Mike Ramsay

GEOG 461

11 February 2025

**Assignment #5: Report on De-identifying Data**

Maintaining confidentiality with personal and identifiable records can be a big challenge when dealing with sensitive data, particularly when it contains geospatial elements. There are several strategies used which modify this data in such a way that it is still useful for data analysis, yet protects the privacy of the individuals. One of those being Geomasking or changing the location of data in an unpredictable way through the use of algorithms (Allshouse, 2010). Also, there are rules and requirements that covered entities must follow to further protect sensitive data, these entities include health plans, clearinghouses and providers, like doctors and pharmacists (CMS, n.d.). Failure to protect this sensitive personal information can result in legal repercussions.

These de-identification processes are crucial in preserving the privacy and protection of sensitive data. There are several legal frameworks in place to protect this data, such as HIPAA or the Health Insurance Portability and Accountability Act (CMS, n.d.). Under this standard, certain identifiable data types are stripped away. The two methods for de-identification are safe harbor which strips away a lot of the identifying metadata/residual information and then expert determination which employs statistical and scientific principles to de-identify the data (OCR, 2025). The goal of data de-identification is to separate the identifiable information from individual’s health information to reduce the risk of re-identification. However, there is always the risk of re-identification but these practices are used to minimize this risk.

**Overview of Map-Based Visualizations**

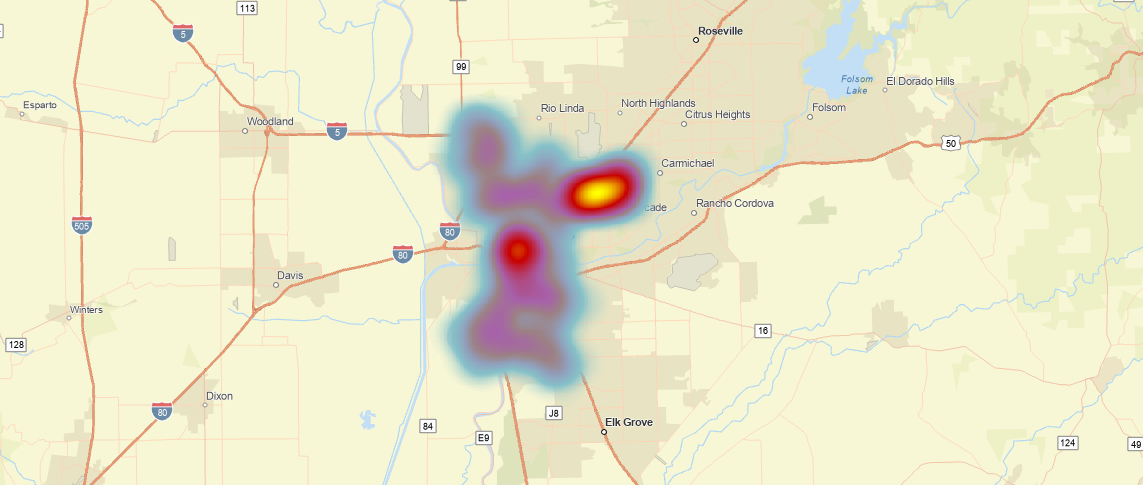


Figure 1. Heatmap

One of the primary map-based visualizations are heat maps, as seen in Figure 1 below. One of their main advantages is that they can help maintain confidentiality of sensitive geospatial data. These maps utilize a gradient, depicting variations in densities in a region, usually without the need for a key as colors can help convey this idea. They achieve this by removing individual data points, meaning exact counts and locations of the data are hard to pick out. However, a weakness of this is that at some scales, individual data points become more visible, as seen in Figure 2, causing heatmaps to lose some of their confidentiality. But further removing administrative boundaries like county lines to further reduce risk of re-identification. Cluster point maps, like the one in Figure 3, are also useful map-based visualizations. This method is best to show exact numbers at different scales, without sharing the individual point locations. However, they experience the same issues as heat maps at smaller scales with individual points becoming more visible. But this type of mapping is good for hospital leadership to identify hotspots without giving away protected health information.

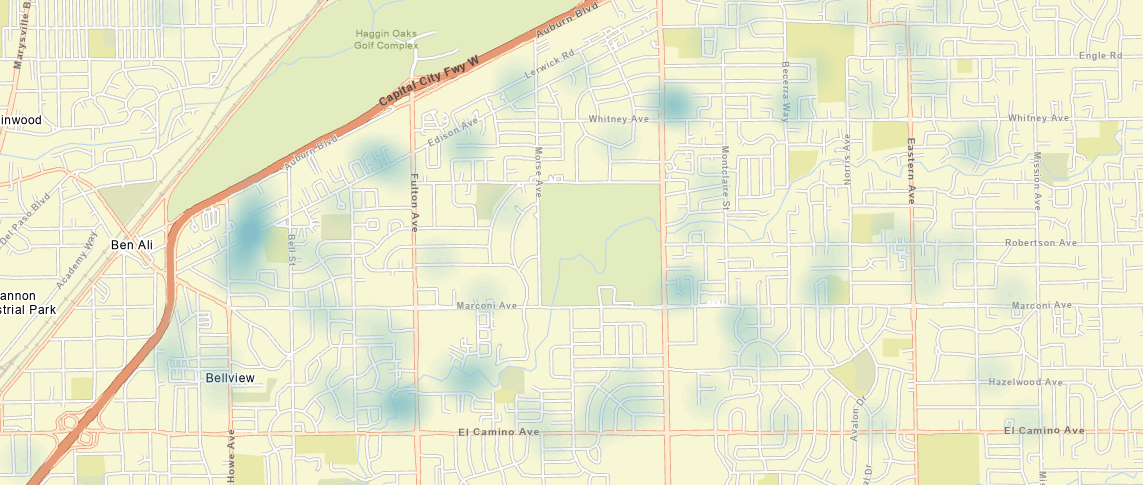


Figure 2. Zoomed into heatmap

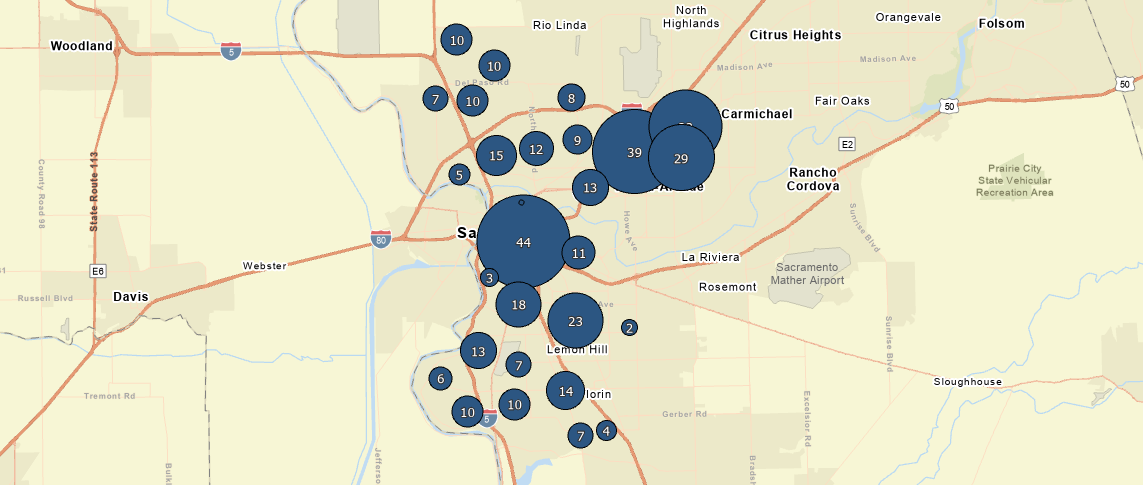


Figure 3. Point Cluster map

**Overview of Small-cell Suppression**

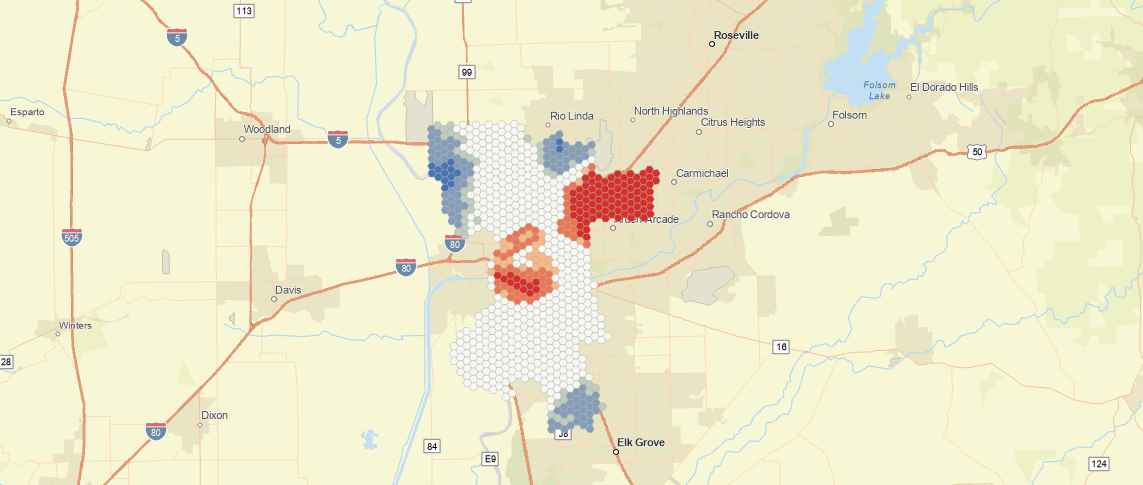


Figure 4. Hexbins

Small Cell Suppression is another method to protect sensitive geospatial health information. The most common of which are hexbins, seen in Figure 4. Hexagons are used as they have more flush sides, as opposed to a square or grid-like structure. This tool is useful to show the exact number of data points at different scales but not identifying individual points. These hexbins can also be reclassified for coins, like high blood lead counts in Figure 5 below. This functions as a kind of density map, while still protecting the privacy of the individual data points. Some of the disadvantages of hexbins are that there can be potential classification issues when data points fall on the boundaries of the bins and the hexbin grid can also obstruct the underlying map and geography but this can be worked around by adjusting the grid’s opacity.

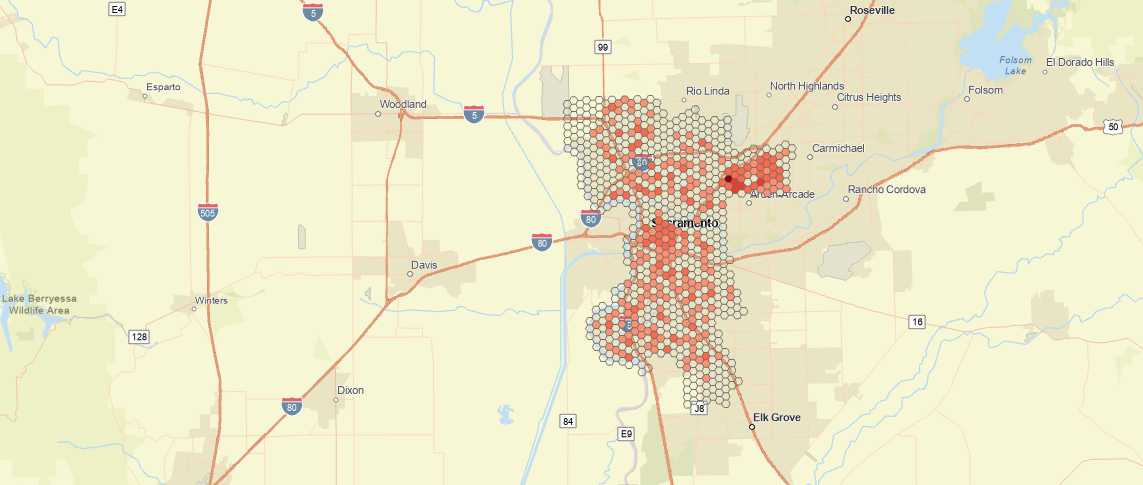


Figure 5. Symbolizing hexbins by count

**Overview of Generalization and Aggregation**

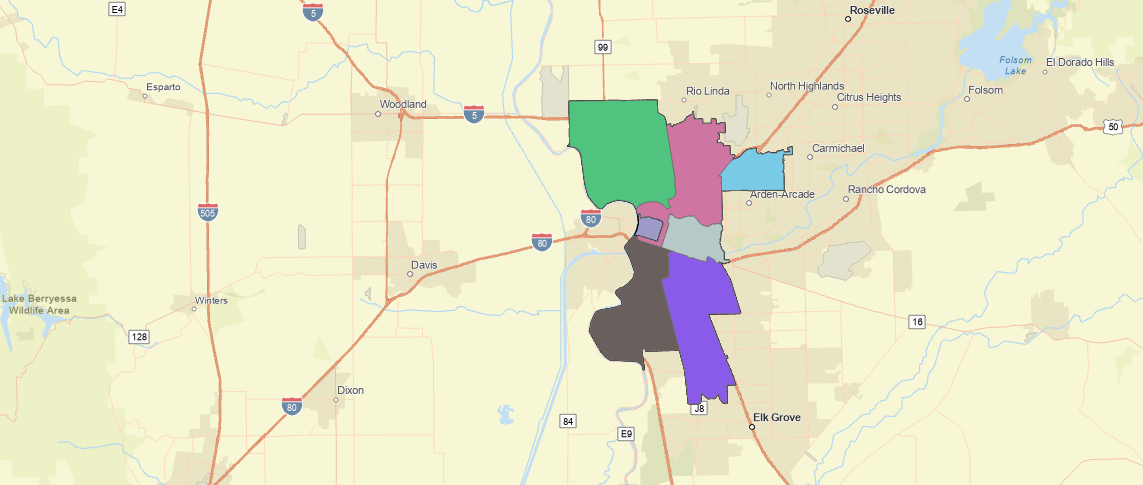


Figure 6. Merged zip codes for

Generalization and aggregation of data shows how to protect sensitive information using methods that still show relevant patterns in the data. In Figure 6 above, the zip codes are merged together and data from different years is used to obtain a minimum of 5 lead cases in each zip code. This approach decreases temporal resolution to maintain spatial resolution. As seen below Figure 7 depicts the use of rounded coordinate maps which protect the exact locations of the data, seen in Figure 8. A stacked cluster map, as in Figure 9, can also be used to generalize and protect sensitive geospatial data. The strengths of this are to protect potentially sensitive geospatial data, like addresses, however a weakness is that these techniques alter and can tend to oversimplify the data which may not represent the real distribution of data from the untouched dataset.

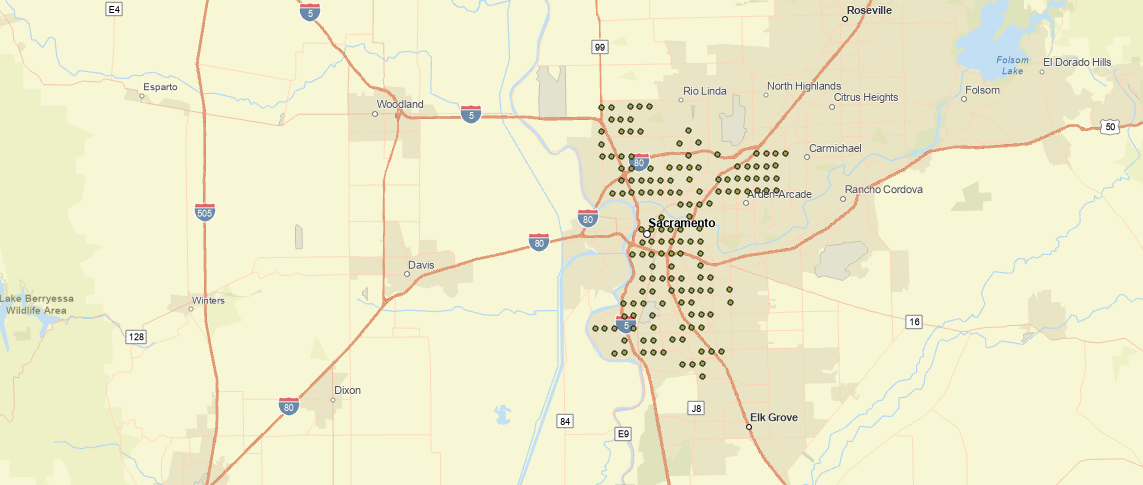


Figure 7. Rounded Coordinated Map



Figure 8. Displacement Map for Rounded Coordinates

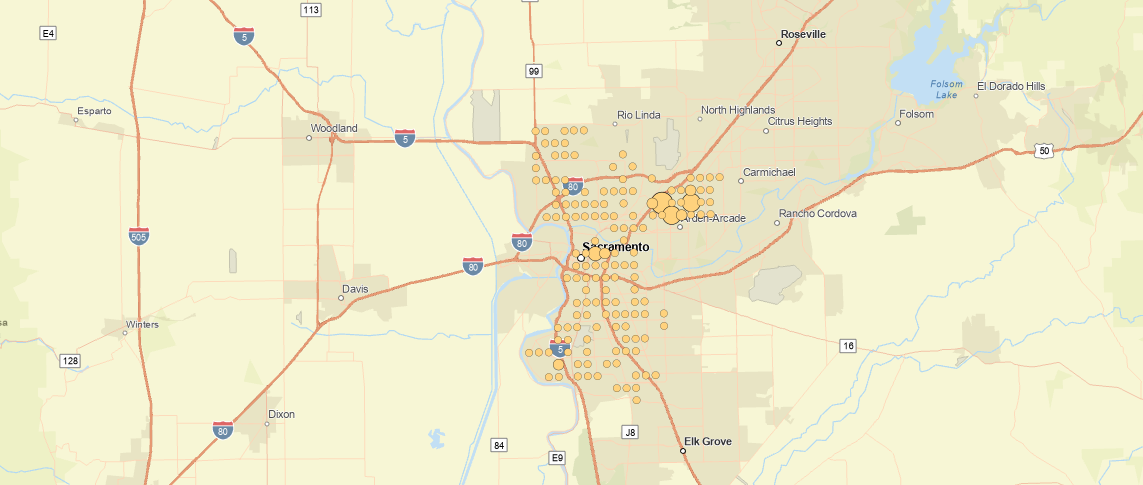


Figure 9. Stacked Clusters Map

**References:**

Allshouse, William B., et al. “Geomasking sensitive health data and privacy protection: An evaluation using an E911 database.” *Geocarto International*, vol. 25, no. 6, Oct. 2010, pp. 443–452, https://doi.org/10.1080/10106049.2010.496496.

CMS, “Are You a Covered Entity?” *CMS.Gov*, Centers for Medicare and Medicade Services, www.cms.gov/priorities/key-initiatives/burden-reduction/administrative-simplification/hipaa/covered-entities. Accessed 11 Feb. 2025.

OCR, “Methods for De-Identification of Phi.” *HHS.Gov*, Office for Civil Rights, 3 Feb. 2025, www.hhs.gov/hipaa/for-professionals/special-topics/de-identification/index.html#standard. Accessed 11 Feb. 2025.