases it is obvious that one or a few particular counties saw the highest levels of positive cases. Separability among counties is therefore not an issue, and it is improved when the viewer hovers over each rectangle and sees the relevant statistics for the county and time period become visible to the viewer.

In order to understand the disease one should focus on the time variable. Thus, the fourth idiom (a Gantt chart) displays the aggregate incidence for each year by months calculated as the incidence over total state population. This idiom highlights the fact that the end of the Summer and early Fall months were the ones when the highest numbers of cases were reported, although years 2009-2011 saw a lower degree of incidence during the same time periods. Color is used to show higher intensity as it is done in other visualizations that are part of the story, with deeper reds representing higher amounts. The viewer can hover over each rectangle to learn more about each month aggregate measures.

Finally, the last idiom in the story shows counties as circles in a 2D plot with years and population on the x and y axes respectively. Each county where cases were reported for a particular year are shown on a vertical axis. Circle size reflects the cumulative sum of cases reported during the entire year for the respective county, while color intensity reflects incidence. Thus, the viewer learns that although at first sight the most populous counties may seem to be where the virus originated (higher positive case counts), the incidence in these areas is actually lower that in smaller counties. Kern county, where the virus originated in most years, is the county displaying both an above average number of positive cases and incidence, indicating that it is one of the areas that suffered the most from the virus.

Conclusion

The visualization shows temporal and spatial progression of the West Nile virus. The story is made up of time idioms which sequentially provide more information to the viewer. The viewer is first provided with a cumulative sum of all cases for each county and might intuitively think that Orange County and Los Angeles County were the areas where the virus originated. However, we have shown that the virus seemed to spread from Kern county every year and that, as a percentage of the total population, Los Angeles County and Orange County were below average in terms of virus incidence. The counties where the virus was more problematic are the ones where the incidence was highest, i.e. the central interior counties of the state.