Visualization of Space-Time Patterns of West Nile virus

Alan McConchie*

Department of Geography, University of British Columbia

Center for Advanced Research of Spatial Information (CARSI) Lab, Hunter College, City University of New York

ABSTRACT

A new technique for visualizing West Nile virus (WNV) activity is presented. Given a database of the space-time point locations of human WNV cases and a series of daily rasters representing modeled WNV risk areas, time slices are extracted from the rasters at the x, y location of the human onset. These time slices represent the "risk history" of a given geographic point. Juxtaposing risk histories in a variety of ways can reveal patterns of risk that are similar from the human's point of view, but dissimilar or disjointed in real space and time. Multiple linked views supplement the risk history display, allowing mapping of detected patterns back to real space and time. Identifying characteristic patterns of risk that precede human WNV cases will increase understanding and future prediction of WNV outbreaks.

1 INTRODUCTION

1.1 West Nile virus

Since its introduction into North America in 1999, West Nile virus (WNV) has spread rapidly across the continent, decimating bird populations and infecting thousands of people. It is primarily transmitted in a cycle between mosquitoes and birds, although infected mosquitoes can also spread the virus to humans. While it is only fatal to humans in rare cases, it can have serious health consequences. Among corvid species such as crows and jays, there is a very high mortality rate. Reports of observed dead birds by the public are used to model areas where WNV outbreaks are occurring and, as a result, where humans are at risk of catching the virus.

Some studies have theorized an amplification cycle occurs as mosquitoes feed on infected birds, thereby becoming infected and spreading the disease to yet more birds. As birds die off in large numbers, there may occur a "spillover effect" as infected mosquitoes switch to feeding on humans. Such an amplification cycle would result in an observable lag between peak bird deaths and human infections. However, there is also an inherent lag due to the incubation period in humans, where several days pass between human infection and the onset of symptoms. In general the lag between avian mortality and human onset is poorly characterized.

1.2 Task

This paper describes a visualization system designed to support exploratory data analysis in a model of WNV transmission. The goal is to assist domain experts in exploring the relationship between human WNV cases and the preceding space-time distribution of high-risk areas. The system should provide a large-scale overview of the pattern of WNV activity in the entire study area, and also facilitate the exploration of small-scale epidemiological differences in the relationship between clusters of dead birds ("risk areas") and points of human disease onset. My working assumption is that the lag time between elevated risk and human onset may vary significantly, either in different areas or at different times in the season. Understanding these differences in lag could lead to deeper understanding of how WNV may be influenced by underlying environmental conditions, and improve our techniques for modeling and predicting the spread of the disease.

The risk model used is DYCAST, the Dynamic Continuous-Area Space-Time system. [1] The DYCAST system identifies non-random clustering of dead bird reports by the public, and produces a continuous surface of human risk of WNV infection. For the purposes of this visualization system, the DYCAST model will be treated as a black box that produces a daily raster of WNV risk.

Currently the system has an ability to determine an overall relationship, averaged between all human cases, but no way to explore relationships within subsets of the data.

1.3 Data

The dataset for this problem consists of two parts, a series of daily rasters of areas at high-risk of human infection (henceforth called "risk"), and the location and date of onset for a number of human WNV cases.

1.3.1 Risk rasters

This system was tested using three years of daily DYCAST risk rasters (2004 through 2006) covering the entire state of California. There are approximately 540 rasters, because the model was not run during the winter, and the grid size is 0.5 miles, resulting in dimensions of 1421 x 1512 pixels. The value of each cell is between 0 and 1, representing decreasing likelihood that dead bird activity in that cell is caused by random chance (or increasing likelihood that the area is at risk to humans). For analysis purposes, this value is classified into a binary risk/no-risk value based on a fixed certainty cutoff.

1.3.2 Human cases

The rasters were compared against the space-time location of every human case of WNV for those three years. For the purposes of visualization, the human cases are placed at the center of the risk cell they fall within, and are then rendered as rectangular glyphs that fill at minimum the entire 1/2 mi square cell, although they can be scaled interactively to appear even larger. While the dataset has greater spatial accuracy for the human cases, using the coarser raster resolution is sufficient, since the relationship between the human case and the risk cell is the primary concern. In any case, the exact locations of human cases cannot be publicly released in order to protect patient privacy.

*e-mail: almccon@interchange.ubc.ca

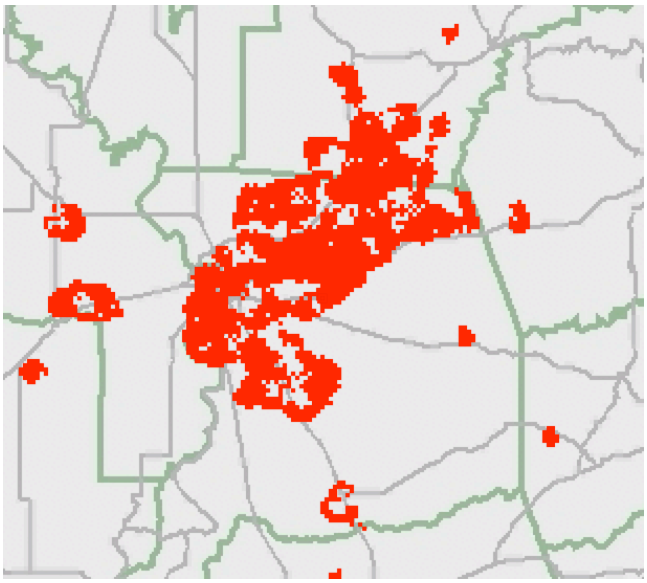

Figure 1. DYCAST risk raster

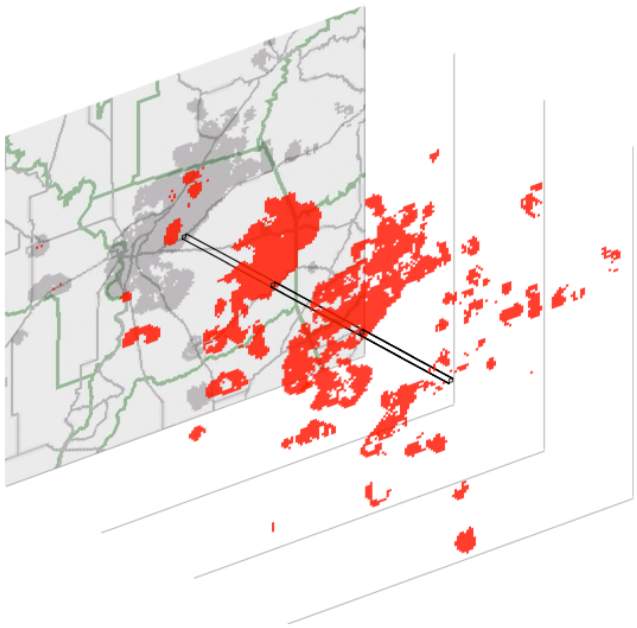

Figure 2. Extracting risk history

2 Previous Work

Current WNV modeling systems such as DYCAST or the Space-Time Scan Statistic (SaTScan) [2] use statistical techniques to assess overall rates of prediction. Recent work with DYCAST has attempted to characterize the overall best-fitting lag time, again using a statistical approach. None has attempted to apply statistical measures to spatial subsets of the data.

The previous extent of visualization of these risk models has been limited to producing daily maps, (Figure 1) or at best, animations of the rasters in sequence. Animations give a sense of the movement of the disease, but are difficult to use to view the structure of the beginning and end of the season simultaneously. Assessing the relationship between risk areas and human cases is also difficult, requiring the user to remember what the map looked like earlier in the animation. Interactive scrubbing of the animation helps somewhat, but is not a satisfactory solution.

3D visualization of the risk rasters is one possible solution. Mapping time as the third dimension would produce risk "clouds" which could be visualized along with the scattered 3D points of the human cases. The shapes of these risk clouds could be compared with one another, and with the location of human cases. 3D visualization would introduce problems of occlusion, however, causing objects behind or within the risk clouds to be obscured, and it would remain difficult to determine the exact history of a single human case out of the middle of a large cloud of risk. Furthermore, it would be difficult to identify areas with similar risk conditions if they are spaced far apart in the 3D space.

3 A new approach: Risk Histories

I propose a novel technique for extracting only the most important information from the risk raster data and presenting it spatially in an epidemiologically relevant layout. This derived "risk history" view transforms the display space, allowing similarities to be discovered that were previously obscured by spatial or temporal distance.

3.1 Human WNV risk histories

To extract meaning from the large amount of data inherent in the risk rasters, it is necessary to reduce the dimensionality of the problem. A derived variable or variables must be found for each human case that represents only the most relevant parts of the risk dataset. To this end, I make the assumption that only the risk at the point location of the human case is relevant to the process of disease transmission. Therefore, only a single raster cell needs to be considered from each raster file: the geographic raster cell in which the human case occurred. Extracting the value of this cell from the risk raster for each day, we obtain time series of binary risk/no-risk values associated with each human case. This time series represents the "risk history" of that point in space, both before and after human disease onset. (Figure 2)

These risk histories are self-contained slices from the risk dataset, representing the risk conditions that led up to human WNV infections. (Figure 3) From the point of view of each human case, the data in the risk history is a proxy for all other environmental contributing factors. Because risk histories are self-contained, we are not restricted to displaying them in their original spatial and temporal context. By placing these risk histories in different spatial layouts according to variables other than their space and time coordinates, it may be possible to find similarities between risk histories that would otherwise be far from one another in space and time.

Figure 3. Risk history. Red squares are days that were at risk, white were not at risk. The purple square is the date of human WNV onset.

This data display technique is inspired by van Wijk and van Selow's clustered calendar visualization, [3] where a time series of univariate data was split apart into day-long segments that were grouped into clusters based on a measure of similarity. These clusters were then color-coded and these colors used to fill in the appropriate days on a calendar display. By viewing a calendar with days color-coded according to similar daily patterns of some variable (such as changing usage of electricity over the course of a day), it is easy to see not only patterns that are expected (such as lower power usage on weekends) but also unusual events that occur irregularly but have a characteristic pattern when they do occur (such as power spikes). The strength of this technique lies in liberating each day's data from its natural temporal neighbors and juxtaposing it with other days that may be more similar according to the variable of interest. Extraction of risk histories from the WNV risk raster data represents an analogous approach, tailored to a different type of data.

3.2 Juxtaposed risk histories

Instead of automatically clustering the risk histories, my visualization allows the user to experiment interactively with different juxtapositions. Automated clustering would require definition of what makes two risk patterns similar with respect to their effect on human infections, a definition that is difficult to determine because the exact relationship between risk patterns and human disease onset is unknown. By contrast, my visualization focuses on the exploratory data analysis paradigm, allowing the user to theorize and experiment with different definitions of similarity (by visually comparing risk histories next to one another) and to evaluate which similarity metrics will best match the data.

3.2.1 Sorting

Spatial juxtaposition is accomplished by displaying all of the risk histories in a list view. The user can choose to sort this list by a variety of metrics that identify features in the data that may be epidemiologically relevant. In the current system is it possible to sort according to:

- position of human onset
- total number of days at risk.
- position of the first day at risk
- position of the last day at risk
- position of the average of all risk dates

The final sort sums the day numbers (the number of days since January 1) for each risk day, and divides by the total number of days at risk, producing the mean risk date. This can be thought of as the "center of gravity" of risk. Figure 4 shows the effects of each sort on a set of simulated risk histories.

3.2.2 Time-shifting

In addition to re-arranging the risk histories spatially by re-sorting the list, it is also possible to juxtapose them in different temporal ways. Currently one type of time-shift has been implemented: the ability to align the date of human onset in all risk histories. The resulting time values (shown in figure 5) are relative to each human case, not relative to real-world calendar time. Since the primary goal of this visualization tool is to explore what patterns of risk precede human cases, this shifted time scale may provide greater insight into similarities between risk histories. The time-shifted view can be used with any of the sort methods listed above, and sorts based on a temporal index into the risk sequence (such as the first, last and average risk dates) can be recalculated relative to the shifted time values.

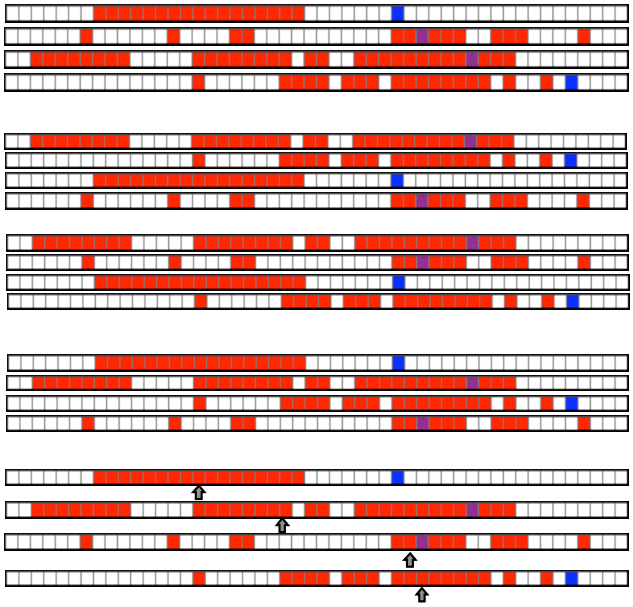

Figure 4. Risk histories, sorted according to a) human onset b) total days at risk c) first day at risk d) last day at risk, and e) average day at risk

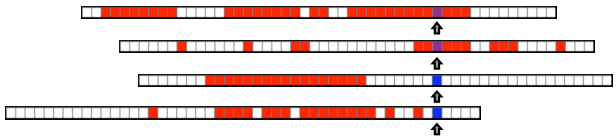

Figure 5. Risk histories, shifted to align human onset

3.2.3 Sorted risk histories as global-scale context

Sorting and time-shifting were designed to allow comparison between individual risk histories, but the sorted list also provides large scale context view of the entire WNV season. Figure 6 shows the risk histories in California for 2004, and Figure 7 shows them shifted to align human onsets. The human cases form a vertical line through the center of the chart. It is simple to observe that the majority of risk days occurred before human onset, suggesting that there is indeed a lag between peak bird deaths and human WNV onset. By examining the chart sorted according to the last day lit, and observing where the human case line intersects the edge of the sorted risk histories, we observe that for a large minority of human cases (those above the intersection point) risk activity ceased well before human onset.

Figure 8 displays the same sort applied to the 2004, 2005 and 2006 datasets in California. It is easy to tell that 2006 not only had many fewer human cases, but those that did occur were poorly predicted by the risk model (generally, risk was detected after human onset had already occurred). Between the 2004 and 2005 datasets, it can be seen that the 2005 graph appears more dense, suggesting that there was less variation in the shape of the risk histories. In 2005, the human onset line passes through the center of the risk history curve, showing that there were as many risk days that occurred after human onset as those that occurred before.

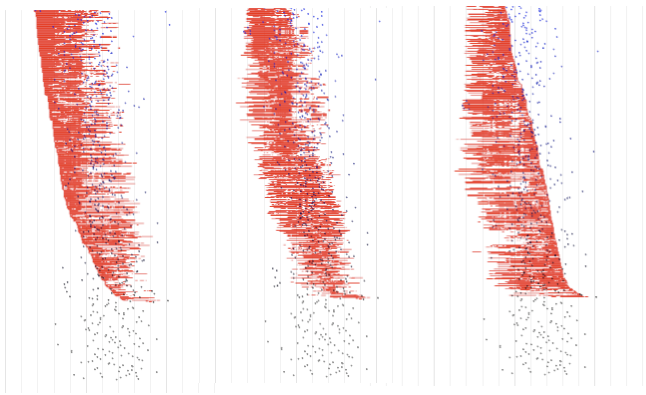

Figure 6. 2004 California dataset, 553 human cases. Sorted by: a) first, b) average, and c) last day at risk.

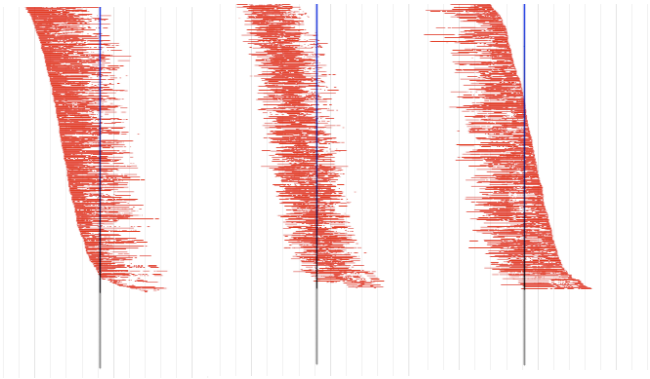

Figure 7. Time-shifted to align human onset

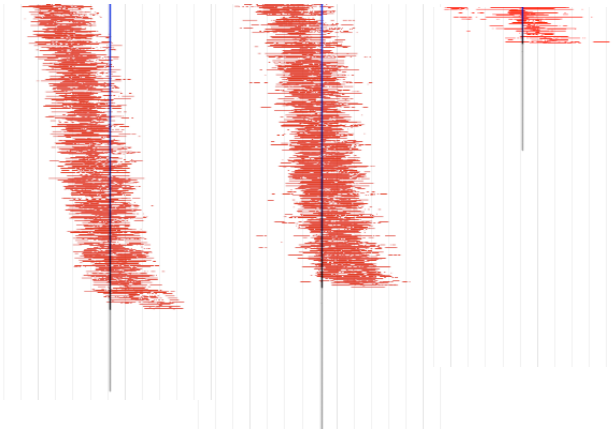

Figure 8. Comparison of different datasets. California WNV activity in: s) 2004, b) 2005, and c) 2006

3.3 Linked Views

The sorting and time-shifting tools allow display and exploration of risk histories freed from the constraints of their original spatial and temporal position. If any patterns are discovered in the sorted list view, however, there needs to be a way to transfer that knowledge back to real-world space and time. Also, during exploration of the transformed data space, a user may want to know the original time and location of one or more risk histories. To provide real-world context, the visualization system uses multiple linked views, supplementing the sorted list view with a map view, a timeline view, and combined space-time views. The axes of all the views are coordinated, such that panning or zooming in one view will adjust the other views accordingly. Also, interactions such as selection and mouseover highlighting are linked, so it is easy to identify interesting features in one view and see their location in all the other views.

3.3.1 Map View

The map is an essential tool for all geographical analysis, and is a natural view to use as an anchor and reference in the center of the visualization environment. A map provides a familiar starting point, and way for the user to bring their own geographical experience and intuition into the data exploration process; additionally, the map is a natural ending point for the data exploration, where observations are synthesized and transferred out of the visualization environment into other applications. For these reasons, the map view is given the same visual weight as the sort view, even though it contributes nothing unique or new in terms of visualization.

The map content comprises the point locations of the human WNV cases, displayed on a base map of the area of interest. Coloring of the human cases is linked to a color-ramp that is in turn linked to the order of human cases in the current risk history sort. Changes in the sort window automatically update the color encoding in the map view, making it possible for the user to see if risk histories that are similar in the sort view are spatially correlated in the map view.

3.3.2 Supplementary views

A third view will show the total number of raster cells that are at risk for each date, for the entire study area. This profile view affords the real-world temporal context for the sort view, just as the map view provides the real-world spatial context. The risk profile only displays aggregate risk, serving as a counterpoint to the de-aggregated sort view, and allowing the user to identify the location of the cursor in the overall continuum of the WNV outbreak.

Combining time and space context together, there are also X vs. time and a Y vs. time scatterplots, joined to the X and Y axes of the map view. These views represent a compromise attempt to display the three-dimensional nature of the risk “clouds” without requiring the complicated interactions that a true 3D display would require. Because the viewing angle is fixed along lines of latitude and longitude, these views can suffer from serious occlusion problems. Nonetheless, they enable a reasonable amount of space-time context without placing additional demands on the user’s attention.

Taking all these linked views together, the user has a wide range of techniques for visualizing the data, ranging from views that are deeply rooted in space-time coordinates to those that are completely unbound from the constraints of space-time proximity.

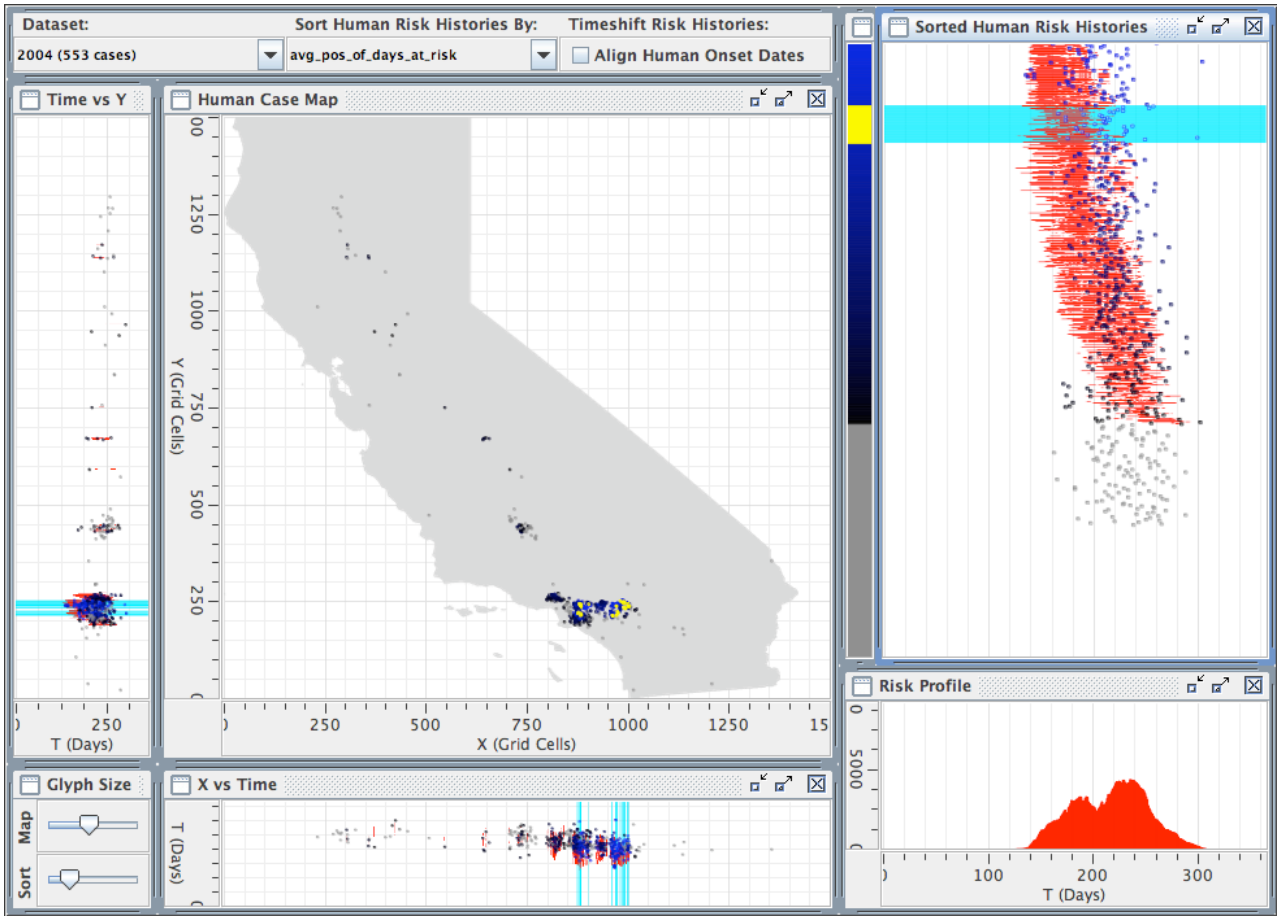

Figure 9. The visualization layout

4 Scenarios of use

4.1 Assessing prediction rates

Suppose that during or following the West Nile virus season, a public health biologist wants to assess how accurately their risk model predicted human cases before onset. Loading the dataset into the visualization system, clicking "align humans" and choosing "average day at risk", the biologist can quickly see which humans were best predicted by the risk model (humans where the average risk occurs several days or weeks before onset). Curious to see where the model did poorly, the biologist drags the selection tool in the sort window, selecting all of the human cases where the average risk occurred after human onset. (Figure 10) Checking the map, these cases are not clustered in any one place, but they do tend to fall on the fringes of large groups of human cases. However, the group of humans in the Northwest appears to have more selected cases than the other groups. Perhaps all the human cases in that area performed poorly, but some happened to fall just outside the selection in the sort window. To test this theory, the user can drag a selection around those human cases in the map view. (Figure 11) While those human cases appear to be weighted toward the bottom of the sort, many of them do appear fairly high up, implying that the risk modeling was not entirely bad.

4.2 Testing theories of virus transmission

Now suppose the biologist theorizes that some species of mosquitoes will only switch to feeding on humans when the bird population is depleted. This might manifest as an area where dead bird activity stops and then human cases occur several days later. To check this, the biologist sorts according to the last day at risk, while keeping the human cases aligned. At the top of the resulting sort (Figure 12) it is clear that a large number of human cases occurred several days after all risk activity in their raster cell ceased. Selecting these human cases in the sort window results in highlighted human cases across most of the map. However, there are very few highlighted cases in the Northwest cluster (the San Fernando Valley) while nearly all of the human cases in the Southeast cluster (centered on Riverside) are selected. The biologist then might want to check to see if there is a species of mosquito that is dominant in Riverside but absent in San Fernando.

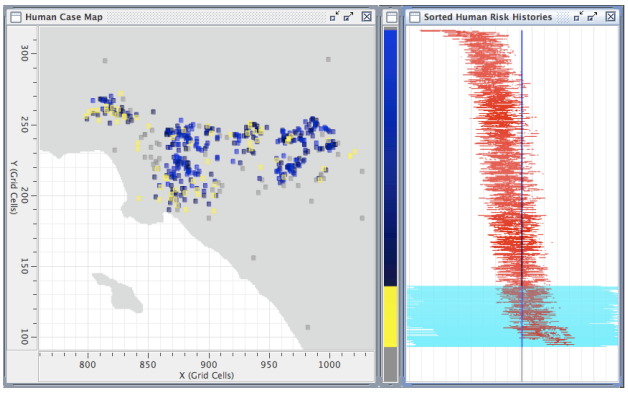

Figure 10.Sort view selection: poorly predicted humans

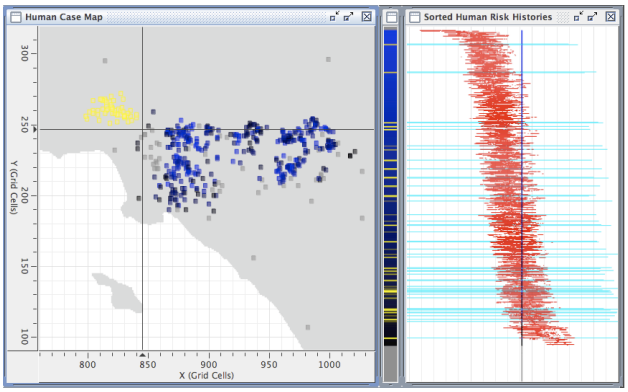

Figure 11.Map based selection

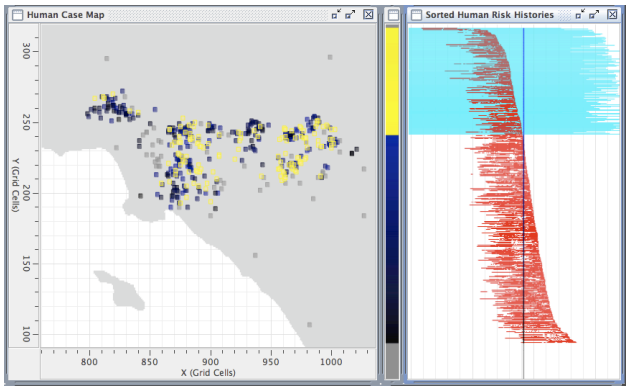

Figure 12.Sort view selection: humans occurring after risk ends

5 IMPLEMENTATION

This visualization environment was created in two parts—a suite of command-line utilities for creating and sorting the risk histories, and a visualization layer created in the Java-based Improvise visualization environment.

5.1 Command-line interface

The command-line utilities were written in Perl, using the PerlMagick bindings to ImageMagick to read and write image files, and custom written Perl module to achieve the rest of the functionality. In the first step, the original risk rasters are read once to extract the risk values for each human onset cell, which are then stored in an intermediate data structure or written to a file to save on future processing. In the second step, the user or the visualization application requests a sort, giving as parameters which dataset to sort, by what sort method, and whether to align the human cases. The sorting script can either produce an image of the sort result, or a sorted sequence of human case IDs. The software was developed to allow easy addition of more sort metrics in the future.

The original interface with the visualization layer involved creation of the sorted image in Perl, and then using that as a background upon which to draw interactive human case glyphs. However, this approach proved too inflexible. The alternative, loading each risk cell into Java appeared to be overkill with a likely negative effect on rendering time and responsiveness. Loading the risk histories as strips or lines with start and end points would also be unnecessarily complicated and perhaps not significantly faster, especially given that each risk history usually has multiple separate periods of risk. The compromise approach was to insert another step in the Perl processing chain above, and create a separate image file for each risk profile at the point the risk profiles are extracted from the risk rasters, not when they are sorted. The application needs only to request the sort order, and then these individual risk history images can be rendered as a single glyph in Java, creating clearer algorithms and reasonably quick rendering.

Because the number of available sorts is currently small, I also pre-calculated the results for each possible sort combination and stored the data as another dataset in the visualization environment. Since the dataset and the types of available sorts are not changing, pre-calculating these results saves time, and frees the application from its dependence on Perl. As a pure Java application, this software can be packaged with pre-processed datasets and made available to epidemiologists or other users using any operating system.

5.2 Improvise visualization

Improvise is an integrated visualization environment that supports interactive building of highly coordinated visualizations. I chose to use Improvise because of this emphasis on coordination and rapid prototyping, and also because it includes the ability to import geographic layers such as shapefiles, which may be necessary for future development of my visualization system. Improvise was created by Chris Weaver, who is currently based at Penn State's GeoVISTA center for geovisualization. [4]

Improvise is advertised as requiring little or no programming experience. However, I found the Improvise codeless development environment quite confusing at first. Improvise's potential for flexible, interactive development is based on a strict model of interaction, where views coordinate through Live Properties. A Live Property is a variable that is shared between multiple views, and which notifies all of the views that reference it whenever its value changes. The result is that changes happen instantaneously, whether during the construction of views, or during exploration of data in a finished visualization layout. Elaborate interactions can be built by combining Live Property-

based expressions with the display components provided by Improvise, although there are some common interactions that cannot be implemented. For example, the range of a viewport cannot be changed programmatically (or, as a result, through something like a zoom button)—it can only be changed through user interaction in that viewport or in another view linked to it. Creating new variables and displays can also be frustrating, as all objects require the creation of a Live Variable in order to perform any action, as well as clearly stated schemas during creation, which can be difficult to change later.

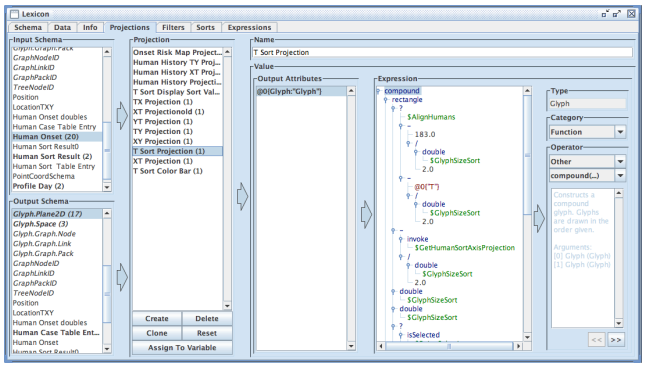

Figure 13.Creating a projection in Improvise

After coming to understand the Improvise style of development, I was impressed by the power of Improvise, and appreciated how easy it became to swiftly create or modify views. However, despite the strengths of Improvise, I am worried that further extension of my visualization environment will increasingly run into the inherent constraints of Improvise. Despite the ability to load shapefiles, Improvise has no other geographic functionality, which may limit future possibilities. I have not yet explored how easy it would be to add new Java components to Improvise, or to integrate existing Java components. If this is attempted, care must be taken to integrate correctly into the elaborate network of relationships inside Improvise.

6 Discussion

I showed an earlier version of the visualization environment to a Public Health Biologist who is familiar with the California WNV dataset. In addition to providing valuable suggestions for major and minor usability improvements, he found the coordinated views to be very useful, and described the profile view (the timeline view used for temporal context) as the most informative view. He felt that the sorted view was confusing at first and would require more time to become acquainted with, but he also found the s-curves striking and thought they had good potential for further study.

The profile view has been used frequently in our previous analyses of WNV activity, but this is the first time is has been rendered interactively, and with coordination with other views. Since no other infovis techniques (such as multiple linked views) had yet been applied to this dataset, I was in effect presented him with several "new" visualization techniques at once. It is perfectly reasonable that an enhanced version of an existing chart would be the most informative.

This raises the question of whether those who are studying WNV are ready for (or need) a brand-new visualization technique, when no other established techniques have been tried yet.

7 Conclusion and Future Work

I have presented a novel way of visualizing disease progression. By isolating only the raster data that directly relates to human onset and extracting that data from its original space-time position, it is possible to search for similarities in the data that might otherwise go undetected in fixed-space or fixed-time analyses. Future improvements to this technique will primarily involve improved sorting and similarity measures. User-defined sorts could be made more like a query language. For example, one might sort the risk histories according to the number of risk days within a five day window, ten days before human onset. As the relationship between risk and human onset becomes better understood, automated clustering algorithms could be applied to group the human cases into epidemiologically relevant groups.

However, it appears that the wise application of existing information visualization techniques may be the more pressing need in the WNV community. Additionally, understanding the large-scale progression of the disease may be more important than exploring the fine scale relationship between risk areas and human cases. To that end, this visualization system could be refocused to emphasize the full season overview, providing easier comparisons with previous years, both in the sort view and the profile view, and displaying the risk rasters in the map view.

Finally, to meet the needs of a large percentage of end users, namely the heads of mosquito control agencies, the map view will need to incorporate the shapefiles of individual counties and mosquito control districts, and allow exploration of the dataset based on these subsets. These administrative divisions have no ecological relevance to the spread of West Nile virus, but they have great significance when it comes to policy decisions regarding WNV remediation efforts.

References

[1] C.N. Theophilides, Ahearn, S.C., Binkowski, E.S., Paul, W.S. and Gibbs, K., First evidence of West Nile virus amplification and relationship to human infections. *International Journal of Geographic Information Science,* **20**, pp. 103-115. 2006

[2] M. Kulldorff, A spatial scan statistic., *Communications in Statistics: Theory and Methods*, **26**, pp. 1481–1496. 1997

[3] J. J. van Wijk and E. R. van Selow, Cluster and calendar based visualization of time series data. In *Proc. IEEE Symposium on Information Visualization*, pp. 4-9, 1999.

[4] C. Weaver. "Building Highly-Coordinated Visualizations In Improvise". *Proceedings of the IEEE Symposium on Information Visualization* 2004