Identifying Vaccination Amenable Populations

Matthew Lawson and Benjamin Morris

When a large percentage of a population is vaccinated against a communicable disease, such diseases are less able to spread. This is called herd immunity and is important for protecting those in the population that can't be vaccinated due to age, health condition, or allergies. However, people who are able to get vaccinated can opt to not get vaccinated. This threatens herd immunity and puts those who are unable to get vaccinated at risk.

There are two general categories of people who opt to not get vaccinated: people who are opposed to vaccinations (anti-vaxers) and people who are economically unable to get vaccinated. These two categories of unvaccinated people tend to respond to outreach differently. Anti-vaxers tend to still not get vaccinated while the economically unable population will get vaccinated if they are able. If we can identify the two different communities, then we can focus outreach efforts and resources on the economically unable population instead of the anti-vax population.

The main question is this: How can we identify whether areas with low vaccination rates can be fixed? If we can identify the regions that are fixable, we can be more efficient with resource spending. This will raise the overall percentage of vaccinated people which leads to stronger herd immunity.

About the Datasets:

To answer this question, we combined two datasets - California 7th Grade school immunization rates in 2015 (can be downloaded as a CSV file here) available publicly from the California Health and Human Services Open Data Portal and a list of median household income by city from 2010-2014 census from the Wikipedia page 'List of California locations by income' (downloaded as a CSV with wikitable2CSV). We cleaned and combined these two datasets to examine median family income, percent overdue, school status (public vs. private), and percent with a Personal Belief Exemption (PBE), for each city in our dataset. For cities with multiple schools, each city's average is an average weighted by school population.

Real-world impact:

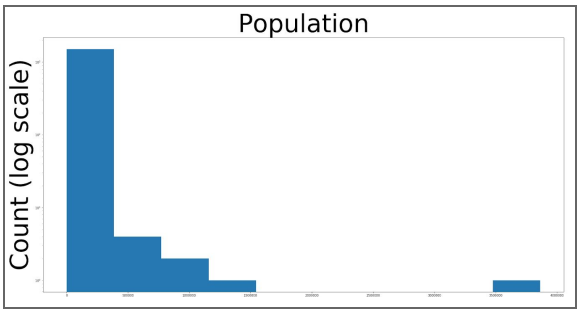
While real world companies may not be interested in categorizing types of vaccination hesitant populations, doing so is important for public health and there is a large literature in characterizing these groups. The literature seems to be divided into

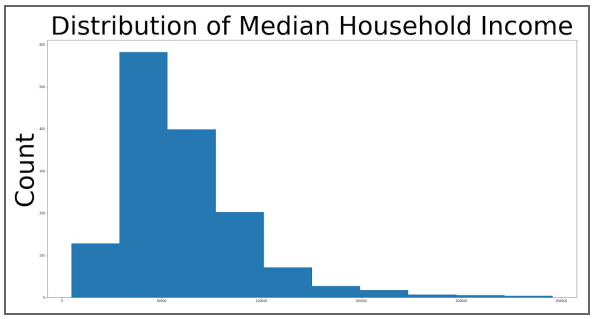
two categories: characterization by attitudes or by attributes (i.e. individual perspectives on vaccines versus demographic traits). At the scale of identifying possible vaccination amenable populations, demographic characterizations are more useful, as this data is widely available, whereas surveying attitudes is time-consuming and difficult.

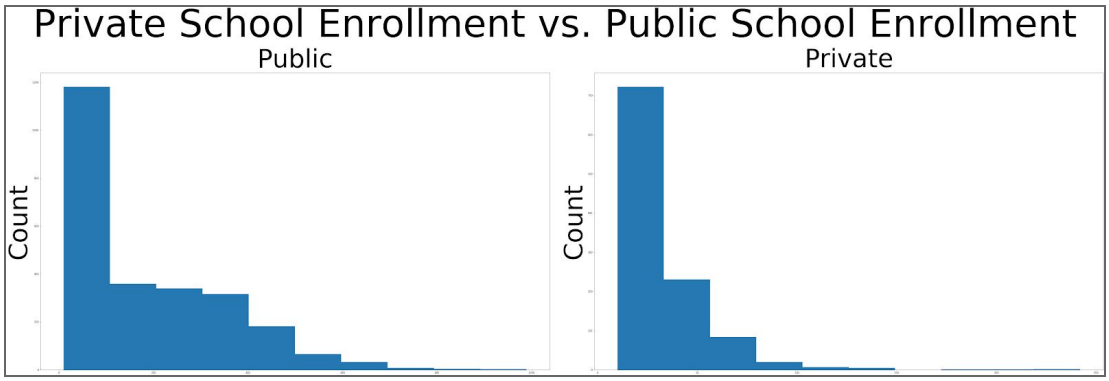
One attribute-based meta-analysis found socioeconomic status (high or low) as a barrier to vaccination in the United States, and several studies with a more global scope have found similar results. For example, a 2009 study of infant medical records by Wei et. al found that parents from well-educated, high SES areas were more likely to refuse vaccination, while another study by Wu et. al found that low-income families without access to Women, Infant, and Children (WIC) support were also likely to refuse vaccines.

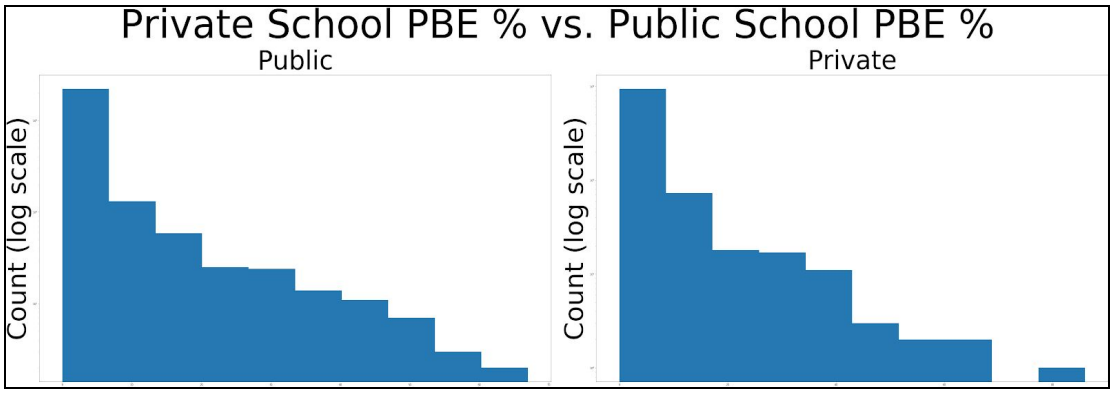
Recent data science research in this area has primarily concerned predicting outbreaks and their severity, however, we think it is more important to find ways to prevent outbreaks by improving vaccination programs in high-risk areas. This is an area that has not been as closely identified by data scientists.

Exploratory results:









We explored the income dataset before any analysis to make sure that it would work for our purposes. Nearly all cities have a population of fewer than 500,000 people. The mean population is 23,750. The median household income looks a lot more normal with a mean of $60,765. The immunization dataset is just as interesting as the income dataset. Public schools have larger enrollment on average than private schools. PBE (percentage of people with personal belief exemption) has many values at 0 and a few non-zero values. They are distributed the same for private schools and public schools. PBE has a mean of 3, a standard deviation of 7, and a maximum value of 86.

Methods:

In order to answer our questions, we relied on clustering methods like hierarchical and k-means combined with IForests to remove outliers. To analyze our results, we used a one way ANOVA and a post-hoc Tukey's Honestly Significant Difference test. As we were looking for latent communities based on demographics like SES and vaccination behaviors, we expected that these clustering algorithms would help find maximally similar communities.

In our initial analyses, we did not remove outliers but noticed that our dataset had a number of extreme outliers such that in hierarchical clustering, each of our k points returned were outliers. As our k points from hierarchical were used as the centroids for k-means initialization, this meant that including outliers gave us poor clustering results as measured by the Davies-Bouldin Index. We used the scikit-learn implementation of iForests and increased the number of estimators to 1000 in order to reduce inter-run variability in which points were designated outliers and removed. Each variable was then 0 centered and adjusted to have standard deviation 1 in order to reduce the impact of high-valued variables like income. Hierarchical clustering was then conducted and returns a dictionary of centroid locations for each k in the range provided. These centroids were then passed to k-means, which assigns each data point to one of the k clusters. For both clustering methods, the L2 norm was used to calculate distance. We then plotted each choice of k versus that clustering's Davies-Bouldin Index, which assesses the ratio of the variability within clusters to the spacing between clusters using the following formula:

$$DBI = \frac{1}{N}\sum_{i=1}^{N} max_{i\,!=j}\frac{S_i+S_j}{M_{i,j}},\ where\ S_i = (\frac{1}{T_i}\sum_{j=1}^{T_i}(X_j - A_i)^2)^{\frac{1}{2}},\ and\ M_{i,j} = ||A_i - A_j||_2$$

With i and j the clusters, T_i the size of cluster i, and A_i the centroid of cluster i. A lower DBI indicates better clustering (smaller spread within clusters and larger distance between clusters).

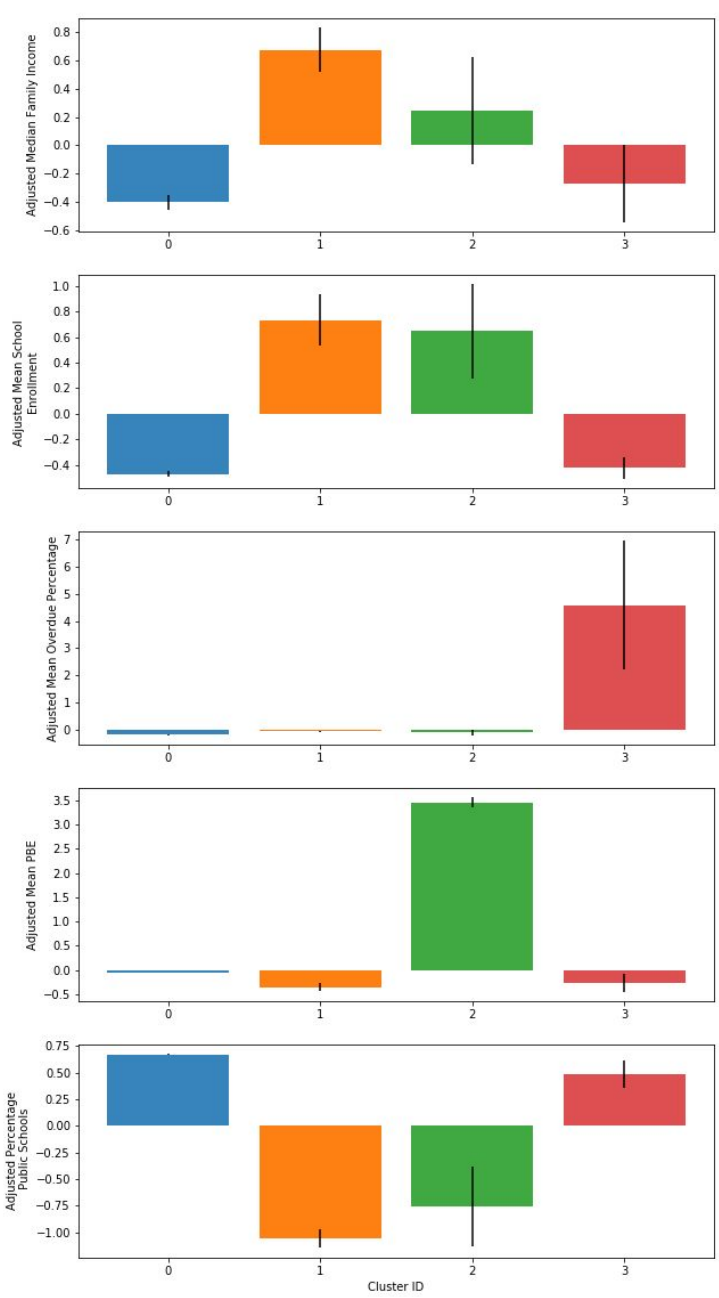
We then selected the optimal k based on an elbow in the k vs. DBI plot that maximizes interpretability and granularity. Values on each of our five variables of

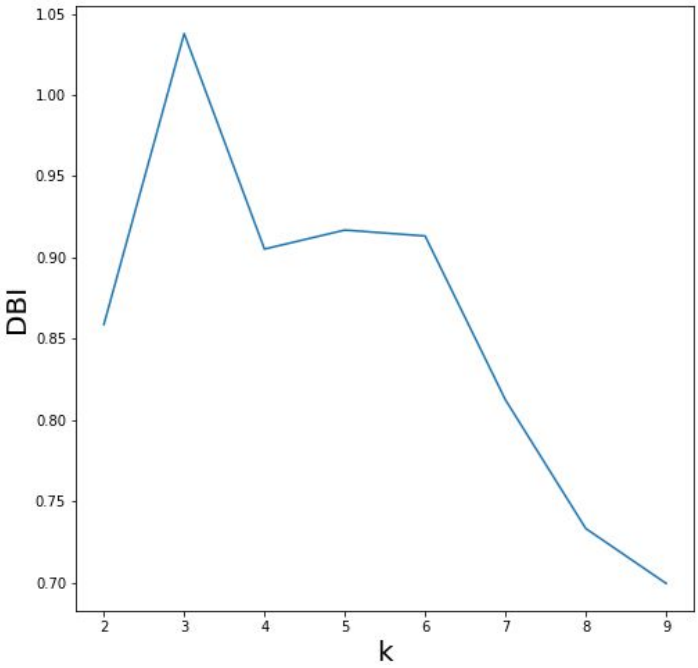
interest were then plotted for each cluster on a radar chart and an ANOVA and pairwise Tukey HSD conducted to examine differences between clusters.

Results

Our plot of DBI vs. k revealed an elbow at k = 4 (Fig. B). Using k=4 yielded clusters of size 352, 206, 35, and 19. One way ANOVAs by cluster on each of the five clustering variables revealed significant differences between clusters (all $p<0.001$). A further Tukey HSD test (family-wise error rate = 0.05) revealed several interesting results: clusters 2 and 3 had significantly higher rates of personal belief exemptions and overdue vaccinations than all other clusters. We then wanted to examine demographic differences between these groups that could explain their distinct vaccination profiles. Clusters 2 and 3 differed in the percentage of public schools and school enrollment, but, contrary to what we had predicted, did not differ in median household income.

Although clusters 0 and 1 were not interesting in terms of their vaccination profiles, the clustering does seem to have successfully differentiated between wealthy cities with lots of private schools, and less wealthy cities with lots of public schools as the two groups with high vaccination rates.





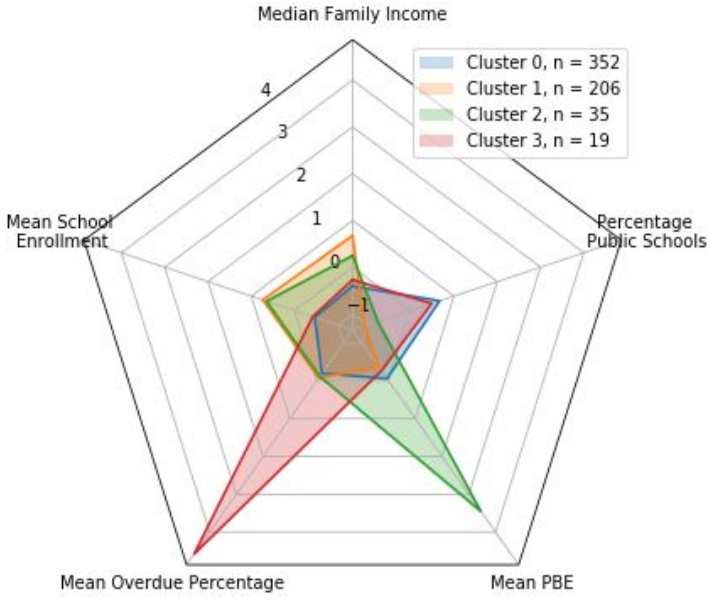


Fig. A Z-scored data attributes grouped by cluster. Data are presented as mean +/- SEM.

Fig. B Davies-Bouldin Index by k

Fig. C Radar plot of data attributes grouped by cluster. Each line is one z score.

Conclusion

Although our methods did successfully find clusters of California cities with low vaccination rates due to overdue vaccines or personal belief exemptions, we were unable to attribute any differences in these clusters' vaccination profiles to socioeconomic status. Several factors could help improve the demographic characterization of high PBE and high overdue populations. As an example, including religious beliefs, political orientation, age, and level of education could be important variables in determining between-cluster differences. As this analysis was conducted at the city-wide level, more granular school-by-school data could also improve detection of demographic predictors of vaccination hesitancy and help inform outreach programs to improve public health and preserve herd immunity.

From an algorithmic level, further work would have to be done in inter-run consistency. Although increasing the number of estimators in the iForest reduced some variability, the DBI vs. K profile changes considerably between runs and as a result the optimal k value and the clusters. The discovery of the two notable vaccination profiles remains relatively consistent between runs, but the small changes are still concerning.

References

Kestenbaum, L. A., & Feemster, K. A. (2015, May 1). Identifying and addressing vaccine hesitancy. *Pediatric Annals*. Slack Incorporated. https://doi.org/10.3928/00904481-20150410-07

Larson, H. J., Jarrett, C., Eckersberger, E., Smith, D. M. D., & Paterson, P. (2014, April 17). Understanding vaccine hesitancy around vaccines and vaccination from a global perspective: A systematic review of published literature, 2007-2012. *Vaccine*. Elsevier BV. https://doi.org/10.1016/j.vaccine.2014.01.081

Wei, F., Mullooly, J. P., Goodman, M., McCarty, M. C., Hanson, A. M., Crane, B., & Nordin, J. D. (2009). Identification and characteristics of vaccine refusers. *BMC Pediatrics*, *9*(1), 18. https://doi.org/10.1186/1471-2431-9-18

Wu, A. C., Wisler-Sher, D. J., Griswold, K., Colson, E., Shapiro, E. D., Holmboe, E. S., & Benin, A. L. (2008). Postpartum mothers' attitudes, knowledge, and trust regarding vaccination. *Maternal and Child Health Journal*, *12*(6), 766–773. https://doi.org/10.1007/s10995-007-0302-4