Ecological Factors Associated with Self-Reported Mental Health Status

Alyssa Berger, Andrew Cistola

# Introduction

## Background

Ecological factors are known predictors of many health outcomes (1–7).

Neighborhood level deprivation is a useful measure in assessing risk for negative health outcomes and creating targeted interventions within population health management (8–13).

### Machine Learning and Spatial Analysis

Machine learning with spatial analysis has potential to improve these population health management (14–17).

Machine learning shows little benefit over traditional approaches (18,19).

Machine learning research needs methodological improvement (20–23).

Spatial analysis within mental health is limited (24).

## Purpose of this Study

There is a need for studies that can inform public health data science (25–27).

There is a need for studies that can inform translational science in mental health (28–30).

### Specific Aims

Specific Aim 1Identify local socioeconomic factors associated with a higher prevalence of poor self-reported mental health status.

Specific Aim 2 Identify area health resources that are associated with a lower prevalence of poor self-reported mental health status.

# Methods

## Data Collection

Collect estimated zip code level self-reported poor mental health status as percent of population from the 2020 release of the CDC and RWJF PLACES dataset. Collect approximately 400 zip code level socio-economic variables as percent estimates from the 2020 release of the US Census American Community Survey. Connect outcomes with predictors, remove observations with missing values, impute missing data for predictor variables using median values, standard scale all variables. Collect approximately 2000 county level health resource variables from the 2020 release of the HRSA Area Health resource File.

***Feature Selection***

Feature selection is a quantitative approach designed to identify predictors that are most important for a given outcome of interest among a large number of candidates. FS techniques fall are often described as *wrapper*, *filter,* and *embedded* methods(31–33)*.* Wrappers select subsets of features for classification to compare results and include approaches such as Recursive Feature Elimination with Support Vector Machines (SVM-RFE) (32). Filters use metrics to rank features individually and include Linear Discriminant Analysis (LDA) (34,35) and certain applications of Principal Component Analysis (PCA) (36). Embedded methods combine qualities of both by including ranking into the classification process and include Random Forests (RF) (37).

Random Forests. RF create an ensemble of decision trees with bootstrapped sampling of predictors fit to a given outcome (38,39) providing an aggregate measure of importance for each candidate predictor that is useful for variable selection (40–44). RF rely on less assumptions that traditional variable selection methods (40,42) and can integrate multiple statistical approaches (45,46) allowing RF to remain highly accurate with high collinearity (43), significant noise (42,47), or complex interactions (48–50). RF are easily adaptable for various applications (51,52), are relatively resource efficient (38), and available in many open-source libraries (52). RF have bene used widely in bioinformatics (53) for disease risk prediction (54–57) and genome wide association studies (GWAS) (41,43,47,50,58), but recently have begun to be used for spatial analysis in public health research (59,60).

**Principal Component Analysis.** Principal Component Analysis (PCA) constructs orthogonal linear combinations of features that account for a given amount of variation within a dataset (61,62) and is commonly used in bioinformatics for dimensionality reduction through variable aggregation (63–65) or component loading (66–68). PCA is applicable to high dimensional data (69), does not require distribution assumptions (70), and can be conducted with widely available computational resources using open-source software (71).

**Recursive Feature Elimination with Support Vector Machines**. Recursive feature Elimination (RFE) is a simple approach to feature selection to identify a reduced set of predictors that provide the highest prediction accuracy (72–75). RFE can be deployed in similar settings to stepwise regression but utilizes cross-validation with Support Vector Machines to improve reliability (76).

### Specific Aim 1

Utilize open-source machine learning algorithms to identify socioeconomic variables with both high variation and high importance. Conduct cross-validated prediction to identify the smallest number of variables that will achieve the best fitting model for zip codes. Calculate spatially adjusted rates and use an artificial neural network with backwards elimination to predict zip codes in the top quartile. Compare the predictive capability of all predictors, a random set of predictors, other known predictors, and the variables obtained in step 4. Test for OLS assumptions and use domain knowledge to develop an appropriate mixed effects regression model for the purpose of identifying parameter estimates for socioeconomic predictors.

### Specific Aim 2

Using local Empirical Bayes smoothing and LISA quadrants, identify ‘hot and cold spot’ regions and assign nominal labels to contained counties. Use algorithms capable of multi-nominal prediction to identify health resources associated with each category. Using zip code predictors, conduct geographic weighted regression to identify regions where each predictor has significantly higher or lower coefficients. Assign nominal labels and use algorithms capable of multi-nominal prediction to identify health resources associated with each category. Test for OLS assumptions and use domain knowledge to develop an appropriate mixed effects regression model. The purpose is to identify parameter estimates for each county predictor identified in step 8 and an interaction term for county and zip code predictors in step 9.

# Bibliography

1. Casper M, Kramer MR, Peacock JM, Vaughan AS. Population health, place, and space: spatial perspectives in chronic disease research and practice. Prev Chronic Dis. 2019 Sep 5;16:E123.

2. Bozigar M, Lawson AB, Pearce JL, King K, Svendsen ER. A Bayesian spatio-temporal analysis of neighborhood pediatric asthma emergency department visit disparities. Health Place. 2020 Nov;66:102426.

3. Cromer SJ, Lakhani CM, Wexler DJ, Burnett-Bowie S-AM, Udler M, Patel CJ. Geospatial Analysis of Individual and Community-Level Socioeconomic Factors Impacting SARS-CoV-2 Prevalence and Outcomes. medRxiv. 2020 Sep 30;

4. Richardson AS, Collins RL, Ghosh-Dastidar M, Ye F, Hunter GP, Baird MD, et al. Improvements in neighborhood socioeconomic conditions may improve resident diet. Am J Epidemiol. 2020 Oct 13;

5. Holder AL, Wallace DJ, Martin GS. Hotspotting sepsis: applying analytic tools from other disciplines to eliminate disparities. Ann Transl Med. 2016 Aug;4(15):295.

6. Beck AF, Anderson KL, Rich K, Taylor SC, Iyer SB, Kotagal UR, et al. Cooling the hot spots where child hospitalization rates are high: A neighborhood approach to population health. Health Aff (Millwood). 2019;38(9):1433–1441.

7. Eibich P, Krekel C, Demuth I, Wagner GG. Associations between Neighborhood Characteristics, Well-Being and Health Vary over the Life Course. Gerontology. 2016 Jan 29;62(3):362–370.

8. Miles JN, Weden MM, Lavery D, Escarce JJ, Cagney KA, Shih RA. Constructing a Time-Invariant Measure of the Socio-economic Status of U.S. Census Tracts. J Urban Health. 2016 Feb;93(1):213–232.

9. Diez Roux AV, Mair C. Neighborhoods and health. Ann N Y Acad Sci. 2010 Feb;1186:125–145.

10. Scaria E, Powell WR, Birstler J, Alagoz O, Shirley D, Kind AJH, et al. Neighborhood disadvantage and 30-day readmission risk following Clostridioides difficile infection hospitalization. BMC Infect Dis. 2020 Oct 16;20(1):762.

11. Kind AJH, Buckingham WR. Making Neighborhood-Disadvantage Metrics Accessible - The Neighborhood Atlas. N Engl J Med. 2018 Jun 28;378(26):2456–2458.

12. Messer LC, Laraia BA, Kaufman JS, Eyster J, Holzman C, Culhane J, et al. The development of a standardized neighborhood deprivation index. J Urban Health. 2006 Nov;83(6):1041–1062.

13. Walker AF, Hu H, Cuttriss N, Anez-Zabala C, Yabut K, Haller MJ, et al. The neighborhood deprivation index and provider geocoding identify critical catchment areas for diabetes outreach. J Clin Endocrinol Metab. 2020 Sep 1;105(9).

14. Krittanawong C, Virk HUH, Bangalore S, Wang Z, Johnson KW, Pinotti R, et al. Machine learning prediction in cardiovascular diseases: a meta-analysis. Sci Rep. 2020 Sep 29;10(1):16057.

15. Benedetto U, Dimagli A, Sinha S, Cocomello L, Gibbison B, Caputo M, et al. Machine learning improves mortality risk prediction after cardiac surgery: Systematic review and meta-analysis. J Thorac Cardiovasc Surg. 2020 Aug 10;

16. Hu L, Liu B, Ji J, Li Y. Tree-Based Machine Learning to Identify and Understand Major Determinants for Stroke at the Neighborhood Level. J Am Heart Assoc. 2020 Nov 3;e016745.

17. Ji J, Hu L, Liu B, Li Y. Identifying and assessing the impact of key neighborhood-level determinants on geographic variation in stroke: a machine learning and multilevel modeling approach. BMC Public Health. 2020 Nov 7;20(1):1666.

18. Angier H, Jacobs EA, Huguet N, Likumahuwa-Ackman S, Robert S, DeVoe JE. Progress towards using community context with clinical data in primary care. Family Med Commun Hlth. 2019;7(1):e000028.

19. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019 Feb 11;110:12–22.

20. Walsh CG, Ribeiro JD, Franklin JC. Predicting risk of suicide attempts over time through machine learning. Clinical Psychological Science. 2017 May;5(3):457–469.

21. Wang W, Kiik M, Peek N, Curcin V, Marshall IJ, Rudd AG, et al. A systematic review of machine learning models for predicting outcomes of stroke with structured data. PLoS One. 2020 Jun 12;15(6):e0234722.

22. Wellner B, Grand J, Canzone E, Coarr M, Brady PW, Simmons J, et al. Predicting unplanned transfers to the intensive care unit: A machine learning approach leveraging diverse clinical elements. JMIR Med Inform. 2017 Nov 22;5(4):e45.

23. Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Sci Rep. 2016 May 17;6:26094.

24. Forthman KL, Colaizzi JM, Yeh H-W, Kuplicki R, Paulus MP. Latent Variables Quantifying Neighborhood Characteristics and Their Associations with Poor Mental Health. Int J Environ Res Public Health. 2021 Jan 29;18(3).

25. Bian J, Buchan I, Guo Y, Prosperi M. Statistical thinking, machine learning. J Clin Epidemiol. 2019 Aug 16;116:136–137.

26. Wang F. Why public health needs GIS: A methodological overview. Ann GIS. 2020;26(1):1–12.

27. Prosperi M, Min JS, Bian J, Modave F. Big data hurdles in precision medicine and precision public health. BMC Med Inform Decis Mak. 2018 Dec 29;18(1):139.

28. Weaver A, Lapidos A. Mental Health Interventions with Community Health Workers in the United States: A Systematic Review. J Health Care Poor Underserved. 2018;29(1):159–180.

29. Bidargaddi N, Schrader G, Klasnja P, Licinio J, Murphy S. Designing m-Health interventions for precision mental health support. Transl Psychiatry. 2020 Jul 7;10(1):222.

30. Campion J, Knapp M. The economic case for improved coverage of public mental health interventions. Lancet Psychiatry. 2018;5(2):103–105.

31. Roffo G, Melzi S. Features selection via eigenvector centrality. Proceedings of new frontiers in mining complex patterns (NFMCP 2016)(Oct 2016). 2016;

32. Sanz H, Valim C, Vegas E, Oller JM, Reverter F. SVM-RFE: selection and visualization of the most relevant features through non-linear kernels. BMC Bioinformatics. 2018 Nov 19;19(1):432.

33. Chandrashekar G, Sahin F. A survey on feature selection methods. Computers & Electrical Engineering. 2014 Jan;40(1):16–28.

34. Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell. 2005 Aug;27(8):1226–1238.

35. Xanthopoulos P, Pardalos PM, Trafalis TB. Linear Discriminant Analysis. Robust Data Mining. New York, NY: Springer New York; 2013. p. 27–33.

36. Song F, Guo Z, Mei D. Feature selection using principal component analysis. 2010 International Conference on System Science, Engineering Design and Manufacturing Informatization. IEEE; 2010. p. 27–30.

37. Saqib P, Qamar U, Aslam A, Ahmad A. Hybrid of Filters and Genetic Algorithm - Random Forests Based Wrapper Approach for Feature Selection and Prediction. In: Arai K, Bhatia R, Kapoor S, editors. Intelligent computing: proceedings of the 2019 computing conference, volume 2. Cham: Springer International Publishing; 2019. p. 190–199.

38. Breiman L. Random Forests. Springer Science and Business Media LLC. 2001;

39. Fawagreh K, Gaber MM, Elyan E. Random forests: from early developments to recent advancements. Systems Science & Control Engineering. 2014 Dec;2(1):602–609.

40. Sandri M, Zuccolotto P. Variable selection using random forests. In: Zani S, Cerioli A, Riani M, Vichi M, editors. Data analysis, classification and the forward search. Berlin, Heidelberg: Springer Berlin Heidelberg; 2006. p. 263–270.

41. Díaz-Uriarte R, Alvarez de Andrés S. Gene selection and classification of microarray data using random forest. BMC Bioinformatics. 2006 Jan 6;7:3.

42. Hapfelmeier A, Ulm K. A new variable selection approach using Random Forests. Comput Stat Data Anal. 2013 Apr;60:50–69.

43. Nicodemus KK, Malley JD, Strobl C, Ziegler A. The behaviour of random forest permutation-based variable importance measures under predictor correlation. BMC Bioinformatics. 2010 Feb 27;11:110.

44. Archer KJ, Kimes RV. Empirical characterization of random forest variable importance measures. Comput Stat Data Anal. 2008 Jan;52(4):2249–2260.

45. Wongvibulsin S, Wu KC, Zeger SL. Clinical risk prediction with random forests for survival, longitudinal, and multivariate (RF-SLAM) data analysis. BMC Med Res Methodol. 2019 Dec 31;20(1):1.

46. Sariyar M, Hoffmann I, Binder H. Combining techniques for screening and evaluating interaction terms on high-dimensional time-to-event data. BMC Bioinformatics. 2014 Feb 26;15:58.

47. Yang W, Gu CC. Selection of important variables by statistical learning in genome-wide association analysis. BMC Proc. 2009;3(Suppl 7):S70.

48. Kim Y, Wojciechowski R, Sung H, Mathias RA, Wang L, Klein AP, et al. Evaluation of random forests performance for genome-wide association studies in the presence of interaction effects. BMC Proc. 2009 Dec 15;3 Suppl 7:S64.

49. Winham SJ, Colby CL, Freimuth RR, Wang X, de Andrade M, Huebner M, et al. SNP interaction detection with Random Forests in high-dimensional genetic data. BMC Bioinformatics. 2012 Jul 15;13:164.

50. Lunetta KL, Hayward LB, Segal J, Van Eerdewegh P. Screening large-scale association study data: exploiting interactions using random forests. BMC Genet. 2004 Dec 10;5:32.

51. Lynn Speiser J. A random forest method with feature selection for developing medical prediction models with clustered and longitudinal data. J Biomed Inform. 2021 Mar 26;103763.

52. Speiser JL, Miller ME, Tooze J, Ip E. A comparison of random forest variable selection methods for classification prediction modeling. Expert Syst Appl. 2019 Nov 15;134:93–101.

53. Boulesteix A-L, Janitza S, Kruppa J, König IR. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. WIREs Data Mining Knowl Discov. 2012 Nov;2(6):493–507.

54. Xiao M, Yan C, Fu B, Yang S, Zhu S, Yang D, et al. Risk prediction for postpartum depression based on random forest. Zhong Nan Da Xue Xue Bao Yi Xue Ban. 2020 Oct 28;45(10):1215–1222.

55. Faramawi MF, Abouelenein S, Johnson E. A case-control study of occupational risk factors for pancreatic cancer in poultry plant workers: a random forest approach. J Public Health. 2021 Feb 25;

56. Wang X, Zhai M, Ren Z, Ren H, Li M, Quan D, et al. Exploratory study on classification of diabetes mellitus through a combined Random Forest Classifier. BMC Med Inform Decis Mak. 2021 Mar 20;21(1):105.

57. Li Y, Zhang L, Zhang Y, Wen H, Huang J, Shen Y, et al. A Random Forest Model for Predicting Social Functional Improvement in Chinese Patients with Schizophrenia After 3 Months of Atypical Antipsychotic Monopharmacy: A Cohort Study. Neuropsychiatr Dis Treat. 2021 Mar 19;17:847–857.

58. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet. 2012 Jan 13;90(1):7–24.

59. Nau C, Ellis H, Huang H, Schwartz BS, Hirsch A, Bailey-Davis L, et al. Exploring the forest instead of the trees: An innovative method for defining obesogenic and obesoprotective environments. Health Place. 2015 Sep 19;35:136–146.

60. Basu S, Siddiqi A. Geographic disparities in US mortality: “hot-spotting” large databases. Epidemiology. 2014 May;25(3):468–470.

61. Hotelling H. Analysis of a complex of statistical variables into principal components. J Educ Psychol. 1933;24(6):417–441.

62. Pearson K. On lines and planes of closest fit to systems of points in space. Philosophical Magazine Series 6. 1901 Nov;2(11):559–572.

63. Zhang Z, Castelló A. Principal components analysis in clinical studies. Ann Transl Med. 2017 Sep;5(17):351.

64. Wood MD, Simmatis LER, Gordon Boyd J, Scott SH, Jacobson JA. Using principal component analysis to reduce complex datasets produced by robotic technology in healthy participants. J Neuroeng Rehabil. 2018 Jul 31;15(1):71.

65. Kandel BM, Wang DJJ, Gee JC, Avants BB. Eigenanatomy: sparse dimensionality reduction for multi-modal medical image analysis. Methods. 2015 Feb;73:43–53.

66. Giuliani A. The application of principal component analysis to drug discovery and biomedical data. Drug Discov Today. 2017 Jan 19;22(7):1069–1076.

67. Eyal E, Bloch BN, Rofsky NM, Furman-Haran E, Genega EM, Lenkinski RE, et al. Principal component analysis of dynamic contrast enhanced MRI in human prostate cancer. Invest Radiol. 2010 Apr;45(4):174–181.

68. Peres-Neto PR, Jackson DA, Somers KM. GIVING MEANINGFUL INTERPRETATION TO ORDINATION AXES: ASSESSING LOADING SIGNIFICANCE IN PRINCIPAL COMPONENT ANALYSIS. Ecology. 2003 Sep;84(9):2347–2363.

69. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philos Trans A, Math Phys Eng Sci. 2016 Apr 13;374(2065):20150202.

70. Groth D, Hartmann S, Klie S, Selbig J. Principal components analysis. Methods Mol Biol. 2013;930:527–547.

71. Ma S, Dai Y. Principal component analysis based methods in bioinformatics studies. Brief Bioinformatics. 2011 Nov;12(6):714–722.

72. Akay MF. Support vector machines combined with feature selection for breast cancer diagnosis. Expert Syst Appl. 2009 Mar;36(2):3240–3247.

73. Escanilla NS, Hellerstein L, Kleiman R, Kuang Z, Shull JD, Page D. Recursive feature elimination by sensitivity testing. Proc Int Conf Mach Learn Appl. 2018 Dec;2018:40–47.

74. Peng Y, Wu Z, Jiang J. A novel feature selection approach for biomedical data classification. J Biomed Inform. 2010 Feb;43(1):15–23.

75. Djellali H, Zine NG, Azizi N. Two stages feature selection based on filter ranking methods and SVMRFE on medical applications. In: Chikhi S, Amine A, Chaoui A, Kholladi MK, Saidouni DE, editors. Modelling and implementation of complex systems. Cham: Springer International Publishing; 2016. p. 281–293.

76. Mao Y, Pi D, Liu Y, Sun Y. Accelerated recursive feature elimination based on support vector machine for key variable identification. Chin J Chem Eng. 2006 Feb;14(1):65–72.