Community Profiles of Army Installations

SDAL SOCIAL & DECISION ANALYTICS LABORATORY
DATA SCIENCE FOR THE PUBLIC GOOD

Kelsey McMahon (Virginia Tech), Megan Grondine (Virginia Tech), Emily Sheen (Pennsylvania State University), Daniel Liden (University of Minnesota) with Joshua Goldstein and Stephanie Shipp (SDAL)
Joel Thurston and Nathaniel Ratcliff, US Army Research Institute for Behavioral and Social Science Research

Project Description

The goal of this project is to understand and quantify embeddedness of Army Soldiers in their jobs and communities.

Research Questions:

- **Can we relate the characteristics of community embeddedness back to metrics of Soldier performance?**
- **Can similar community profiles be clustered into categories using unsupervised learning methods?**

We searched the web for publicly available data sources that could help expand the installation community profiles that were created by DSPG 2017. After creating a clean and detailed data product, we examined relationships among variables that may indicate how social behaviors and external influences drive Soldiers to successfully perform their jobs. We did not address soldier performance within installations due to delays in obtaining data from ARI.

Data Sources

Publicly available data sources

1. County Health Rankings - overall health data
2. Simply Analytics - popular data, business data, location data
3. NCES - public & private elementary through high school in US
4. ACS - from last year's ARI profiling project

Reviews of embeddedness literature (Shown in References below)

References/Acknowledgements

- Holtom, Mitchell, and Lee: "Increasing human and social capital by applying job embeddedness theory"
- Kiazad, Holton, Hom, and Newman: "Job Embeddedness: A Multifoci Theoretical Extension"
- Lee et. al.: "The Effects of Job Embeddedness on Organizational Citizenship, Job Performance, Volitional Absences, and Voluntary Turnover"
- Zhang, Fried, and Griffeth: "A review of job embeddedness: Conceptual, measurement issues, and directions for future research"

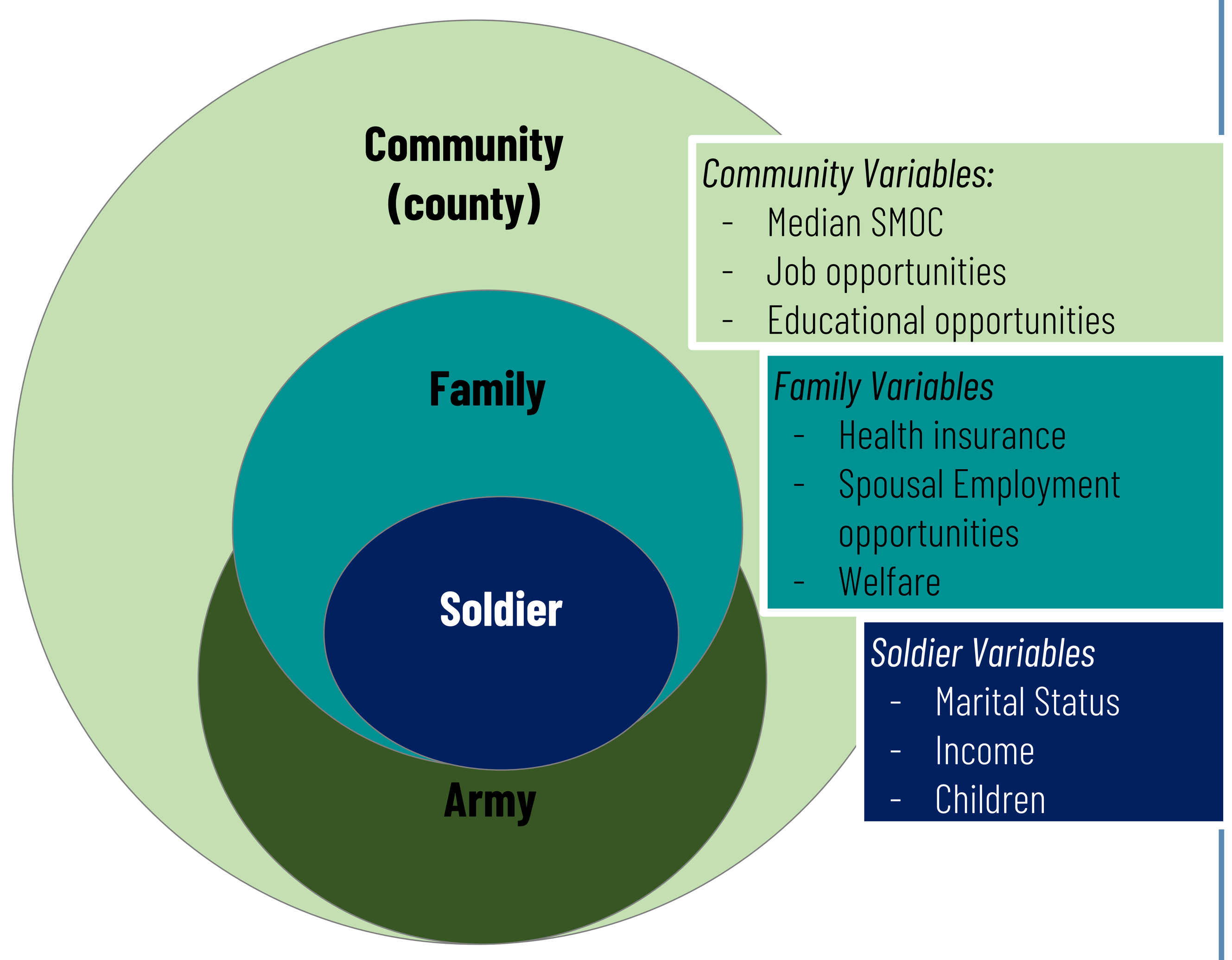


Figure 1: Graphic showing embeddedness of the Soldier within the Army, the family, and the community, and relevant variables related to each. Note SMOC is "Selected monthly owner costs" - the sum of payments for mortgages, including home equity loans and other junior mortgages; for example insurance and utilities.

Clustering Results

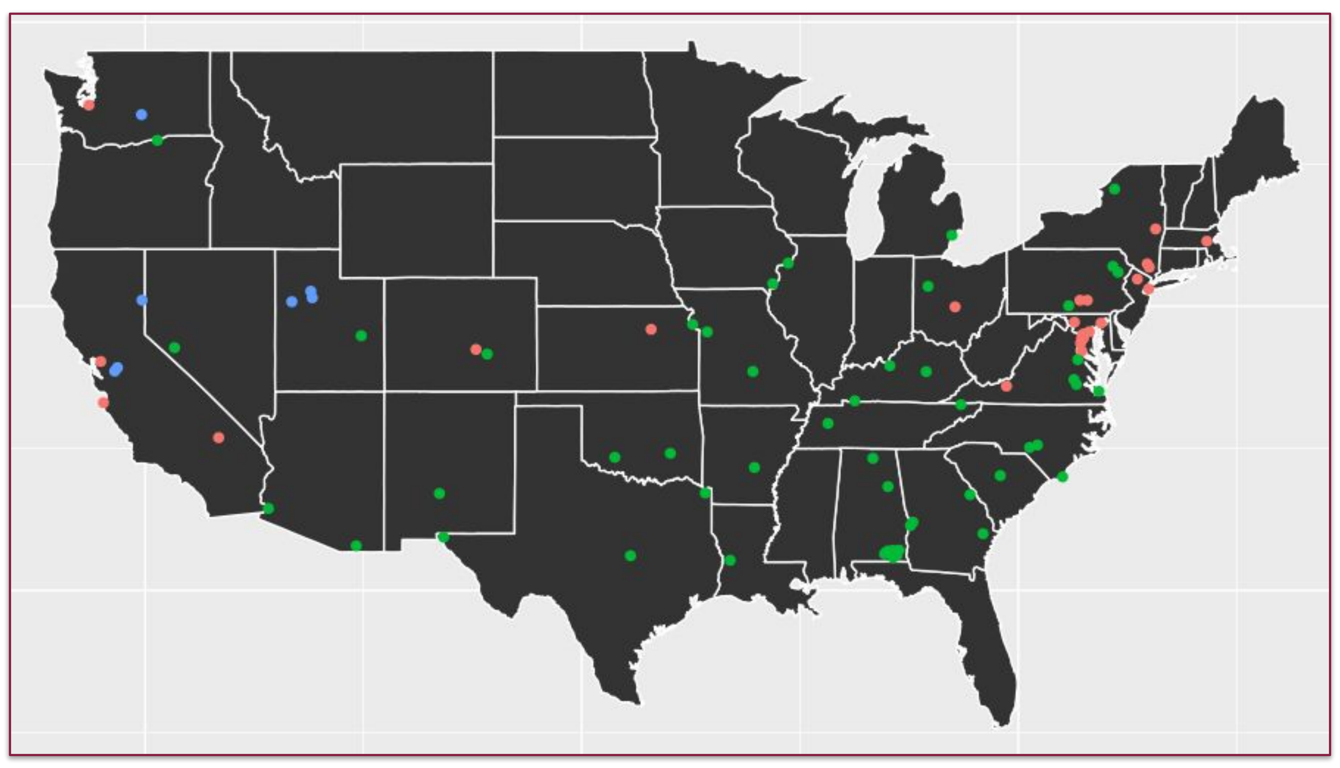


Figure 2: Map of Installations in the contiguous US, colored by random forest clustering algorithm with K = 3 clusters. Notice the geographic correlation of the clustering, with the orange installations clustered in more major metropolitan areas, green scattered across the country, and blue in the Northwestern US.

We implemented K-means and Partitioning Around Medoids (PAM) clustering with Random Forest, extracting variable importance measures from the Random Forest. We chose the PAM random forest clustering method as optimal because it is less sensitive to outliers than K-means, and we can extract variable importance measures to guide our exploratory analysis.

We selected the optimal number of clusters as 3 using the elbow, silhouette, and gap statistic methods. The PAM Random Forest technique results in installation clustering shown in the graph on the left. As you can see, there is a strong geographic relationship within clusters.

The most significant variables identified by the Random forest clustering algorithm include:

- Median home ownership costs, Air pollution, Transportation costs, Exercise access, Obesity, Household density, Mentally unhealthy days, Intersection density, Public transportation use, Smoking

Exploratory Analysis

We started our exploratory analysis with the important variables identified by the random forest clustering algorithm. We plotted every combination of these top 10 variables in a matrix, and identified trends and correlations between them.

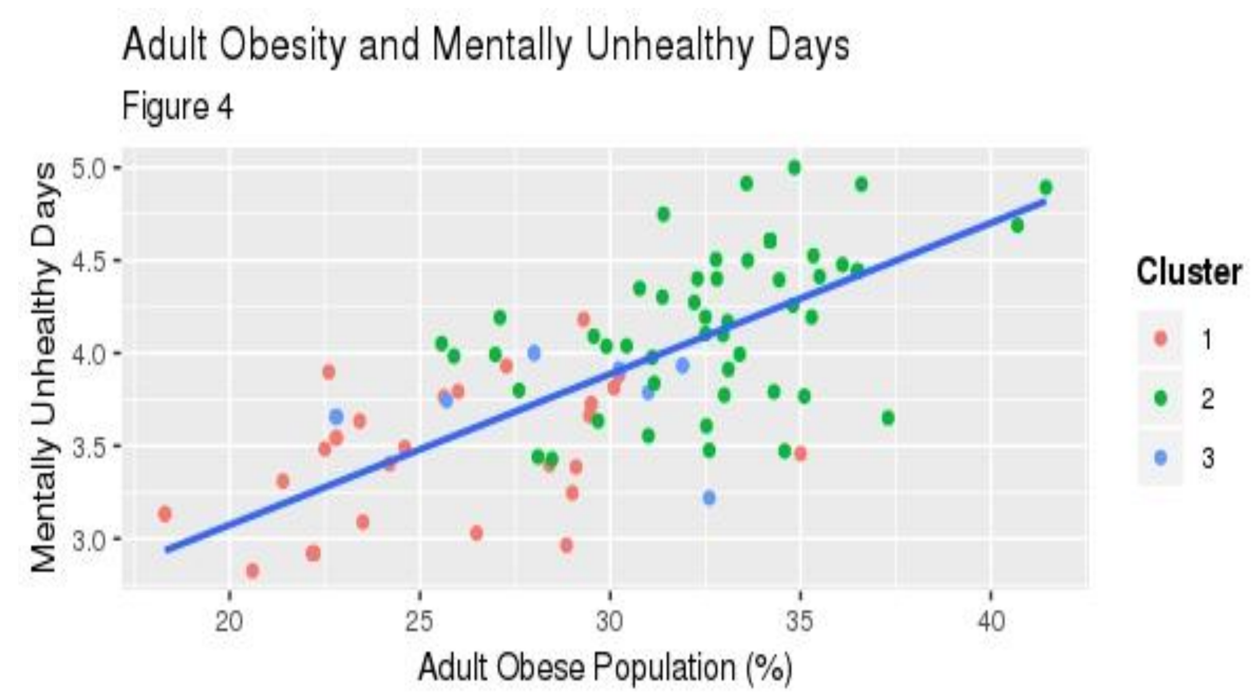
We also plotted other variables not identified by random forest to gain additional insights the algorithm may have missed. We explored histograms and scatter plots of many chosen variables we expected may be important for future analysis.

Separate from the data exploration, we performed a detailed qualitative analysis on a subset of 12 installations, diving into the history, population, installation resources, and community characteristics.

Conclusions and Next Steps

Our research provides detailed community profiles for the Army installations across the United States, as well as a preliminary look at which communities are most similar, identified by our clustering methods.

Now that the community profiles have been expanded to include many interesting characteristics, the community data must be linked back to Soldier performance data. Future research should attempt to link soldier performance to community health and risk factors, and identify which community variables share the strongest relationship with Soldier performance.



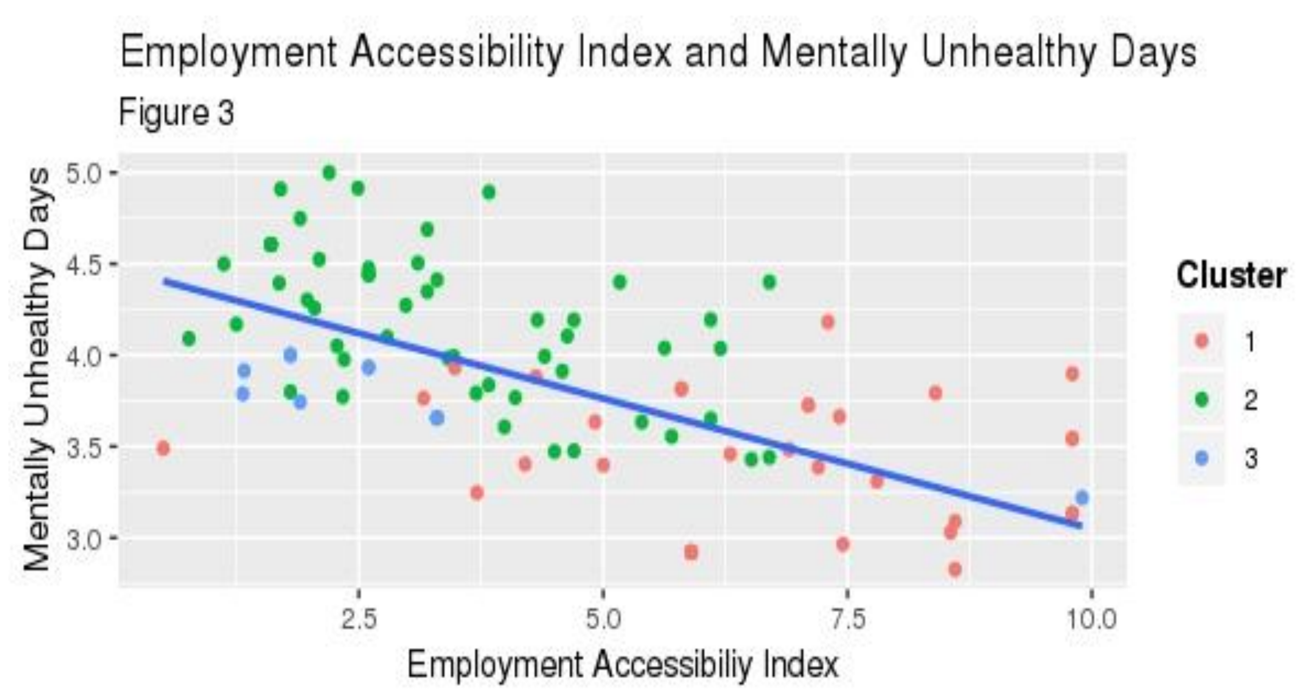


Figure 3: *Scatter plot of Employment Access Index vs. Mentally Unhealthy Days, colored by Random Forest Cluster. Employment Access Index is a Job Access Score from 0 to 10 computed by H+T based on access to jobs and variety of employment types. Mentally Unhealthy Days is from the Behavioral Risk Factor Surveillance System (BRFSS) telephone survey, giving the average number of mentally unhealthy days reported per month for respondents from the county.*
Figure 4: *Scatter plot of Obese Population vs. Mentally Unhealthy Days, colored by Random Forest Cluster. Adult Obesity is calculated by the CDC Diabetes Interactive Atlas and represents the percentage of adults that report a BMI of 30 or more. Mentally Unhealthy Days is from the Behavioral Risk Factor Surveillance System (BRFSS) telephone survey, giving the average number of mentally unhealthy days reported per month for respondents from the county.*
Figure 5: *Scatter plot of Median SMOC vs. Annual GHG per Acre, colored by Random Forest Cluster. Median Selected Monthly Ownership Costs are from the American Community Survey (ACS) and represent owners with a mortgage. According to H+T, ownership costs include mortgage payments, real estate taxes, various insurances, fuels, mobile home costs, and condo fees. Annual Greenhouse Gas Emissions per Acre is from the H+T Index. According to the H+T Glossary of Terms, "This variable is calculated* using the values for Vehicle Miles Traveled, national average fuel efficiency (20.7 mpg), and an average emissions factor (0.438 metric tonnes of CO2 per mile). This per acre measure divides the total block group emissions by the total land acres in the block group."

BIOCOMPLEXITY INSTITUTE

bi.vt.edu/sdal
Stay Connected: *facebook.com/biocomplexity*
twitter.com/biocomplexityvt
youtube.com/c/biocomplexity
google.com/+biocomplexity

VIRGINIA TECH™