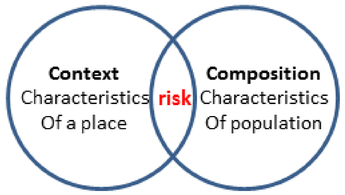
Uncovering the hidden patterns in complex epidemiological datasets

Abhishek Kala*, Samuel F. Atkinson, Chetan Tiwari

University of North Texas, Advanced Environmental Research Institute

Purpose

- This study postulates that underlying environmental conditions and a susceptible population's socio-economic status should be explored simultaneously to adequately understand a vector borne disease infection risk.
- Here we focus on West Nile Virus (WNV), a mosquito borne pathogen, as a case study for spatial data visualization of environmental characteristics of a vector's habitat (contextual factors) alongside human demographic composition (compositional factors) for understanding potential public health risks of infectious disease.
- Multiple efforts have attempted to predict WNV environmental risk, while others have documented factors related to human vulnerability to the disease. However, analytical modeling that combines the two is difficult due to the number of potential explanatory variables, varying spatial resolutions of available data, and differing research questions that drove initial data collection.
- We propose that the use of geovisualization tools may provide a glimpse into the large number of potential variables influencing the disease and help distill them into a smaller number that might reveal hidden and unknown patterns.
- This geovisual look at the data might then guide development of analytical models that can combine environmental and socio-economic data.



Challenges with multivariate analysis

- Although compositional and contextual factors are individually modeled using a variety of statistical models, they are usually not integrated due to the high dimensionality of associated datasets. Thus the potential patterns lurking from such complex multivariate datasets can be difficult to identify.
- Analysts need to find interesting patterns out of huge possible combinations from these complex datasets. Even in a selected subset of dataset it is still a challenge to discover the hidden relationships among those variables, as potential patterns may take various forms, linear or non-linear, spatial or non-spatial. The attribution of meaning to discovered patterns therefore requires input from experts who have the domain knowledge.
- It is also important that analysis of such complex datasets is performed using methods that are computationally efficient. Tools developed for geovisualization can be used to support multivariate analysis of geospatial data.

Human population characteristics (demographic composition)		Mosquito habitat characteristics (environmental context)	
Factors studied (reference)	**Relation to WNV risk**	**Factors studied (reference)**	**Relation to WNV risk**
Old age (Jean et al., 2007; Ruiz et al., 2004)	Weakened immune system	Stream, Vegetation, Road (Cooke, Grala & Wallis, 2006; Kala et al., 2017)	Sites for breeding and resting.
Male sex (Murray et al., 2006)	Social history or lifestyle.	Temperature (Kala et al., 2017; Wimberly et al., 2008)	Increases growth rate of vector, decreases egg development cycle and shortens extrinsic incubation period of vector.
Race/Ethnicity (Ruiz et al., 2004)	Increased risk from behaviors linked to their lifestyle.	Surface slope (Ozdenerol, Bialkowska-Jelinska & Taff, 2008)	Water stagnation creating mosquito breeding ground.
Income (Ruiz et al., 2004)	Increased risk from behaviors linked to their lifestyle.	Cultivated land, Developed land (Kilpatrick, 2011)	Preferred natural ground pools in cultivated land and warmer micro-climates in developed lands.

Variables related to susceptible human population characteristics (composition) and vector habitat characteristics (context) utilized in this study

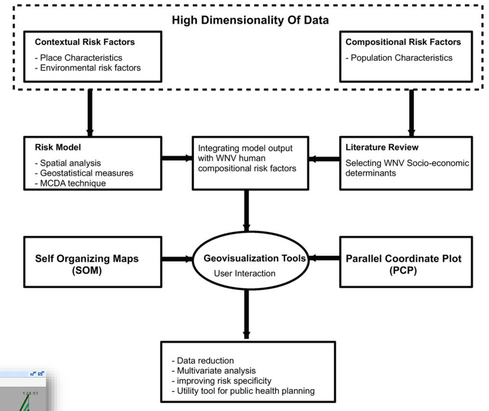
Methodology

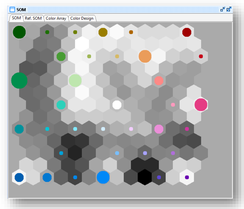
- This study utilized a spatially explicit exploratory approach for identifying the interaction between different environmental (mosquito habitat) and socio-economic (human demographic) processes occurring in each census tract of our study area in the state of California.
- The analysis was facilitated with SomVis, originally an open source Java application, that has now been ported to a web-based service (zillioninfo.com). SomVis was/is an integrated software tool consisting of three interactively linked visualizations that can help focus attention on patterns of similarity in complex data sets.
- The three visualizations used were:

1. **Self organizing maps, SOM** (Kohonen, 2001) to perform multivariate analysis, dimensional reduction, and data reduction;
2. **Parallel coordinate plot, PCP** (Inselberg, 2002) to visualize the multivariate patterns with display;
3. **Geographic mapping (GeoMap)** to highlight clusters of specific interrelationships.

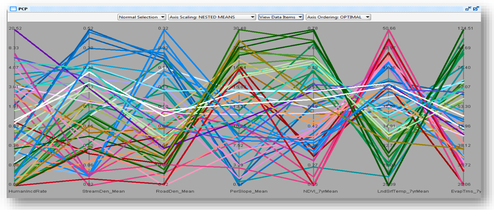
- These tools help to display the high-dimensional datasets, search for hidden relations among the complex set of variables and transform them into a 2-D pattern recognition problem.

WNV risk and susceptibility geovisualization modeling framework



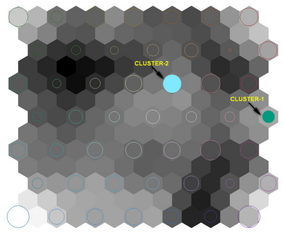


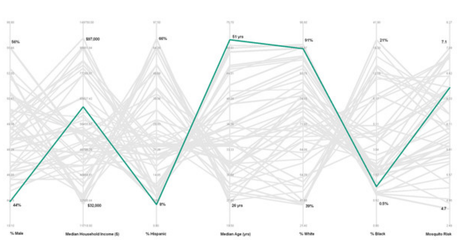
SELF ORGANIZING MAP

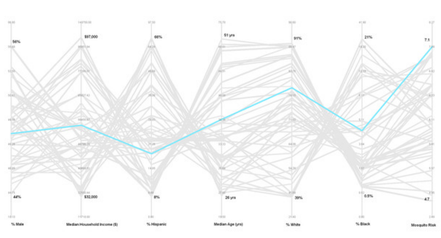


PARALLEL COORDINATE PLOT

Results



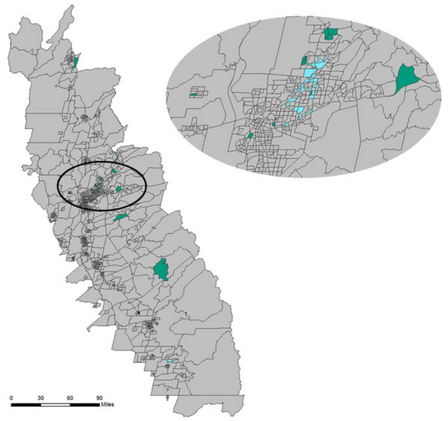




- Environmental and socio-economic data were extracted for all census tracts within the study area and considered simultaneously with SOM analyses. The resultant SOM identified 49 distinct nodes of census tracts
- Each SOM node (indicated with colored circles) represents a cluster of census tracts that are most similar in terms of all seven variables. The diameter of each node represents the number of census tracts in the node.
- To illustrate how geovisualization can be used by public health planners, two specific nodes are highlighted for discussion.

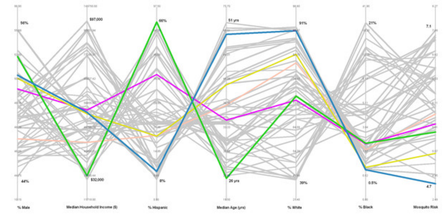
1. First, the cluster that contains census tracts with the highest median age is highlighted (labeled as cluster 1 and green in color), and is of interest because it is a variable that has been described as representative of the most vulnerable population (the elderly) to WNV health issues (e.g., Campbell et al., 2002).
2. Second, the cluster that contains census tracts with the highest environmental WNV risk (mosquito habitat) based on the GWR model is highlighted (labeled as cluster 2 and blue in color) because of the statistically significant relationship to WNV infected dead bird count (*reference-2*) .

- Once SOM nodes are defined, a PCP was used to explore the interaction between different risk factors. The PCP shows seven vertical axes representing each of the variables under consideration, and 49 polylines representing clusters of census tracts that are most similar to each other for those seven parameters.



Discussion and Conclusion

- We have shown a few examples of how geovisualization could be used by public health planners to better understand and respond to an infectious disease outbreak.
- Examples of several interesting patterns were revealed. For example, the cluster that had census tracts with the highest average mosquito habitat risk only had mid-level median age levels. Had there been a cluster that had both the highest mosquito habitat risk and the highest median age, public health planners might choose more intense intervention measures in those census tracts.
- Another interesting pattern uncovered was that census tracts in the county that had the highest reported incidence of WNV had relatively low mosquito habitat risk. This might lead to a speculation that demographic and socio-economic parameters should be weighted more importantly than mosquito habitat risk when developing public health plans. Likewise, this pattern might suggest other factors like poor links between modeled mosquito habitat risk and WNV risk in areas outside the training set data or spatial biases in recording effort operating differently at the county level and the census tract level could be at play. Focusing on those ideas through geovisualization may reveal other unknown patterns.
- Our study highlights the potential of combining geovisualization tools along with GIS to detect and analyze different hidden patterns within the complex multivariate data.
- The coupling of these techniques provides an interesting platform for analyzing larger datasets by integrating it into a spatially-explicit disease model or by using it for near-real time disease monitoring.
- This user interactive data exploration platform helps identify clusters of complex high dimensional datasets while preserving the topological relationships between data vectors.
- These techniques can reveal how the interactions between the contextual and compositional variables vary locally across geographic space.
- They also help to conduct exploratory analyses by identifying the causes and correlates of health problems and by identifying the populations at risk.
- Such an integrated approach facilitates the analysis of complex data and supports reasoning about the underlying spatial processes that result in differential risks.



References

1. Kala AK, Atkinson SF, Tiwari C. 2020. Exploring the socio-economic and environmental components of infectious diseases using multivariate geovisualization: West Nile Virus. *PeerJ* 8:e9577 https://doi.org/10.7717/peerj.9577
2. Kala AK, Tiwari C, Mikler AR, Atkinson SF. 2017. A comparison of least squares regression and geographically weighted regression modeling of West Nile virus risk based on environmental parameters. *PeerJ* 5:e3070 https://doi.org/10.7717/peerj.3070