# DataSpeak—Clusters, Maps, and Hotspots: Small Area Analysis in Maternal and Child Health

August 12, 2015

## Questions and Answers

**Q: Can you recommend an introductory textbook or website for someone interested in learning more about spatial analysis?**

A: (Russ Kirby) There are several good resources to learn more about spatial analysis. ESRI has published a number of texts, including *GIS, Spatial Analysis, and Modeling*, Maguire, Batty and Goodchild, Eds (2005). A brand new text that looks fairly good is *Spatial Analysis: Statistics, Visualization, and Computational Methods*, by Tonny Oyana and Florence Margai (CRC Press, 2015). But for those just beginning to work with GIS, there are several exercises that introduce aspects of spatial analysis with health data in the *GIS Tutorial for Health, 4th Ed,* Kristen Kurland and Wilpen Gorr (ESRI); this does require access to ArcGIS 10.

**Q: Can you speak to Birth Defects Programs about what you focus on for Neural Tube defects and CCHDs since both of these are pretty spotty in terms of where they show up in a state? Our numbers for counties are often less than five.**

A: (Michael Kramer) Pooling across years and/or aggregating from small spatial units (e.g. tracts/zip codes) to larger (e.g. counties or PH districts) may be necessary for extremely rare events. The methods described in the webinar including Empirical Bayesian small area estimation will make the most of the data available, but there's no denying that we are limited in saying much with extremely few events. Another version of these methods is to map something like the SMR: what is the observed case count in an area compared to that expected from national or regional data. It's akin to the EB rate estimation except expressed as a ratio. More complex methods could be (and have been) considered including multivariate Bayesian estimation. Here I use "multivariate" in its proper sense meaning there are multiple OUTCOMES, rather than a common misuse which is as a substitute for multivariable. So the idea is that there may be a class of congenital anomalies each of which is rare, but are thought to have some commonality in their etiology (stage of development, exposures, genetic traits, etc.) Bayesian modeling can take the information for multiple extremely rare events and ask whether they co-occur (in a possibly slightly less rare manner) in space. The bottom line is that with rare events distinguishing between "chance" or spatially random distribution vs a shared or spatially-represented exposure is challenging. We should certainly be cautious infusing too much certainty about a spatial cluster when based on very few events.

(Russ Kirby) This is a perennial issue, and not related only to rare birth defects. Many agencies managing confidential data are concerned about privacy issues related to small numbers of events mapped across small areas. Some have policies that limit access to the actual number of events when it is less than 5, while others will allow the researcher or analyst to use all of the data, but put procedures in place to ensure that individual cases cannot be identified in any map output viewable by the public. Often one must pool data across years—this is not a bad idea even when numbers are larger as random variability is reduced with 3-, 5-, 7-year temporal aggregation.

If the researcher has access to geocoded lat-long coordinates, one solution is the create maps using that data rather than the more traditional choropleth maps. Isopleth maps interpolate spatial surfaces from the distribution of points and their characteristics; common examples include topographic maps and weather maps (there is a great example today (8/28/15) for the state of Florida showing the projected rainfall associated with tropical storm Erika).

Another option is spatial aggregation. Establish a minimum number of events that must be present for an area to be mapped, and combine adjacent areas until the minimum is achieved. There is a very nifty tool for this, developed by Tom Talbot and colleagues at the New York State Department of Health, available at http://www.albany.edu/faculty/ttalbot/GAT/.

**Q: Can you ask the speakers if they can give us references to papers using these techniques in the area of birth defects?**

A: (Russ Kirby) There are a number of examples of GIS used in birth defects research. Here are a few:

- Delmelle EM, Cassell CH, Dony C, Radcliff E, Tanner JP, Siffel C, Kirby RS. Modeling travel impedance to medical care for children with birth defects using Geographic Information Systems. *Birth Defects Res A Clin Mol Teratol.* 2013 Oct;97(10):673-84. doi: 10.1002/bdra.23168
- Yazdy MM, Werler MM, Feldkamp ML, Shaw GM, Mosley BS, Vieira VM; National Birth Defects Prevention Study. Spatial analysis of gastroschisis in the National Birth Defects Prevention Study. *Birth Defects Res A Clin Mol Teratol.* 2015 Jun;103(6):544-53. doi: 10.1002/bdra.23375
- Case AP1, Canfield MA, Barnett A, Raimondo P, Drummond-Borg M, Livingston J, Kowalik J. Proximity of pediatric genetic services to children with birth defects in Texas.
- *Birth Defects Res A Clin Mol Teratol.* 2008 Nov;82(11):795-8. doi: 10.1002/bdra.20515.

The issues involved in mapping birth defects data are generally similar to other perinatal outcomes, so it's worthwhile also looking at studies on low birth weight, preterm birth, and infant mortality.

(Michael Kramer) There are dozens. I think the best approach to see a snapshot of literature is to go to www.scholar.google.com and search "gis birth defects"—I have many papers in my personal library and they all show up in this search plus many others I didn't have.

**Q: Could you please speak to natural breaks categorization?**

A: (Michael Kramer) Jenks Natural Breaks is the default categorical scheme applied to continuous data in ArcGIS when creating a choropleth map. There is accessible background on it here: https://en.wikipedia.org/wiki/Jenks_natural_breaks_optimization. In sum it is a method to identify the natural groupings or clusters of values along a continuous scale and thus make what are hopefully "meaningful" cutpoints for symbolizing in various colors on a map. I rarely (never?) find it useful! The reason is because there rarely is anything truly "natural" about local minima/maxima in a given dataset and the algorithm often makes what look like extremely arbitrary choices. The ideal symbolization scheme displays something interesting about the data and is amenable to transparent disclosure of the assumptions built into the cutpoints. I think if Natural Breaks meets the first criteria it is most by chance, and it rarely meets the second criteria. I often use quantile cutpoints (e.g., 5 quintiles, 4 quartiles, etc.) because the "arbitrary" component of the cutpoints can be understood. However if you have a very narrow distribution then quantiles can still give the misinterpretation of the extremes (e.g. 1st vs 5th quintile) being very different when in fact they might be very similar. So it is not perfect. There is a literature on cartography and visualization that tries to manage this tension in health data. Here is one paper that recommends combining the color legend of the choropleth with the distribution (in this case cumulative distribution although a histogram is another possibility) of the underlying data: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2760860/ Such an approach provides the reader with knowledge about the full underlying distribution as well as the location (and hopefully rationale) for the cutpoints.

(Russ Kirby) Dr. Kramer has explained this fairly well. It is a technique developed many years ago by Dr. George Jenks, a well-known professor of cartography at University of Kansas. It is a tool built into many GIS software applications, including ArcGIS.

**Q: Hotspot/cluster analysis seems an extremely smart and efficient way to deliver service. Are there significant or hidden barriers that are preventing states/federal programs from administering/evaluating all MCH programs (WIC, Title V, etc.) this way?**

A: (Russ Kirby) These are indeed useful techniques, but as with any scientific approach, the real world often gets in the way. Firstly, one needs data geocoded to points or small areas (tracts, block groups, etc.) from the program, either for participants or based on claims/use of services. Secondly, the program managers must be willing to allow the analyses to be done, and to realize that the process will inevitably reveal issues in data collection and management previously not known and that the exercise should be thought of in the context of program quality improvement. Thirdly, these analyses provide only one form of evidence that should be considered in many decisions concerning location/allocation of services.

An anecdotal example tells the tale. In a southern city (location withheld to protect the innocents) it was found that many women did not start prenatal care till late in their pregnancies. Birth certificate data were mapped across small areas, and the location of public and private prenatal clinics identified on the same map. One area stood out as having both a high proportion of mothers with late prenatal care, and no clinics nearby. The health department built a new prenatal clinic close to those with late prenatal care. Three years later, there was essentially no change in the pattern when the data were mapped again. The lesson is, the map shows the pattern, but not all of the factors associated with the processes that created the pattern. Mixed methods research including both quantitative and qualitative approaches is necessary to garner all the necessary findings to make reasoned decisions concerning location and management of MCH programs and services.

(Tom Stopka) ArcGIS is an expensive software. Some local, county and state agencies may not have ample budgets to cover the costs. There are examples of GIS freeware (Quantum GIS [QGIS]; geoda, R-spatial packages) that may meet local needs. Ample training and expertise is also needed. For one to conduct and understand hotspot cluster analyses, and other spatiotemporal analyses, one would typically need to take at least a couple of GIS and spatial analysis courses to thoroughly understand the details. Having a good mentor along the way is also key. Not all local, county, state, and national agencies have such expertise in house. There are, however, a growing number of GIS and spatial analytical experts around the globe!

**Q: What packages are available in Stata for small area estimation?**

A: (Michael Kramer) I am not a Stata user, but a Google search for stata spatial packages gave me this short list: spmap; shp2dta; mif2dta; spatcorr; spatreg; spatgsa; spatlsa. I think the spatcorr and spmap are probably good starting places with respect to content covered in the webinar. Other packages assist with more complex spatial regression methods.

**Q: Could Dr. Kramer please elaborate on the specific R packages that may be used to conduct empirical Bayes spatial estimation? How does the empirical Bayes spatial regression differ from a spatial error model (a multilevel model with a spatial error random effect)?**

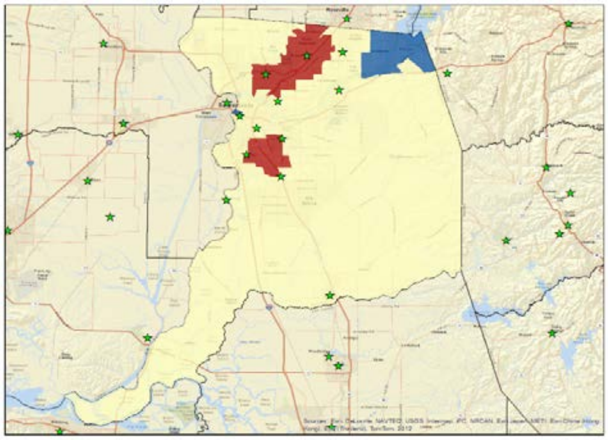
A: (Michael Kramer) Probably the best one-stop shopping for spatial packages in R is the Spatial Task View on the CRAN website (https://cran.r-project.org/web/views/Spatial.html). There are literally dozens of packages reviewed there, but a good starting place for importing and managing spatial data are these: sp, rgeos, maptools, and rgdal. For creating weights matrices and cluster analysis consider spdep, DCluster, SpatialEpi. For more complex spatial regression consider spgwr, GWModel, spdep, CARBayes, McSpatial. A very valuable into book for spatial analysis in R is Rober Bivand, et al's, Applied Spatial Data Analysis with R from Springer Press.

**Q: With the overall hotspot analysis, it didn't look like there were big problems between location and underserved. Are you sure the new enrollees were for that WIC new or new to WIC altogether? Can you show us where on the map the center was entered?**

A: (Tom Stopka) Good question. Overall, there appeared to be good overlap with WIC Centers in a number of WIC unmet need hotspots. So why do those areas remain as hotspots? There may be several reasons. Perhaps the WIC hours of service, the WIC locations, and the WIC staff were insufficient to cover the overall burden of unmet need. In other words, demand for services may have been greater than supply. This may have indicated that additional hours of service, additional WIC staff, and additional WIC Centers were needed to meet unmet needs. It might also indicate that some WIC services could be better placed in a location within or more proximal to a hotspot for unmet needs.

We did not know how many of the new enrollees at the new WIC Center were new to WIC altogether. It is possible that some of the new enrollees at the newly opened WIC Center previously received services at another WIC Center. We had hoped to tease this apart in subsequent analyses but I then moved to the East Coast. It is possible that some of the new enrollees at the new WIC center found the new WIC center to be more conveniently located compared to other WIC centers they might have attended previously. This could support more frequent and consistent attendance at WIC support and learning sessions for mothers who otherwise might have had to travel unreasonable distances to receive WIC services.

The new WIC Center was opened at the location on the map highlighted below:



Sources: Esri, DeLorme, NAVTEQ, USGS, Intermap, iPC, NRCAN, Esri Japan, METI, Esri China (Hong Kong), Esri (Thailand), TomTom, 2012

**Q:** **Does your AJPH article provide sufficient detail for someone to replicate this analysis?**

A: (Tom Stopka) The AJPH article goes into great detail on the five-step geostatistical approach we used to conduct hotspot cluster analyses. We point out the different tools and steps we used in ArcGIS along the way. For someone with strong GIS and spatial analytical skills, I believe the article provides details that can lead to replicated analyses. That said, due to space limitations in AJPH, we needed to be succinct. Additional instructions about some of the tools and parameters may help guide spatial analysts through the process even more efficiently. I would be happy to provide additional support and guidance should you have interest in replicating the analyses in your local area. Esri's ArcGIS Online and Spatial Resources can also connect you to a number of tutorials for spatial analyses and hotspot cluster analyses that you might find useful.

**Q: In your research, you matched WIC data with birth data. How were you able to obtain the address level data for both datasets? I ask this because many government and health care agencies have HIPAA regulations that restrict them from sharing sensitive PHI such as address.**

A: (Tom Stopka) At the time I conducted the analyses, I was working at the California Department of Public Health with the Office of Maternal, Child, and Adolescent Health and the State WIC Program so we had access to the data within CDPH.

**Q: How did you link the birth data with the WIC information?**

A: (Tom Stopka) Please take a look at our AJPH article for details.

Stopka TJ, Krawczyk C, Gradziel P, Geraghty EM. Use of spatial epidemiology and hotspot analysis to target women eligible for prenatal women, infants, and children services. *Am J Public Health.* 2014 Feb; 104 Suppl 1:S183-9. doi: 10.2105/AJPH.2013.301769. Epub 2013 Dec 19. PubMed PMID: 24354821.

If you still have questions after reading the linkage information presented in the paper, please let me know.

**Q: I have no experience with this type of mapping, but It seems to me that the determination of the "center" from which the polygon will start "spreading" could make a difference in how far we should be cover (5, 10, 25km). Is that impression correct? If so, how is the "center" chosen/defined?**

A: (Tom Stopka) The geocentroid, or the geographic center of a polygon (in my example, the census tract) is calculated in ArcGIS. It is important to keep in mind that the local mean value for the variable of interest (in my example, density of WIC eligible women [i.e., unmet need]) is calculated for each census tract and its neighboring census tracts within the exemplar sphere of spatial influence. So the example I showed with the expanding concentric circles would be happening over and over again with each census tract, with ArcGIS calculating a local mean value for WIC unmet need in each census tract and its neighboring census tracts. We then compare each local mean for unmet need to the global mean for WIC unmet need in all census tracts in the area of focus (e.g., all census tracts in CA). I realize this can be hard to follow without talking through the details and/or using a visual. I would be glad to talk with you further to explain these details, if it would be useful.

## About DataSpeak

The Maternal and Child Health Bureau's DataSpeak webinar series is dedicated to the goal of helping MCH practitioners on the Federal, State, and local levels to improve their capacity to gather, analyze, and use data for planning and policymaking. DataSpeak is funded by the Maternal and Child Health

Bureau's Office of Epidemiology and Research under the supervision of Gopal Singh, PhD. This question and answer sheet was created by moderator Sarah Lifsey, MPP.

*August 28, 2015*