Maps, Rates, and Fuzzy Mountains: Generating Meaningful Risk Maps

Tamara Jimenez, Armin R. Mikler, Marty O'Neill II
Department of Computer Science & Engineering
University of North Texas
Denton, USA
{*tamara.jimenez, mikler, martyo*} *@unt.edu*

Chetan Tiwari
Department of Geography
University of North Texas
Denton, USA
chetan.tiwari@unt.edu

***Abstract*—Creating meaningful maps that represent rates and risks in the population is a challenge. Risk rates are often computed for small area units such as census entities that may contain small population counts. Due to the unstable nature of such estimates, maps produced using such data are likely to misrepresent the risk of an event's occurrence over geographic space. This paper introduces two systems based on distinct approaches to generate risk maps that are not biased by the underlying population distribution of a given region: the adaptive kernel density estimation procedure implemented in WebDMAP and the population-uniform partitioning method included in UPAS. Comparison of both systems shows that qualitatively similar results can be obtained by both approaches.**

***Keywords*-risk maps; risk representation; epidemiology; disease maps; public health;**

I. Introduction

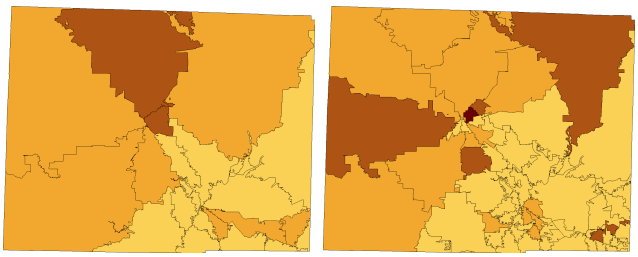
Risk maps are maps designed to visualize the spatial risk distribution of specific events, such as disease incidence (cancer, influenza) and environmental hazards (exposure to chemicals, biological agents). Once a meaningful risk map has been created, it can be used as a tool for public health disaster preparedness. One of the earliest known risk maps dates back to 1854 when a cholera epidemic broke out in London. The English physician John Snow marked all cholera cases on a map and successfully identified the source of the epidemic as a particular water pump [1].

Since the perception of risk depends on the number of individuals in a region, the generation of risk maps is a function of the proportion of at-risk individuals. When presented a risk map, one is easily prone to accept the presented data as *exact* without questioning the methods used to compile the risk map nor obtaining additional information about the underlying granularity. For instance, Figure 1 illustrates two representations of the risk of an event for a given geographic region. The map in Figure 1a divides the geographic region into 20 sub-regions of approximately the same population size. Analogously, the map in Figure 1b is based on a subdivision into 60 sub-regions. Although both maps have been created using the same underlying data and methodology, the resulting maps lead to different conclusions about the risks in some of the sub-areas.

The main cause for this discrepancy is that risk and population data are usually provided for some polygonal sub-divisions of a geographic region with varying population counts (e.g. census blocks) as opposed to a fine-grained grid structure. For example, the population count of a census block group can range from 600 to 3,000 people, whereby a census block can have a population of 0. The above comparison poses the question of how to construct *meaningful* risk maps with an *adequate* granularity.

Lawson et al. describe how incidence data can be represented [2]. If observations are sparse for a geographic region, smoothing methods, such as Kriging, may be applied. The geostatistical method of Kriging in the context of disease mapping is discussed in more detail by Olaf Berke [3]. Alternatively, incomplete data can be estimated by adapting traditional maximum likelihood estimations via expectation maximization (EM) algorithms [4]. The different mapping approaches can be extended to spatio-temporal models [5], [6]. Disease rates at different levels of a hierarchical health administrative structure and the influence of space and time on disease maps has been investigated by MacNab and Dean [7], [8].

As risk data varies within geographic regions, zones of highest risk may need to be identified. Statistical methods generally do not provide the necessary statistical confidence. Ugarte et al. introduce methodologies to estimate confidence intervals for relative risks [9]. Heterogeneity of disease rates across sub-regions of a study area is investigated by Martuzzi



(a) 20 sub-regions (b) Division into 60 sub-regions

Figure 1: Risk maps of same geographic region with division into different granularities.

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

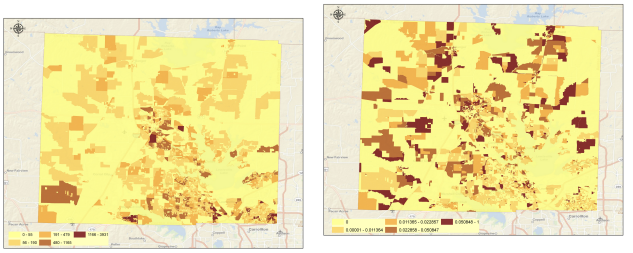
and Hills [10]. Hanafiah et al. investigate the usefulness and comparability of global risk maps for Hepatitis A infections [11] and establish that it is difficult to compare maps.

The earliest disease maps used point symbols to represent the locations of diseases across geographic space. Such maps become inappropriate for representing phenomena where the outcomes are measured as ratios rather than counts. Disease maps, or risk maps more generally, are commonly constructed by computing rates in a population rather than raw counts. It is desirable that disease maps portray risk as a continuum across geographic space rather than discrete points of observations. Choropleth maps are commonly used to create such representations of disease risk. They are constructed by shading areas (typically representing administrative units) with colors or intensities of colors that are derived by grouping the observed disease data (i.e. disease rates) into classes. However, choropleth maps are subject to a number of criticisms, namely issues of color schemes and classification, areal bias, map unreliability due to the Modifiable Areal Unit Problem (MAUP), and small numbers. A detailed literature review of these criticisms and methods to address them is beyond the scope of this paper.

In this paper, we are comparing two systems to generate risk maps: The Web-based Disease Mapping and Analysis Program (WebDMAP) and the Universal PArtitioning System (UPAS). WebDMAP implements an adaptive kernel density estimation procedure for spatially indexed data. UPAS is based on the population-uniform partition of a geographic region.

II. Methodology

In the following sections the adaptive kernel density estimation procedure used in WebDMAP, as well as the population-uniform partitioning of UPAS are introduced. As an example for both methods, a specific attribute representing at-risk populations was examined in Denton County, Texas. On all of the following maps darker colors represent higher rates of at-risk individuals, whereas lighter colors correspond to lower risk rates. The map in Figure 2a shows the at-risk population when simply considering quantities of at-risk individuals in census blocks. The map in Figure 2b shows the at-risk population as proportions of the populations of census blocks (i.e. as risk rates). It becomes evident that the overall risk rate for the census block can significantly differ based on the total population of that census block.



(a) Numbers of at-risk individuals in each census block

(b) Proportion of population of each census block comprised of at-risk individuals

Figure 2: Risk maps of same geographic region with division into different granularities.

In many cases risk data is available at some level of subdivision for a study area. Examples for such sub-divisions are census blocks, census tracts, and areas delineated by zip code boundaries. The corresponding population data for the United States can be obtained from the U.S. Census Bureau website at different geographic granularities. However, the sizes of these sub-divisions are not homogeneous.

A. WebDMAP

WebDMAP uses a kernel density estimation approach to address the issue of rate instability that is caused due to small population counts (i.e. denominators). Rates estimated for small areas such as census blocks are known to be unstable in that the addition or removal of a few cases (i.e. numerators) will result in drastic changes in the estimated disease rate. The WebDMAP approach uses kernels or spatial filters of varying sizes to estimate rates of disease burdens at each point of a predefined grid of regularly or irregularly spaced points. Rates are computed by dividing the numerator counts by the denominator counts that fall within the area encompassed by each spatial filter. The size of this filter at each grid point is determined based on the underlying density of the control population and a user defined threshold that defines the minimum number of controls that must be included within the spatial extent of each filter. The grid of points on which the spatial filters are centered are constructed using a quad-tree based approach that recursively divides the study area into polygons that contain some minimum population size [12]. The polygon centroids are used as the grid definition. The reasoning behind the quad-tree approach is to ensure that the grid appropriately represents the population distribution. Since the size of the spatial filter is driven by population density, not controlling the grid definition can result in redundancies or omissions in the rate calculations. Yiannakoulias et al. compared a quad-tree based grid with a uniformly spaced grid to search for simulated clusters using the spatial scan approach to disease clustering [13]. They found the quad-tree based approach was not only computationally more efficient, but was also more sensitive to high-resolution spatial clusters.

An advantage of the WebDMAP approach over other forms of disease mapping is that it minimizes the overall variance in estimated rates, while maximizing the amount of geographic detail portrayed on the map. This is a desirable property as the map obtained provides stable estimates of disease burdens and maintains high levels of geographic resolution. The map in Figure 3a was constructed using

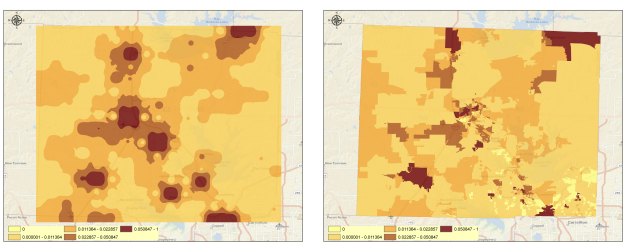
a population threshold of 600. This corresponds to the population sizes that were included in the UPAS approach (Figure 3b).

B. UPAS

As a consequence of the non-uniform population densities, risk maps that are breaking down the at-risk rates at a census block level (shown in Figure 2) may not reflect a *meaningful* picture of the actual risk distribution. Consequently, multiple census blocks have to be combined to construct larger areas of significant population size. Due to the fluctuating populations sizes across census blocks, it becomes evident that naïvely combining census block in an arbitrary or ordered fashion will not yield maps with homogeneous sub-divisions. To construct meaningful risk maps, whereby each at-risk case has approximately the same weight, it is desirable to obtain subdivision with comparable population sizes. Such sub-divisions can be generated by the Universal PArtitioning System (UPAS).

UPAS is a generic partitioning system designed to partition data in a variety of formats. Currently, methodology to partition geographic information in the form of polygonal sub-regions (e.g. census blocks) has been implemented. In particular, a region is partitioned into k sub-regions, such that the population of each of the sub-regions is approximately uniform.

The partitioning approach follows a greedy strategy. Without loss of generality, assume that the underlying geographic region is a county represented by the set of its census blocks. In its first iteration, the partitioning algorithm will split the region into 2 sub-areas with weighted population sizes. The weights are determined by the algorithm based on the desired final number k of sub-divisions. This approach is recursively repeated until k partitions have been obtained. Ideally, all partitions have the same *optimal* size $opt = pop/k$, whereby pop is the sum of the population of the individual census blocks in a given region. It has been shown that the maximum deviation from opt is bounded by the maximum size of a sub-division, and therefore the maximum difference for any two of the k sub-division is twice that amount. However, in practice, the algorithm outperforms this worst-case bound by far. The details of the underlying partitioning algorithm and corresponding formal proofs are beyond the scope of this paper can be found in [14].



(a) WebDMAP (b) UPAS

Figure 3: At-risk population estimation of WebDMAP and UPAS

The algorithm yields k sub-areas of the original region. The sub-areas are delineated by borders that originally belong to the underlying *smallest polygonal unit*. For the example discussed in this paper, this unit corresponds to census blocks. Each of the k sub-areas contains multiple census blocks and no census block is crossed by any of the sub-area boundaries. Therefore, the polygon edges delineating the sub-areas originally belong to census blocks and are the cause of the *jagged* appearance. The advantage of preserving the exact shape of the smallest unit is that exact population and risk counts can be maintained.

The regions of different risk levels are delineated along census block boundaries, since the partitioning algorithm respects census blocks. As census block borders are not displayed, neighboring census blocks within the same risk categories turn into a seemingly larger region. Displaying census block borders would distract from the actual risk map and introduce unnecessary detail. Potentially, the borders between different risk categories could be interpolated.

For the example of Denton County the partitioning algorithm has been applied to generate $k = 720$ sub-regions. The total population pop of Denton County is 432,976, which implies an optimal partition size opt of approximately 601 individuals. 601 closely corresponds to the optimal size of 600 used for WebDMAP. The parameters have been chosen, such that the results of both systems can be compared more objectively. The resulting risk map of UPAS is illustrated in Figure 3b.

C. Comparison of WebDMAP and UPAS

WebDMAP and UPAS use fundamentally different approaches to map risk rates. Initially, it seems difficult to visually compare both approaches. Parameters for WebDMAP and UPAS have been chosen such that the WebDMAP threshold and the number of partitions generated by UPAS correspond. Since the WebDMAP threshold was set to 600, UPAS was set to generate 720 (approximately 432,976/600) partitions. To facilitate an objective comparison, the resulting maps of both systems have been combined into a 3d-plot. Figure 4 shows the risk maps generated by WebDMAP and UPAS as a single graphic. The shades of the figure correspond to Figure 3b, i.e. they illustrate the risks obtained by UPAS. The risks estimated by WebDMAP correspond to Figure 3a and are overlaid in Figure 4 as elevation. The higher the elevation of a given point, the higher the risk as estimated by WebDMAP. Flat areas correspond to regions of low risk. The 3d-plot shows that darker areas mostly coincide with peaks or slopes in the plot, which demonstrates a good correlation of both approaches. Further, flat areas coincide with lighter colors, representing low at-risk rates, respectively. While WebDMAP produces smooth-looking

maps, UPAS respects census block boundaries to preserve exact population and at-risk counts for all geographic sub-regions. Both approaches yield similar maps expressing comparable risk distributions.

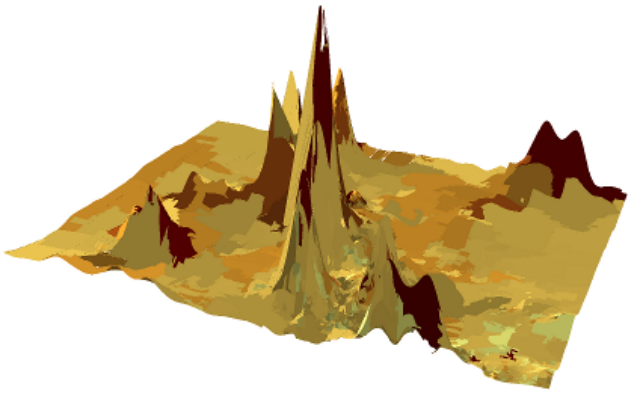


Figure 4: Comparison of WebDMAP and UPAS results

III. Conclusion

While risk maps are used widely in public health and a variety of other applications, it is essential that these maps are *meaningful*, i.e. they adequately represent the distribution of the at-risk population within a geographic region. As uniform population distributions across an entire region rarely occur, a naïve at-risk count at a per sub-division level may convey an inaccurate representation of reality. Both, WebDMAP and UPAS provide approaches to generate risk maps that take into a account the non-uniformity of population distributions. WebDMAP uses a parameter to set a population threshold, while the parameter utilized in UPAS determines the number of population-uniform partitions to be generated. Hence, both systems rely on human input to determine and adequate threshold or number of partitions.

Acknowledgment

Part of this research has been funded by NIH grant NIH 1R15LM010804-01.

References

[1] J. Snow, *On the Mode of Communication of Cholera*. John Churchill, 1855.

[2] A. B. Lawson, D. Boehning, A. Biggeri, E. Lesaffre, and J.-F. Viel, *Disease Mapping and Risk Assessment for Public Health*. John Wiley & Sons Ltd., 1999, ch. Disease Mapping and Its Uses, pp. 3–13.

[3] O. Berke, "Exploratory disease mapping: kriging the spatial risk function from regional count data," *International Journal of Health Geographics*, vol. 3, no. 1, 2004.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[5] T.-H. Wen, N. H. Lin, C.-H. Lin, C.-C. King, and M.-D. Su, "Spatial mapping of temporal risk characteristics to improve environmental health risk identification: a case study of a dengue epidemic in taiwan." *Science of the Total Environment*, vol. 367, no. 2-3, pp. 631–640, 2006.

[6] L. A. Waller, B. P. Carlin, H. XIA, and A. E. Gelfand, "Hierarchical spatio-temporal mapping of disease rates," *Journal of the American Statistical Association*, vol. 92, pp. 607–617, 1996.

[7] C. B. Dean and Y. C. MacNab, "Modeling of rates over a hierarchical health administrative structure," *Canadian Journal of Statistics*, vol. 29, no. 3, pp. 405–419, 2001.

[8] Y. C. MacNab and C. B. Dean, "Spatio-temporal modelling of rates for the construction of disease maps," *Statistics in Medicine*, vol. 21, no. 3, pp. 347–358, 2002.

[9] M. Ugarte, A. Militino, and B. Ibanez, "Confidence intervals for relative risks in disease mapping," *Biometrical Journal*, vol. 45, no. 4, pp. 410–425, 2003.

[10] M. Martuzzi and M. Hills, *Disease Mapping and Risk Assessment for Public Health*. John Wiley & Sons Ltd., 1999, ch. Estimating the Presence and the Degree of Heterogeneity of Disease Rates, pp. 321–327.

[11] K. Mohd Hanafiah, K. Jacobsen, and S. Wiersma, "Challenges to mapping the health risk of hepatitis a virus infection," *International Journal of Health Geographics*, vol. 10, no. 1, pp. 1–8, 2011.

[12] H. Samet, "The quadtree and related hierarchical data structures," *ACM Computing Surveys (CSUR)*, vol. 16, no. 2, pp. 187–260, 1984.

[13] N. Yiannakoulias, A. Karosas, D. Schopflocher, L. Svenson, and M. Hodgson, "Using quad trees to generate grid points for application in geographic disease surveillance," *Advances in Disease Surveillance*, vol. 3, no. 2, 2007.

[14] T. Jimenez, A. R. Mikler, and C. Tiwari, "A novel space partitioning algorithm to improve current practices in facility placement," *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, vol. 42, no. 5, pp. 1206–1215, 2012.