Ecological Factors Associated with Self-Reported Mental Health Status

Alyssa Berger, Andrew Cistola

Abstract

Background: Ecological factors are known predictors of many health outcomes. However, there are many complex ecological factors that have not yet been thoroughly identified and spatial analysis used for non-infectious public health outcomes, particularly for mental health, is limited. ***Aims:*** We aimed to utilize machine learning and spatial statistics to identify significant search among many possible zip code or county predictors (N > 2000). We also aimed to identify informative associations between ecological variables and mental health outcomes available from public data sources (CDC Places). This study seeks to identify local socioeconomic factors and area health resources associated with poor self-reported mental health status. ***Methods:*** This study combines feature selection, spatial regression, to identify a small set of important variables representing possible effects. Hierarchical linear models and artificial neural networks to identify the effects and their predictive ability. ***Results:*** The approach identified a final model consisting of (N = 5) zip code variables related to families, education, or employment. Designation as manufacturing dependent accounted for a significant variation in poor mental health (71% based on ICC). Based on the presence of statistically significant, meaningful, and informative results developed without direct researcher engagement with the content, this process of "algorithmic triangulation" was relatively effective. ***Conclusions:*** Types of employment and educational opportunities may be related to population level mental health, but more research is needed. This methodology can be applied to many other outcomes for the purpose of generating hypotheses that can be further investigated with relevant domain area knowledge.

Ecological Factors Associated with Self-Reported Mental Health Status

Introduction

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). However, there are many complex ecological factors that have not yet been thoroughly identified. With the advent of 'big data,' there is increased availability of ecological data that can measure these effects as well as machine learning algorithms that can analyze large datasets.

Alongside geographic information systems that can model complex associations, new computational approaches have great potential to improve public health research and population health management efforts (14–17). However spatial analysis used for non-infectious public health outcomes is limited (18) and the use of machine learning often provides little benefit over traditional approaches (19,20). Machine learning needs methodological improvement (21–24) to better assist population health management efforts (25–27). There is a need for expanded scope in spatial analysis to better inform translational science in public health (28–30).

Purpose of this Study

The purpose of this study is to utilize machine learning and spatial statistics to identify significant associations between ecological variables and health outcomes available from public data sources. This study seeks to identify local socioeconomic factors associated with a higher prevalence and health resources that are associated with a lower prevalence of poor self-reported mental health status. Once identified, hierarchical linear models can be developed to display these relationships and artificial neural networks can be used to predict future outcomes. By searching among large numbers of candidate predictors, informative relationships that may not be already considered in previous research can be used to develop targeted future population health management efforts.

Methods

Due to the inability for traditional statistical approaches to search among candidate predictors in multi-dimensional data, this study has developed a novel approach that combines several previously established methods for feature selection, spatial regression, hierarchical linear modeling, and artificial neural networks.

Together these methods provide a process for hypothesis generating observational research able to identify candidate predictors, adjust for spatial relationships, model multi-level interactions, predict future outcomes. This process consists of four steps: feature selection, spatial regression, hierarchical linear modeling, and artificial neural networks.

Feature Selection

Feature selection (FS) 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). This study utilizes an approach that combines embedded (RF), filter (PCA), and wrapper methods (SVM-RFE) together to identify the smallest set of predictors that have both high variation and high importance in the context of all other candidates.

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). RF were then used to identify predictors with above average importance measured by Gini impurity. This allowed for variables to be independently evaluated for predictive ability that adjusts for collinearity and confounding.

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). For variables at the zip code level, PCA was used to identify predictors with above average eigenvectors for all components with an eigenvalue above 1.0. This allowed for predictors with higher variation to be identified from among possible candidates.

SVM-RFE is a simple approach to feature selection that identifies a reduced set of predictors that provide the highest prediction accuracy through iterative cross validation (72–75). SVM-RFE can be deployed in similar settings to stepwise regression but avoids the many established drawbacks by utilizing Support Vector Machines which requires minimal assumptions related to the distribution of the data to be met (76). In this study, predictors that had both above average variation (from the PCA model) and importance (from the RF model) were selected for RFE-SVM with cross validation. This process calculates AIC values for all possible combinations of selected predictors and identified the best possible selection with the lowest number of candidate predictors.

Spatial Regression

Since health outcomes and ecological predictors may exhibit geographic similarity, regression methods that display these differences are used to identify areas where multi-level effects may be present. These models use features identified from the feature selection process at some geographic layers and will inform feature selection at other geographic layers.

Multi-scale geographic weighted regression (GWR) is spatial regression (SR) for identifying spatial variations in an independent-variable’s estimated effect on the outcome of interest. Rather than globally identifying parameter estimates, GWR modifies parameters for each individual area based on the relationship to other spatially adjacent areas. This provides for areas to be identified where a given independent variable has a greater or lesser effect on the same outcome, indicating that higher level effects may be present.

The list of zip code variables from the FS process is used for GWR and areas where zip code level predictors had higher predictive ability are identified. These areas represent locations where infrastructural differences have a possible interaction with socio-economic factors leading to an increased effect on the outcome. To identify these interactions, each county is coded with a multi-level categorical variable indicating which zip code variable had the highest GWR coefficient. Due to the ability for SVM to handle a low number of observations with high variable counts, SVM is used to identify county level predictors from the county predictors with the highest coefficients for each zip code variable. These county-zip code predictor pairs are kept as possible multi-layer interactions. Using local Empirical Bayes smoothing and local Moran's test for spatial autocorrelation (LISA) quadrants, 'hot and cold spot' regions are created that represent significant autocorrelation in the outcome. These regions are overlayed to county boundaries and provided with nominal labels. SVM are used to identify independent county level predictors.

Hierarchical Linear Modeling

After significant features are identified and spatial relationships are quantified, a hierarchical linear model (HLM) model is used to identify parameter estimates for ecological predictors that can illustrate the relationship between variables with statistical tests.

Artificial Neural Networks

In order to compare the selected predictors used in the HLM model to all candidate predictors, a multi-layered perceptron (MLP) with an artificial neural network (ANN) architecture is used to predict a zip code being in the 75th percentile for the outcome. An MLP using all possible candidate predictors, a random selection of candidate predictors, and an MLP for the selected multi-level predictors that are included in the HLM are both trained and tested on a 50-50 random split on the original data (with standard scaling and missing data imputation). A Receiver Operator Curve (ROC) and the corresponding C-statistics are used to compare the accuracy of each variable set.

During the FS, SR, and ANN processes, all candidate predictors with less than 75% non-missing data are removed and the remaining missing values are imputed with median values. During the HLM process, a significantly smaller set of predictors are used and observations with missing values are removed due to low infrequency.

Each of the machine learning, spatial regression, and traditional statistical models used open-source libraries available in the Python Programming Language (Python) version 3.6 and R Project for Statistical Computing (R) version 3.0. The ANNs utilized the Keras API for TensorFlow v2.4.1. Datasets and code scripts are available as under open-source license from GitHub for the purpose of dissemination and replication.

Data Collection

US Census American Community Survey (ACS) five-year percent estimates by zip code are collected from the 2020 data release to represent socio-economic ecological predictors. The Health Services Resource Administration (HRSA) Area Health Resource File (AHRF) 2020 data release is collected and five-year averages are calculated by county to represent infrastructural ecological predictors. The CDC and RWJF PLACES 2020 data release (formerly 500 Cities project) provided zip code level prevalence estimates of many common diseases, health conditions, and preventive care services. Zip codes were joined to county using crosswalk files from the Department of Housing and Urban development so that each zip code could be contained within a specific county based on where a majority of the population is located. Together these datasets provide a comprehensive list of approximately 2500 possible candidate predictors at zip code and county levels for all 50 states and include individual demographics, housing types, economic status, education levels, employment types, health workforce availability, insurance markets, and inpatient facilities, and hospital quality metrics.

For each of the datasets, the 2020 release was created based on data collected during 2014-2019. This time period was selected due to its timing immediately after the most significant period of ACA implementation and preceding the SARS-Cov20 pandemic. When compared to other time periods, this 5-year window represents a relatively stable healthcare environment that can be useful in determining the direction for future reforms.

Results

Self-reported mental health status from (N = 969) zip codes in Florida were collected from the CDC PLACES dataset (Figure 1) and descriptive statistics were calculated for crude rate per 1000 residents (quartiles = 0, 128, 149, 146, 167). At the zip code level (N = 366) socioeconomic variables were collected from the ACS and (N = 28) were collected from the CDC. At the county level (N = 1802) variables were collected from the AHRF.

A generalized GWR model was created using the natural log of the crude rate (Adj. R-squared = 0.9577, F-statistic = 10970, p-value < 0.001, AIC -3985) with composite index of both social variables from the ACS (Estimate = 0.005, p < 0.001) and health variables from the CDC (Estimate: 0.021, p < 0.001). The weighted coefficients for local health (Figure 2) were calculated, normalized, and averaged among zip codes by each (N = 67) county (quartiles = 18.38, 18.90, 19.45, 20.12, 22.87). Gini coefficients for each county (Figure 2) were calculated (quartiles = 0.00, 0.03, 0.05, 0.07, 0.09).

Using PCA, RF, and RFE-SVM a set of (N = 21) predictors were selected based on predicting the crude rate (Figure 3). Using SVM, (N = 10) county level predictors were selected based on predicting the different county level rates (Figure 3). The selected zip code predictors were used to create a simple ordinary least squares (OLS) model for crude rates (Figure 4). OLS assumptions were checked by Utt's rainbow test (p < 0.001, indicating non-linearity), the Jarque-Bera test (p < 0.001, indicating non-normal errors), the Anderson-Darling test (p < 0.001, indicating non-normal errors), the Durbin-Watson test (p = 0.761, indicating no generic autocorrelation), the Moran's I test (p < 0.001, indicating spatial autocorrelation), the Breusch-Pagan test (p < 0.001, indicating homoskedasticity), the Goldfield-Quandt test (p = 0.995, indicating no heteroskedasticity). A Q-Q plot and residuals plot (Figure 5) were further used to confirm non-normality and homoscedasticity within the sample. A correlation matrix (Figure 6) was used to identify colinear variables with a Spearman's rank r above 0.5 (N = 7).

These results show that the homoscedasticity, and independent sampling assumptions were met, but non-normality, non-linearity, spatial autocorrelation, and collinearity was present. Multiple generalized linear models (GLM) were created using different approaches to addressing these concerns and compared. A GLM was created for the standardized rate (SMR) with an identity link and normal distribution (AIC = 1304.5), the crude rate with log-link and normal distribution (AIC = 1152.8), the binary outcome (SMR > 1) with a logit link and binomial distribution (AIC = 547.3), the crude rate with a log-link and Poisson distribution (AIC = 10219.0), the crude rate with a log-link and a negative binomial distribution (AIC = 9401.4). The most simple and accurate approach for addressing non-normality was a GLM with a natural log transformation of the crude rate. This model was implemented along within a hierarchical linear model to further account for spatial autocorrelation.

A HLM was created for the natural log crude rate assuming random-effects with fixed coefficients and varying intercepts for each selected county variable. Zip code predictors for the HLM models were excluded if they were collinear (r > 0.5) and less "important" based on scores from the RF model. (N = 1) county variables were significant (alpha = 0.001) based on a One-Way ANOVA test when models with or without the randomized terms were included and (N = 5) zip code variables had significant parameter estimates in each HLM model.

A final model using these selected predictors (Figure 7) was created (AIC = 1283.3, log-likelihood = -633.65) and the adjusted intraclass correlation coefficient was calculated (ICC = 0.714). A MLP with 50-50 random split was trained and used to calculate the C-statistic for all predictors (AUC = 0.91, epochs = 100, loss < 0.01) and the selected predictors (AUC = 0.87, epochs = 100, loss = 0.09).

Immediately upon download (using public APIs) all predictors were given generic labels and blinded to the software user until the above results were collected. Once the final HLM and MLPs were created, the feature labels were used to identify each of the selected zip code level predictor with relevant definitions. This allowed for the final selection to be chosen without bias from the researcher and rely entirely on quantitative assessment. The list of selected zip code predictors (Table 1) with significant parameter estimates in the final model included: average household size (B = -0.075, 95CI = -0.060 to -0.090, p < 0.001), population 25 years and over with a bachelor's degree or higher (B = -0.159, 95CI = -0.141 to -0.176, p < 0.001), total population female (B = 0.249, 95CI = 0.235 to 0.264, p < 0.001), percent of civilian employed population as government workers (B = 0.093, 95CI = 0.078 to 0.108, p < 0.001). The final list of selected county predictors (Table 1) with significant ANOVA tests comparing the presence of a random effect in the model included: manufacturing dependent designation ($p < 0.001$).

Discussion

The purpose of this study was to explore social determinants of health resources associated with self-reported poor mental health status in Florida. A novel approach was utilized which included methodology incorporating machine learning and spatial statistics to search among many possible predictors of hospital discharges. This methodology sought to identify informative associations regarding mental health outcomes that may not be considered in previous research excluding spatial data. Additionally, there were a variety of quantitative methods applied to minimize intrinsic biases among various modeling techniques with a process commonly deployed in qualitative research.

Overall, findings from this study discovered many variables that reflect previous findings, but also variables that provide greater nuance on social factors associated with mental health status. Evidence from this study can be used for hypothesis testing and intervention design for population health management. Based on the presence of statistically significant, meaningful, and informative results developed without direct researcher engagement with the content, this process of “algorithmic triangulation” was relatively effective. Due to the automation involved in the process, this methodology can be applied to many other outcomes for the purpose of generating hypotheses that can be further investigated with relevant domain area knowledge.

The most significant finding in the study was the level of variation in poor mental health (71% based on ICC) attributed to being labeled as manufacturing dependent. Although this is not applicable to all studies, it can provide relevant information when assessing for accessibility and availability of resources in areas of target populations. These findings may also raise important questions regarding finances and economics within residence of these target populations. Types of employment and educational opportunities may be related to population level mental health, but more research is needed to identify these possible effects. Despite more research being needed to draw these direct conclusions and direct causations, studies such as these are useful in obtaining spatial information that may not be known form assessing the literature alone or from talking to community members alone.

Limitations

There are several important limitations to acknowledge, particularly in relation to the exploratory nature of this study. Since this study did not seek to test a hypothesis, involve longitudinal data, or utilize any experimental (or quasi-experimental) designs the results should be interpreted carefully. However, it is important to note the value in hypothesis generating studies for mental health, especially in connection to ecological data. Although this study was broad in nature, it is a first step in further understanding how spatial data can benefit psychiatric epidemiologic research.

The final model is a reflection of what possible effects may be present when these other research processes take place. These results are designed to inform future studies and not to replace or contradict them. Domain knowledge and researcher input is necessary for these results to help translational research. The goal of this study was to simply identify through advanced computational techniques what effects may be present, but not yet fully investigated for mental health outcomes. This hopefully can provide a starting point for future innovative research.

Due to the level of precision in the selected variables, there is a temptation to explain each one as a proxy for broader effects (age, neighborhood deprivation, poverty etc). However, the datasets included robust measures for all general variables and the algorithms were designed to account for each variable in context through multiple and repeated ways. If variables are not fully represented in the final selection, there is strong evidence that they may not be representative of the true possible effect. These would need to be further verified by future study, but dismissal due to other factors considered by the algorithms would not be accurate.

Due to the fact that many different methods were used in the process of "algorithmic triangulation", there is a possibility that a different combination may yield more informative results. Each algorithm or model was chosen due to their previously validated uses and mathematical relevance to the research question, however this exact combination has not been yet utilized.

Florida was chosen as a focus area for two reasons: 1) It is the home state of the research institution and 2) The computational resources required to conduct analysis among multiple states far exceed the capability of everyday devices (The analysis was conducted on an Intel i7 CPU with 16GB DDRAM running Ubuntu 20.04.

Some algorithms took over a week to run when using data from all 50 states). GPUs have shown great potential in advancing processing speed for many algorithms that benefit from the parallel structure used for graphics (RF, ANN, GWR) and NVIDIA's RAPIDS architecture provides a Python library capable of utilizing the same algorithms on their GPUs. If the code scripts could be adapted to GPU use, access to resources that are commercially available as 'gaming' devices (or furthermore advanced research computers) would provide a significant boost in speed allowing for a much wider analysis.

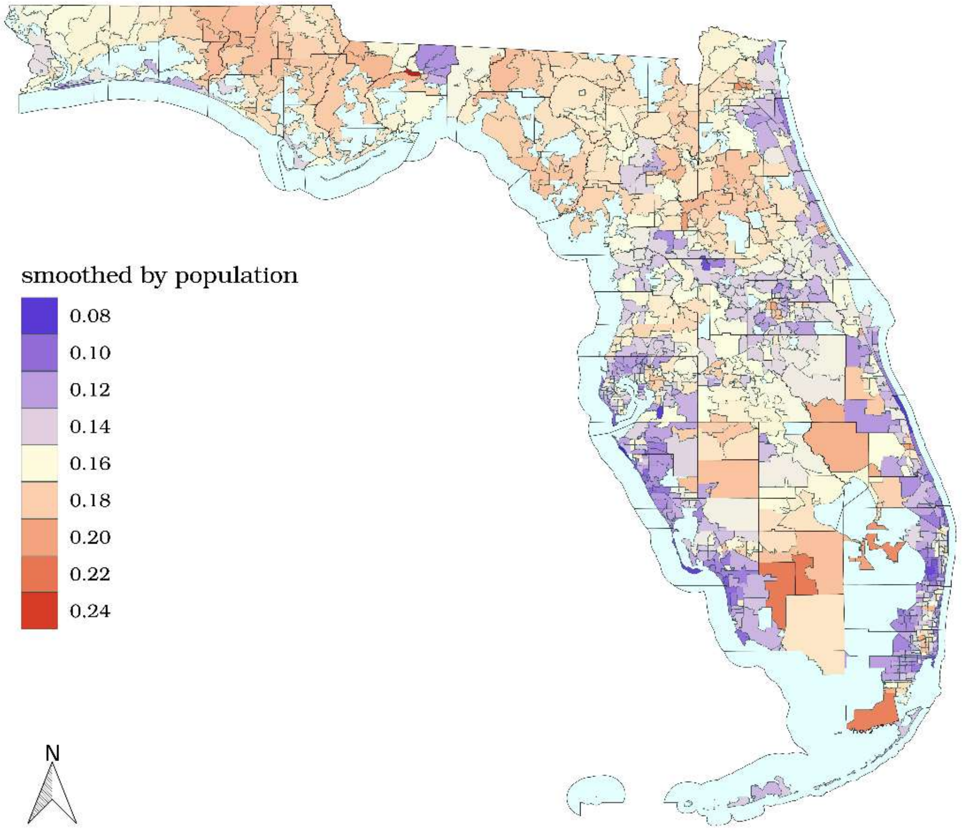
Conclusion

When considering potential social determinants of health and associated resources in Florida, there is significant evidence that types of employment and educational opportunities may be related to population level mental health, but more research is needed to identify these possible effects. Future research should use these results to more specifically identify why there is such variation in self-reported mental health. Understanding the context behind this quantitative data will better inform hospitals, neighborhoods, and public health professionals on how to collaborate in creating enhanced targeted interventions to overall improve population health and other potential impeding financial constraints.

Tables and Figures

Table I. Final list of selected zip code and county predictors used in mixed-effects model using natural log of crude prevalence as outcome.

Feature	Layer	Description
D2_T1_17	Zip Code	Average Household Size
D2_T1_46	Zip Code	Population 25 Years And Over Bachelor's Degree Or Higher
D2_T2_39	Zip Code	Civilian Employed Population Government Workers
D2_T2_90	Zip Code	Percentage of Families with Income Below The Poverty Level
D2_T4_5	Zip Code	Total Population Female
D3_1029	County	Manufacturing-Dependent Designation



PLACES (2020) CDC https://www.cdc.gov/places/index.html Accessed: 2021-04-19

Figure 1. Estimated zip code prevalence of self-reported poor mental health using local empirical bayes smoothing.

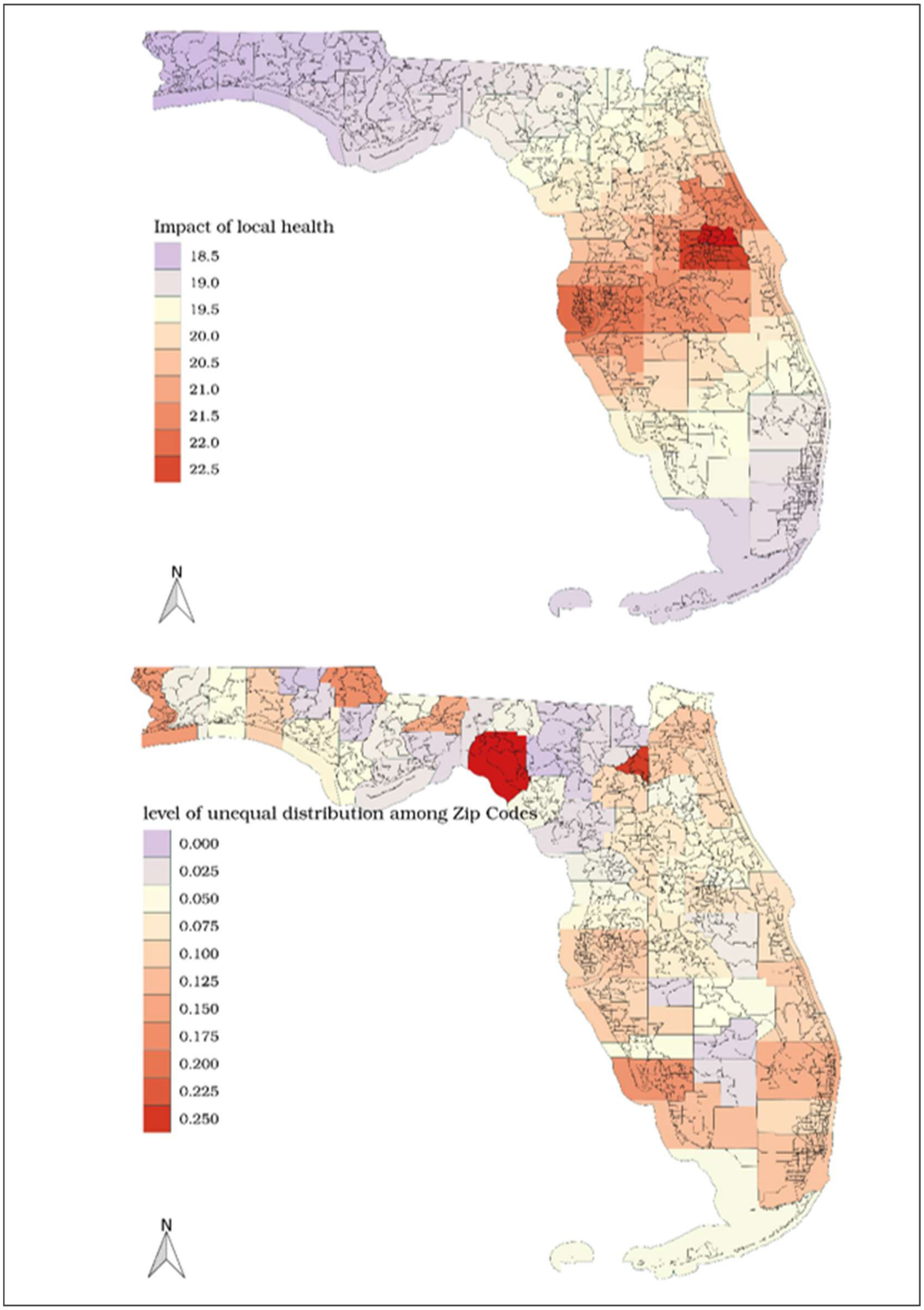
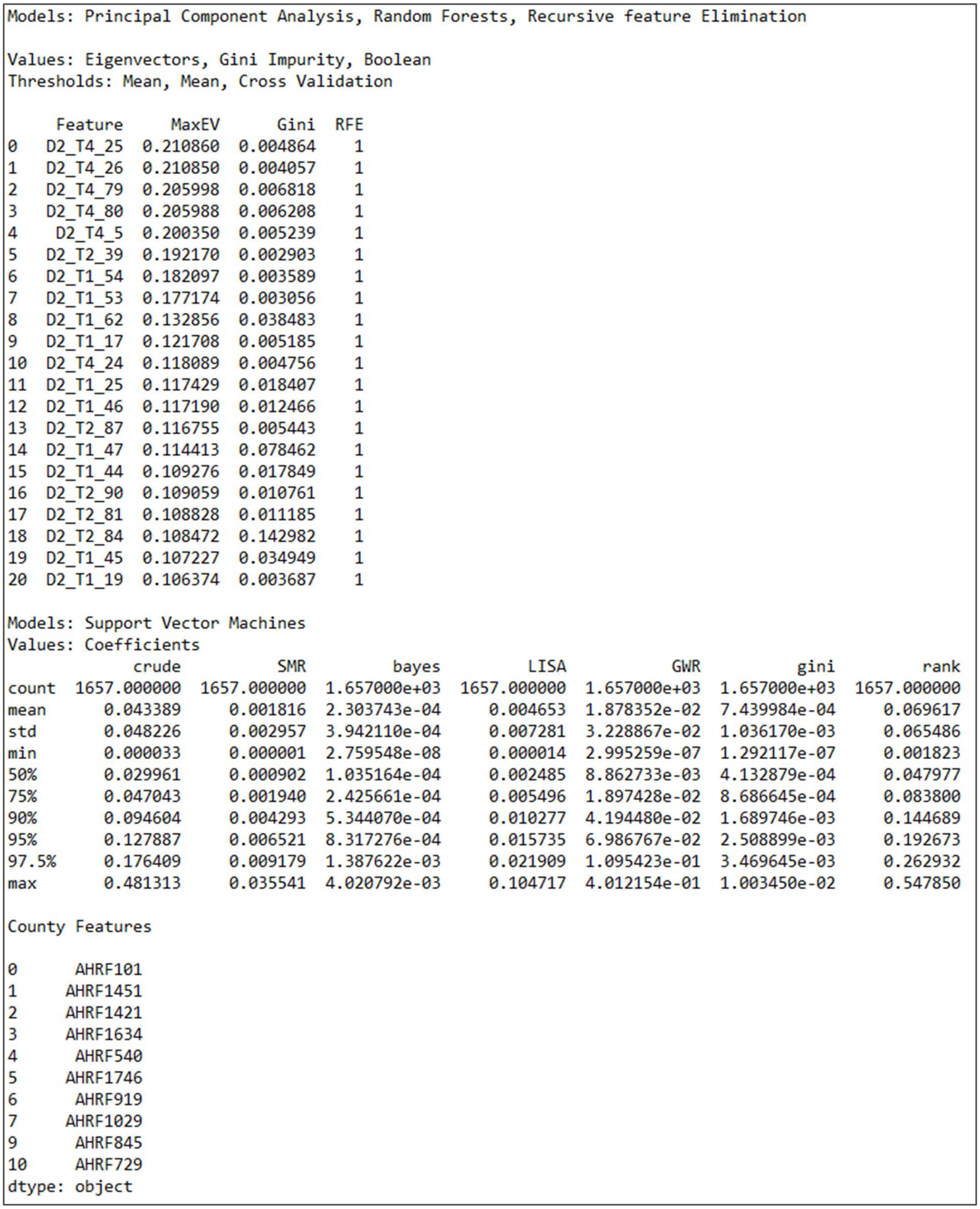
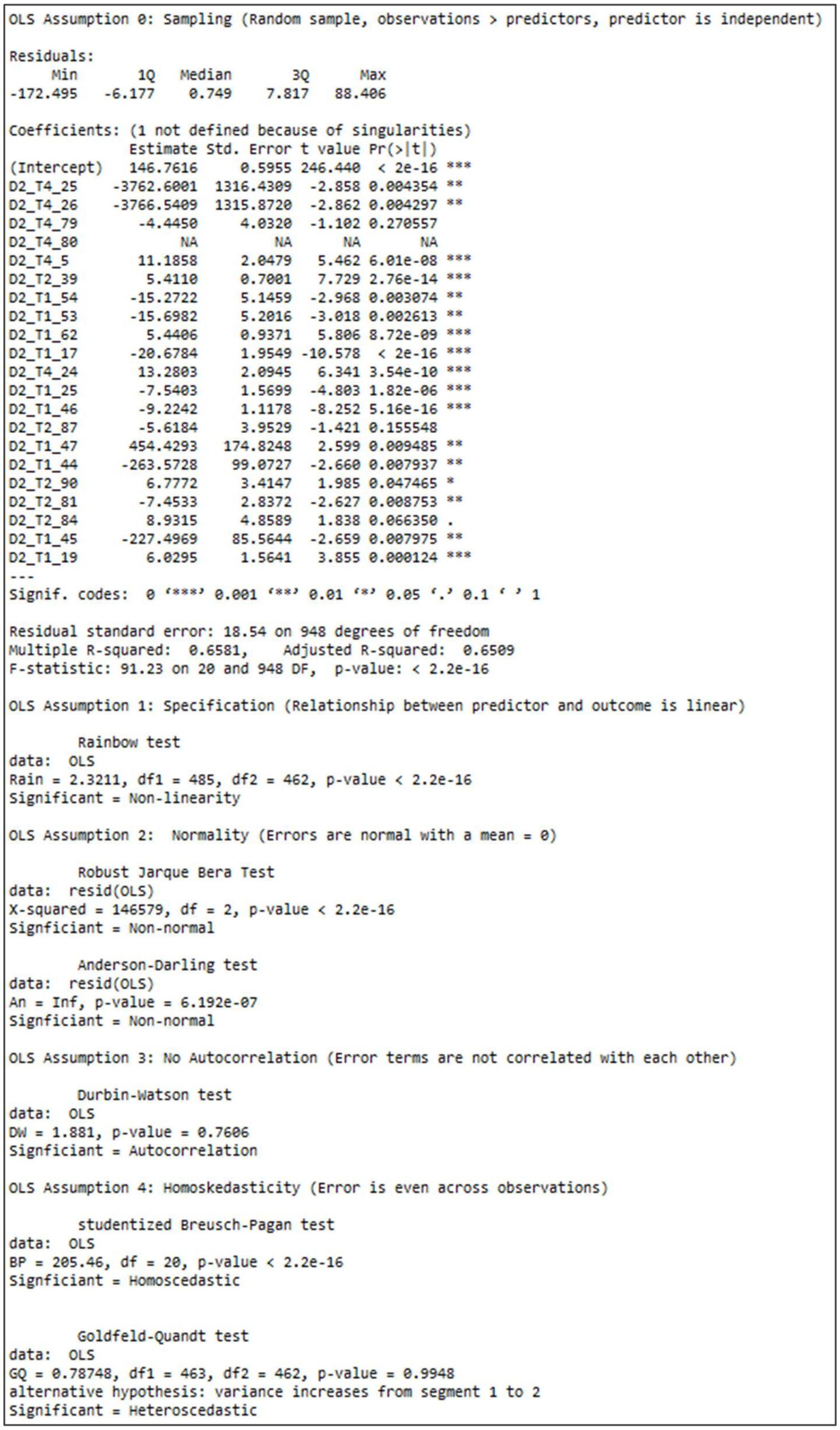


Figure 2. (above) coefficient variation of aggregated local health status in a geographically weighted regression model (below) Gini index measuring level of unequal distribution of higher rates among zip codes within counties.



```
Models: Principal Component Analysis, Random Forests, Recursive feature Elimination

Values: Eigenvectors, Gini Impurity, Boolean
Thresholds: Mean, Mean, Cross Validation

     Feature     MaxEV      Gini  RFE
0   D2_T4_25  0.210860  0.004864    1
1   D2_T4_26  0.210850  0.004057    1
2   D2_T4_79  0.205998  0.006818    1
3   D2_T4_80  0.205988  0.006208    1
4    D2_T4_5  0.200350  0.005239    1
5   D2_T2_39  0.192170  0.002903    1
6   D2_T1_54  0.182097  0.003589    1
7   D2_T1_53  0.177174  0.003056    1
8   D2_T1_62  0.132856  0.038483    1
9   D2_T1_17  0.121708  0.005185    1
10  D2_T4_24  0.118089  0.004756    1
11  D2_T1_25  0.117429  0.018407    1
12  D2_T1_46  0.117190  0.012466    1
13  D2_T2_87  0.116755  0.005443    1
14  D2_T1_47  0.114413  0.078462    1
15  D2_T1_44  0.109276  0.017849    1
16  D2_T2_90  0.109059  0.010761    1
17  D2_T2_81  0.108828  0.011185    1
18  D2_T2_84  0.108472  0.142982    1
19  D2_T1_45  0.107227  0.034949    1
20  D2_T1_19  0.106374  0.003687    1

Models: Support Vector Machines
Values: Coefficients
             crude          SMR         bayes          LISA           GWR          gini         rank
count  1657.000000  1657.000000  1.657000e+03  1657.000000  1.657000e+03  1.657000e+03  1657.000000
mean      0.043389     0.001816  2.303743e-04     0.004653  1.878352e-02  7.439984e-04     0.069617
std       0.048226     0.002957  3.942110e-04     0.007281  3.228867e-02  1.036170e-03     0.065486
min       0.000033     0.000001  2.759548e-08     0.000014  2.995259e-07  1.292117e-07     0.001823
50%       0.029961     0.000902  1.035164e-04     0.002485  8.862733e-03  4.132879e-04     0.047977
75%       0.047043     0.001940  2.425661e-04     0.005496  1.897428e-02  8.686645e-04     0.083800
90%       0.094604     0.004293  5.344070e-04     0.010277  4.194480e-02  1.689746e-03     0.144689
95%       0.127887     0.006521  8.317276e-04     0.015735  6.986767e-02  2.508899e-03     0.192673
97.5%     0.176409     0.009179  1.387622e-03     0.021909  1.095423e-01  3.469645e-03     0.262932
max       0.481313     0.035541  4.020792e-03     0.104717  4.012154e-01  1.003450e-02     0.547850

County Features

0      AHRF101
1     AHRF1451
2     AHRF1421
3     AHRF1634
4      AHRF540
5     AHRF1746
6      AHRF919
7     AHRF1029
9      AHRF845
10     AHRF729
dtype: object
```

Figure 3. Raw output from Python 'scikit-learn' library showing results from algorithms used in feature selection for zip codes (above) and counties (below). These features were selected for use in the ANN and HLM models.



```
OLS Assumption 0: Sampling (Random sample, observations > predictors, predictor is independent)

Residuals:
     Min       1Q   Median       3Q      Max 
-172.495   -6.177    0.749    7.817   88.406 

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   146.7616     0.5955 246.440  < 2e-16 ***
D2_T4_25    -3762.6001  1316.4309  -2.858 0.004354 ** 
D2_T4_26    -3766.5409  1315.8720  -2.862 0.004297 ** 
D2_T4_79       -4.4450     4.0320  -1.102 0.270557    
D2_T4_80            NA         NA      NA       NA    
D2_T4_5        11.1858     2.0479   5.462 6.01e-08 ***
D2_T2_39        5.4110     0.7001   7.729 2.76e-14 ***
D2_T1_54      -15.2722     5.1459  -2.968 0.003074 ** 
D2_T1_53      -15.6982     5.2016  -3.018 0.002613 ** 
D2_T1_62        5.4406     0.9371   5.806 8.72e-09 ***
D2_T1_17      -20.6784     1.9549 -10.578  < 2e-16 ***
D2_T4_24       13.2803     2.0945   6.341 3.54e-10 ***
D2_T1_25       -7.5403     1.5699  -4.803 1.82e-06 ***
D2_T1_46       -9.2242     1.1178  -8.252 5.16e-16 ***
D2_T2_87       -5.6184     3.9529  -1.421 0.155548    
D2_T1_47      454.4293   174.8248   2.599 0.009485 ** 
D2_T1_44     -263.5728    99.0727  -2.660 0.007937 ** 
D2_T2_90        6.7772     3.4147   1.985 0.047465 *  
D2_T2_81       -7.4533     2.8372  -2.627 0.008753 ** 
D2_T2_84        8.9315     4.8589   1.838 0.066350 .  
D2_T1_45     -227.4969    85.5644  -2.659 0.007975 ** 
D2_T1_19        6.0295     1.5641   3.855 0.000124 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.54 on 948 degrees of freedom
Multiple R-squared:  0.6581,    Adjusted R-squared:  0.6509 
F-statistic: 91.23 on 20 and 948 DF,  p-value: < 2.2e-16

OLS Assumption 1: Specification (Relationship between predictor and outcome is linear)

        Rainbow test
data:  OLS
Rain = 2.3211, df1 = 485, df2 = 462, p-value < 2.2e-16
Significant = Non-linearity

OLS Assumption 2:  Normality (Errors are normal with a mean = 0)

        Robust Jarque Bera Test
data:  resid(OLS)
X-squared = 146579, df = 2, p-value < 2.2e-16
Signficiant = Non-normal

        Anderson-Darling test
data:  resid(OLS)
An = Inf, p-value = 6.192e-07
Signficiant = Non-normal

OLS Assumption 3: No Autocorrelation (Error terms are not correlated with each other)

        Durbin-Watson test
data:  OLS
DW = 1.881, p-value = 0.7606
Signficiant = Autocorrelation

OLS Assumption 4: Homoskedasticity (Error is even across observations)

        studentized Breusch-Pagan test
data:  OLS
BP = 205.46, df = 20, p-value < 2.2e-16
Signficiant = Homoscedastic


        Goldfeld-Quandt test
data:  OLS
GQ = 0.78748, df1 = 463, df2 = 462, p-value = 0.9948
alternative hypothesis: variance increases from segment 1 to 2
Significant = Heteroscedastic
```

Figure 4. Raw model summary from R ‘stats’ library for initial OLS model using selected zip code predictors as well as test results checking OLS assumptions.

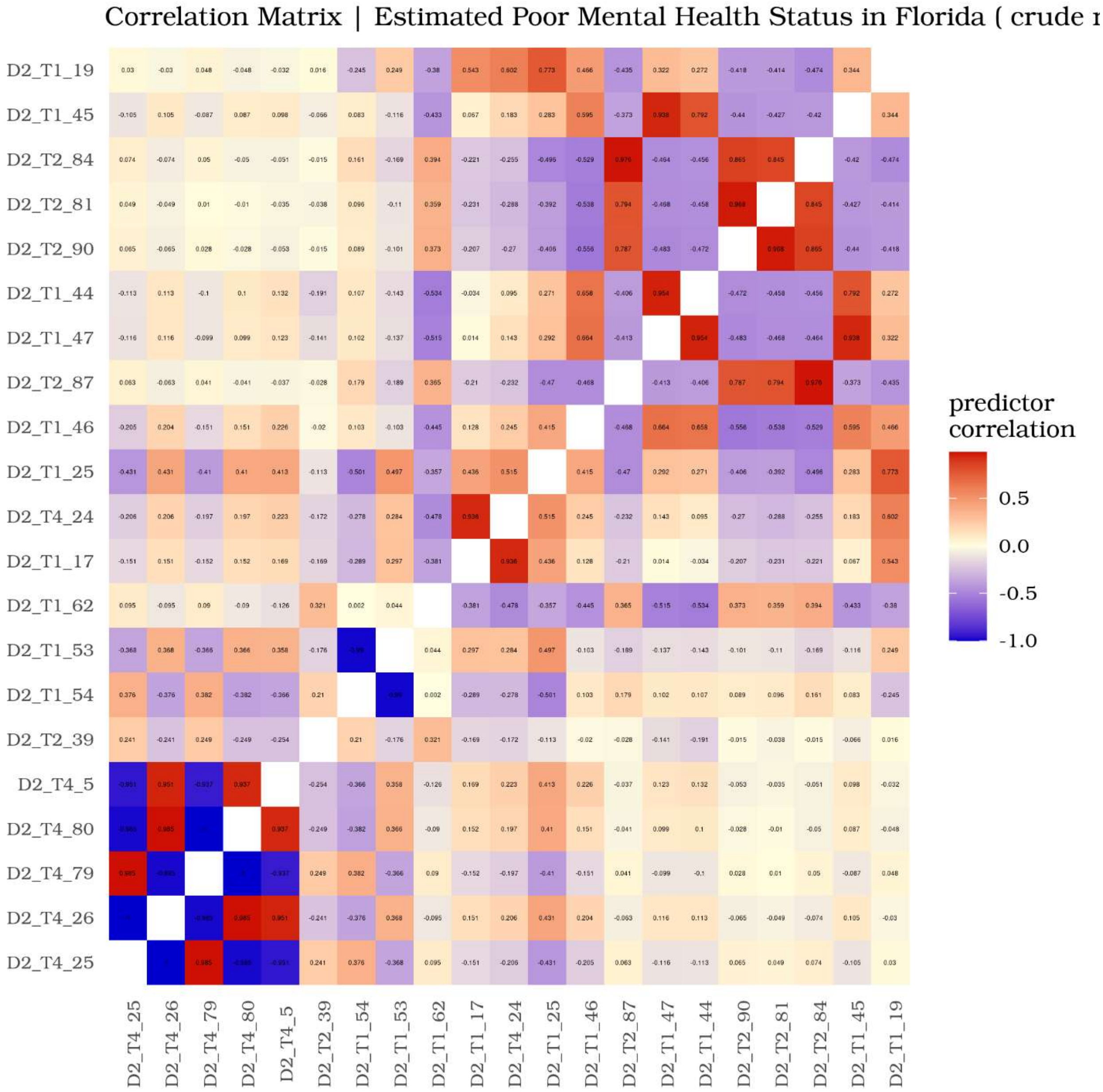


Figure 5. Correlation matrix of zip code level predictors

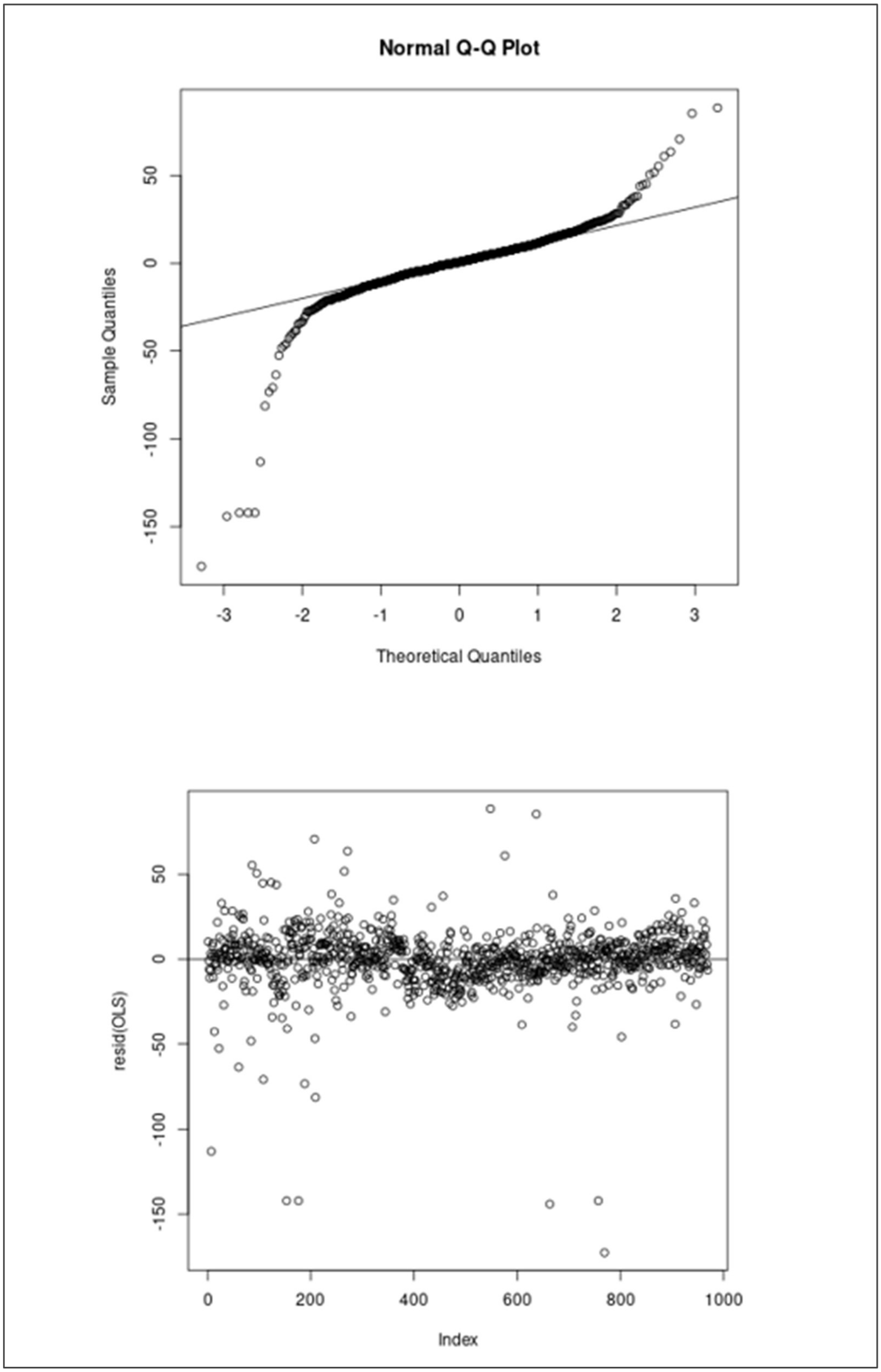
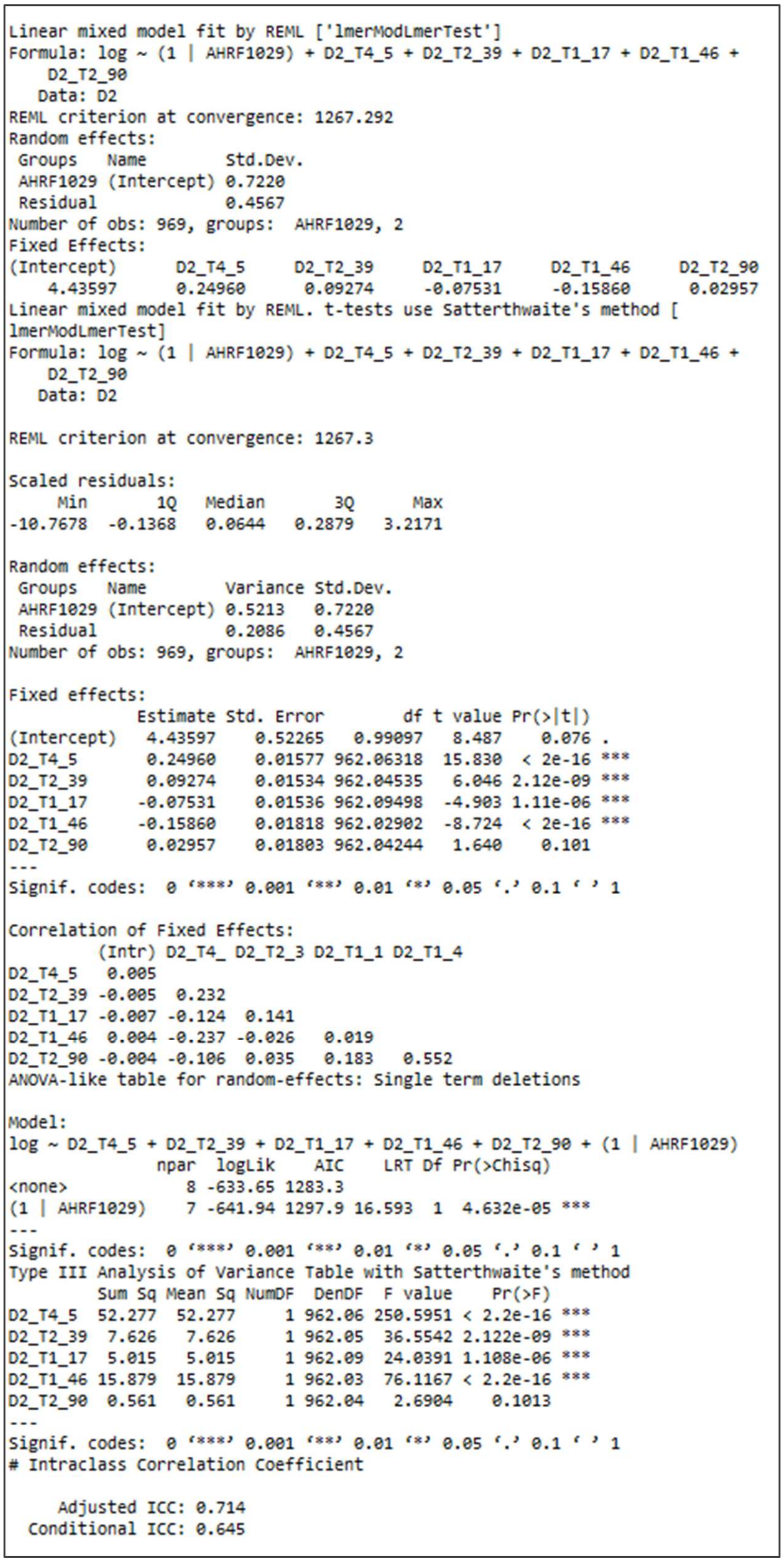


Figure 6. Q-Q (above) and residuals plot (above) from original OLS model.



```
Linear mixed model fit by REML ['lmerModLmerTest']
Formula: log ~ (1 | AHRF1029) + D2_T4_5 + D2_T2_39 + D2_T1_17 + D2_T1_46 +
    D2_T2_90
   Data: D2
REML criterion at convergence: 1267.292
Random effects:
 Groups   Name        Std.Dev.
 AHRF1029 (Intercept) 0.7220
 Residual             0.4567
Number of obs: 969, groups:  AHRF1029, 2
Fixed Effects:
(Intercept)      D2_T4_5     D2_T2_39     D2_T1_17     D2_T1_46     D2_T2_90
    4.43597      0.24960      0.09274     -0.07531     -0.15860      0.02957
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: log ~ (1 | AHRF1029) + D2_T4_5 + D2_T2_39 + D2_T1_17 + D2_T1_46 +
    D2_T2_90
   Data: D2

REML criterion at convergence: 1267.3

Scaled residuals:
     Min       1Q   Median       3Q      Max
-10.7678  -0.1368   0.0644   0.2879   3.2171

Random effects:
 Groups   Name        Variance Std.Dev.
 AHRF1029 (Intercept) 0.5213   0.7220
 Residual             0.2086   0.4567
Number of obs: 969, groups:  AHRF1029, 2

Fixed effects:
             Estimate Std. Error        df t value Pr(>|t|)
(Intercept)   4.43597    0.52265   0.99097   8.487    0.076 .
D2_T4_5       0.24960    0.01577 962.06318  15.830  < 2e-16 ***
D2_T2_39      0.09274    0.01534 962.04535   6.046 2.12e-09 ***
D2_T1_17     -0.07531    0.01536 962.09498  -4.903 1.11e-06 ***
D2_T1_46     -0.15860    0.01818 962.02902  -8.724  < 2e-16 ***
D2_T2_90      0.02957    0.01803 962.04244   1.640    0.101
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
         (Intr) D2_T4_ D2_T2_3 D2_T1_1 D2_T1_4
D2_T4_5   0.005
D2_T2_39 -0.005  0.232
D2_T1_17 -0.007 -0.124  0.141
D2_T1_46  0.004 -0.237 -0.026   0.019
D2_T2_90 -0.004 -0.106  0.035   0.183   0.552
ANOVA-like table for random-effects: Single term deletions

Model:
log ~ D2_T4_5 + D2_T2_39 + D2_T1_17 + D2_T1_46 + D2_T2_90 + (1 | AHRF1029)
               npar  logLik    AIC    LRT Df Pr(>Chisq)
<none>            8 -633.65 1283.3
(1 | AHRF1029)    7 -641.94 1297.9 16.593  1  4.632e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Type III Analysis of Variance Table with Satterthwaite's method
         Sum Sq Mean Sq NumDF  DenDF  F value    Pr(>F)
D2_T4_5  52.277  52.277     1 962.06 250.5951 < 2.2e-16 ***
D2_T2_39  7.626   7.626     1 962.05  36.5542 2.122e-09 ***
D2_T1_17  5.015   5.015     1 962.09  24.0391 1.108e-06 ***
D2_T1_46 15.879  15.879     1 962.03  76.1167 < 2.2e-16 ***
D2_T2_90  0.561   0.561     1 962.04   2.6904    0.1013
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
# Intraclass Correlation Coefficient

     Adjusted ICC: 0.714
  Conditional ICC: 0.645
```

Figure 7. Raw model summary from R ‘lme4’ library for final mixed-effects model and intraclass correlation coefficient result.

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. 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).

19. 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.

20. 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.

21. 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.

22. 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.

23. 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.

24. 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.

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.