Data Analysis Tools for Informing Policy Decision – Embeddings

Opioids are substantially overprescribed, frequently with harmful consequences. Studies show that prolonged opioid prescriptions can increase the frequency of an opioid overdose-related emergency by 60% (Kuo 2016). This epidemic is most prevalent in poor, uninsured, urban residents of generally Caucasian or African-American origin (McDonald 2012).

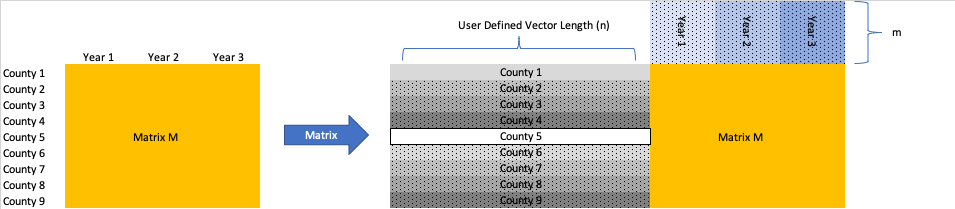
While some policies have reduced the severity of the opioid crisis, new policies will need to be enacted to continue this progress. In order to determine the best policies for the State of Minnesota, the various factors driving this crisis need to be drawn out. Fortunately there are many data analysis tools at the disposal of researchers in this field.

One such tool is matrix factoring, which is frequently used in movie recommendation algorithms. A researcher using this technique to investigate future opioid policy can provide a matrix containing historical opioid use or prescriptions, sorted by two or more data attributes, such as year, location, distance to physician, age, ethnicity, gender, or other metrics. **Table 1**, below, shows an example of such a matrix, shown as a portion of the table correlating opioid use by year and county in Minnesota.



**Table 1:** The top few lines of a table correlating opioid prescriptions sorted by county and year

Once the data is in matrix form, the user can factor the matrix into vectors for each data attribute. In the case of the current analysis, the matrix is factored into a vector for each county and another vector for each year the data covers, as shown in **Figure 1**. The size of the vectors is defined by the analyst before the matrix factoring. Matrix factoring takes place by a steepest gradient optimization. Once the matrix is factored into these vectors, often called latent factors, the researcher can determine the similarity between two different data attributes.



**Figure 1:** Matrix factoring of Matrix M into vectors for each county and each year

In the example from **Table 1** the researcher can take the set of vectors corresponding to the counties in Minnesota and use cosine similarity to identify counties that are facing similar problems. The process described above can be found at the link below, which contains the entire code used to produce the analysis here.

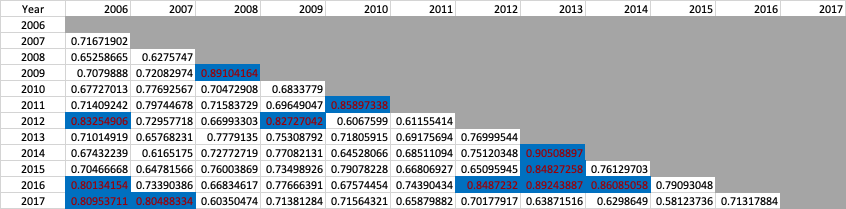
https://github.com/rvatassery/MinnesotaOpioid/

# Discussion

The similarities can be obvious, and the obvious similarities allow for the researcher to validate the result. For example, as expected Hennepin and Ramsey counties are the most similar to each other in terms of their latent vectors, whereas Hennepin and Pine counties are very dissimilar. Anoka, Carver, and Dakota counties are also very similar in this analysis.

The primary benefit from matrix factoring in the context of addressing the opioid epidemic is that the counties can be split up into groups of counties for the purposes of applying various remedies. For example, given that Anoka, Carver, and Dakota Counties are very similar, a program that produces results in one county has a substantial chance of success in the other two. Targeting counties in this way can properly direct funding dollars.

Matrix factoring as demonstrated here can provide guidance based on other data attributes as well. In the data set shown in **Table 2** vectors for years are also output from the same matrix factoring method, so we can assess which years the opioid epidemic saw the same set of factors operating. As expected, the most similar years are the adjacent years, but the years that are non-adjacent provide an opportunity to identify important drivers for the opioid crisis.



**Table 2:** Cosine similarity chart for years in the opioid epidemic

Because the algorithm doesn’t use the data labels (county or year in this case) at all, other than to present back to us what the result is, the cosine similarity results will only find correlations that are relevant to the opioid data set. They will be free of human biases.

