# West Nile virus incidents in California

Michael Gallaspy and Kevin Kannappan

## Motivation

The dataset consists of numbers of positive cases of West Nile virus (WNV) identified in California county and the week of the year in which the cases were observed. The time period covered is from 2008 to 2015. The kinds of questions we are interested in include:

- In which counties do cases occur, and how many cases are observed?
- Which counties consistently have the most cases? The least?
- What trends in cases can we observe over time? In space?
- What time of year do the most cases occur?

## Visualizations

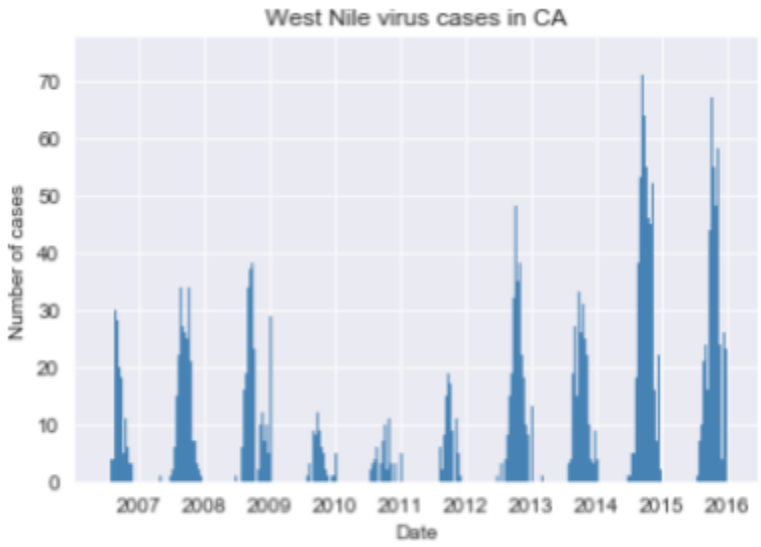


Figure 1: Total number of cases in CA over time.

Our first and simplest visualization (Figure 1) is a histogram meant to illustrate an answer to the last question. The visualization consists of horizontal bars drawn across an axis corresponding to time. The horizontal length channel of bars encodes how many cases occurred at that time, while the position channel of the bars encodes the time. The number of labels for the time axis is small compared to the number of time periods in the dataset, but the effectiveness of the position channel is enough to confidently conclude that most cases occur in the latter half of the year, with an evident peak around September or October. Figure 1 shows that this trend is very consistent for every year in the dataset. Moreover we can see that incidences of cases tend to be very localized in time, meaning that there is evidently one WNV “season” per year.

Our next visualization (Figure 2) dispenses with the week-by-week information and explores the total number of cases per year. The aggregation is justified since WNV seasons are strongly localized within one year. The visualization is an interactive choropleth illustrating

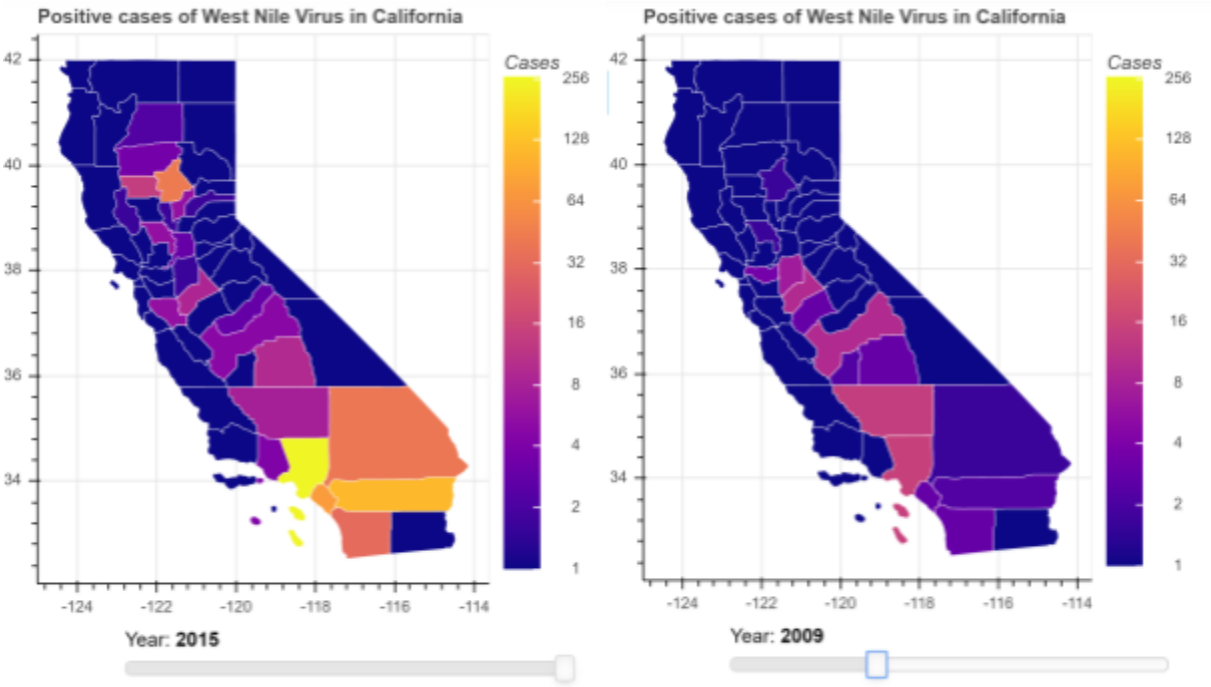


Figure 2: WNV case in California counties per year.

the number of WNV cases per year in California counties. The counties are plotted according to their actual geographic location, so the position channel strongly encodes the proximity of cases. The hue and luminance channels of the counties jointly encode the number of cases along a logarithmic scale. The chosen colormap is matplotlib's “plasma” colormap, a so-called “perceptually uniform” colormap featuring monotonically increasing luminance that permits values at all points in the colormap to be easily distinguished. The hue is easily distinguished even when simulating color blindness (Figure 3). A logarithmic scale was chosen in order to emphasize the difference between counties with no cases and counties with a small number of cases. The visualization includes an interactive slider bar to select the year for which cases should be visualized. The scale corresponding to the colormap remains fixed from year-to-year in order to facilitate comparison across years.

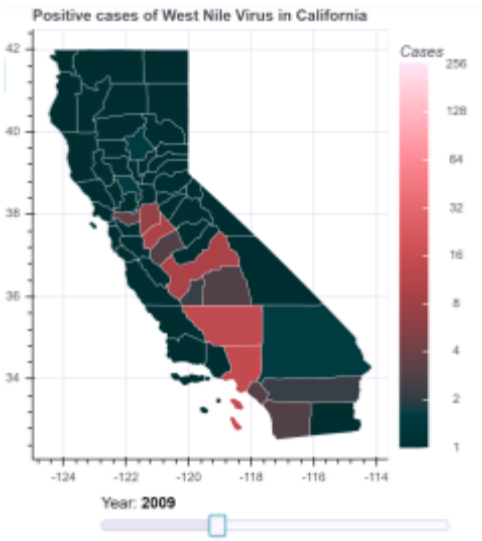


Figure 3: Choropleth visualization with simulated blue-blindness, which has the most severe impact on the visualization, demonstrating the efficacy of the colormap.

# Findings

Many conclusions can be drawn right away from both visualizations. There was a lull in the total number of cases for the 2009 through 2011 WNV seasons,

which is demonstrated quite effectively by using the slider of the choropleth visualization to move quickly through the years, as well as by comparing the height of the bars in the histogram. The choropleth also demonstrates clearly that the most active region is consistently southern California, particularly in LA and Kern counties in 2007 and 2008, and again LA county from 2011 onwards with a large number of cases also occurring in Orange, Riverside, Imperial and San Diego counties.

Moreover counties along the coast and the Nevada border consistently have few or no cases. In fact the incidence of cases in all years suggests that the virus travels approximately along the route of Interstate 5. This observation is easy to arrive at since the positions of the counties are accurately plotted, and suggests further research questions.

One potential problem with this visualization is that it doesn't account for the population of the counties. The population of California counties is highly variable, and one might naturally expect fewer incidences where there are fewer people. One question not clearly answered by the visualizations is whether the lack of cases in many counties can be attributed to a smaller population. One way to address this shortcoming would be to augment the dataset with populations of counties over time, and use the number of cases per capita to select a color.