Investigating Socioeconomic and Climatic Factors of Mental Health Outcomes in United States Counties

Abhishek Babu, Rehaan Bhimani, Alan Liu, Michael Wilson
Paul G. Allen School of Computer Science & Engineering
University of Washington
Seattle, WA 98195
`{babua, bhimar, alnliu11, wilsmic2}@cs.washington.edu`

Abstract

Mental health problems are rising among the general population in the United States, with suicide being a significant cause of death [3] and an increasing number of people having suicidal thoughts [9]. In addition to personal genomics and pre-existing mental health conditions, socioeconomic [10] and climatic [1] factors also impact mental health. Prior work generally examines mental health through suicide rates alone, which is an extreme manifestation of mental distress [17]. In this paper, we pursue insights about the associations of key socioeconomic and climatic factors with both suicide rates and self-reported poor mental health days to examine mental health outcomes more holistically. Studies that employ only regression analyses are limited by the assumptions that are required to disentangle the independent impacts of any discovered associations [20]. We propose a dual approach of regression analysis and matching analysis to understand and disentangle independent associations of key socioeconomic and climatic factors with mental health outcomes. We observe that U.S. counties with higher college attendance rates and income exhibit lower poor mental health days and suicide rates than more high-risk counties. We also found that more urban counties exhibit higher poor mental health days, but reduced suicide rates compared to their more rural counterparts. We also discovered that counties with lower precipitation tend to have fewer poor mental health days, but increased suicide rates. These findings demonstrate quantitative insights that inform further research and policy actions for interventions on poor mental health in the United States.

1 Introduction

1.1 Motivation and Research Question

Mental health problems are rising among the general population, with suicide being the tenth largest cause of death in the United States [3]. The number of people with suicidal thoughts has been increasing since 2011 and currently represents 4.58% of adults in the U.S. [9] In 2019, more than 47,500 deaths occurred by suicide [3], and 24.7% of U.S. adults with mental illness reported that their need for treatment was unmet [9]. We believe that understanding the salient factors impacting mental health is crucial to mitigating such a problem. In addition to personal history and pre-existing conditions, socioeconomic [10] and climatic [1] factors also affect mental health. In this paper, we investigate how these factors contribute to mental health outcomes. Through our results, we hope to provide quantitative and actionable insights that can inform policy and further research into mitigation strategies for poor mental health in the United States.

Previous work analyzing mental health outcomes focuses on suicide rates [10], which is an extreme manifestation of mental distress [17]. Our research aims to clarify the relationship between socioeconomic factors, climatic factors, and both self-reported poor mental health days and suicide rates to examine effects on mental health outcomes more holistically. Specifically, our **research question** is: What socioeconomic and climate factors are effective predictors and/or causes of mental health outcomes among United States counties?

Our key contributions are expanding previous regression analyses to predict not only suicide rates but also poor mental health days, using more granular regression features such as monthly climate data instead of yearly climate data, and performing a matching analysis to further control for confounding effects and disentangle individual impacts of features. We conducted both regression analyses and matching analyses on socioeconomic and climatic features for United States counties over the periods of 2011–13 and 2015–16. In Section 3.1, we describe our regression models and approach and in Section 4.1, we compare performance across models and determine the most important features for predicting mental health outcomes. In Section 3.2 we describe how we determine treatment and control groups for each variable and in Section 4.2, we assess balance in our matching and a summary of the average treatment effect on the treated (ATT) for each treatment variable. Finally, we discuss our findings, limitations, and potential for future extensions in Section 5.

1.2 Hypotheses

From our data exploration (in Section 2), we expect that median household income and some college education will be important features for predicting these mental health outcomes. We also expect that they may have causal effects on both mental health outcomes.

From our review of related work (in Section 6), we expect that more rural counties and counties with higher temperatures will be associated with higher suicide rates and may have some causal effects on one or both of our mental health outcomes.

2 Dataset, preprocessing and visualization

2.1 Data collection

The county-level socioeconomic data was collected from County Health Rankings [4] for the period of analysis from 2010 to 2016. Due to the complicated and high-dimensional nature of socioeconomic status, the factors examined must be relevant. Therefore, we examined variables that represent the “Big 3" measurements of socioeconomic status as defined by the National Center for Education Statistics (NCES) [12]: *family income*, *parental education* and *parental occupation status*. The specific variables of interest from this source that we chose concerning these measurements are median household income, education attainment (high school graduation %, some college attendance %, including all types of education following high school [6]), and unemployment rate. Additionally, we also included the ratio of a county’s population to the number of mental health providers in the county as a means of quantifying access to mental health care. Furthermore, prior work by Mukherjee and Wei [10] indicates that the urbanization levels of counties are associated with suicide rates. To investigate this, we represented the urbanization of each county through the Rural-Urban Continuum Codes (RUCC) obtained from the U.S. Department of Agriculture Economic Research Service (USDA ERS) [19] for the year 2013. RUCC values range from 1 (more urban) to 9 (more rural), with values from 1 to 3 representing metropolitan counties and values from 4 to 9 representing nonmetropolitan counties.

For the climate data, county-level temperature and precipitation data were obtained from the National Oceanic and Atmospheric Administration’s National Climatic Data Center (NOAA NCDC) [13], averaged on monthly and yearly bases over the period of analysis from 2010 to 2016.

For the mental health outcomes, self-reported poor mental health days (in the last 30 days) data was also obtained from County Health Rankings [4] from 2010 to 2016. Suicide mortality data (per 100,000 people) was collected from the CDC’s WONDER database tool [2] from 2010 to 2016. The data was collected on the county level, filtered by *Injury Intent*, which was specified as “Suicide."

2.2 Preprocessing

We joined data from our sources based on the year and county FIPS code. There were 6,323 unique data points, in terms of county, year, for suicide rates from 2010 to 2016, which we then joined with our socioeconomic data. Additionally, 15 counties did not have precipitation or temperature data from 2010 to 2016, which reduced the size of our valid data to 6,279 points. Due to the reporting of the ratio of population to mental health providers as a 0 or negative number when the data is missing for that county or year [5], we dropped rows that did not have a positive ratio of population to mental health providers, which included all of our data from 2010 and 2014, meaning our final dataset was over the ranges of 2011–13 and 2015–16, giving us 4,471 data points. Finally, we dropped rows from our dataset that contained any NaN values for the factors and outcomes described in Section 2.1. Before dropping, we found that the feature distributions across the filtered out rows and the remaining rows were comparable, meaning that we did not change the study population significantly. After preprocessing, we had 4,441 data points to use for our analysis.

2.3 Data exploration and visualization

To profile our data, we plotted all of the features in consideration against each other and the mental health outcomes (see Figure 1). Many of these plots seemed to show no relationship, but we did find linear relationships between median household income and poor mental health days, median household income and suicide rates (Crude Rate), some college and poor mental health days, and some college and suicide rates. Notably, all of these were negative correlations, so it seems that higher median household income and some college are associated with decreasing levels of mental distress. We explore this further in our regression and matching analyses.

We also see that there seems to be no strong relationship between the ratio of population to mental health providers and either of the mental health outcomes. We investigate this further in our matching analysis (Sections 3.2 and 4.2).

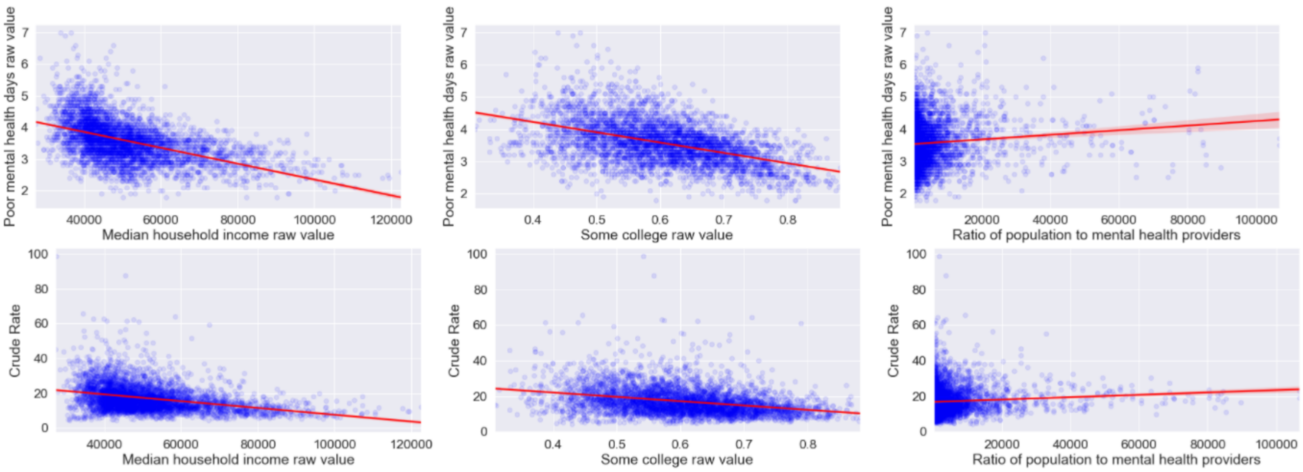

Figure 1: Scatter plots of mental health outcomes and features

3 Analytical methodologies

3.1 Regression

Prior regression work [10] does not take into consideration seasonal climate changes but rather averages climate on an annual basis. The regression analysis that we present is inspired by Mukherjee and Wei [10] but extends it by using monthly averaged climatic features and poor mental health days as an additional outcome.

3.1.1 Model selection process

We considered linear and nonlinear models including Lasso, Ridge, LinearGAM, Random Forest, and Gradient Boosting. Finally, we considered the null model as the benchmark for comparison, which is simply the mean value for the measurements we are predicting.

To determine the best model for our data, we compared three validation metrics for each model including the mean absolute error (MAE), root mean squared error (RMSE), and R-squared (R^2) as they are standard measures to quantify how well the model fits the dataset [10]. The MAE measures the average residuals in the dataset, while the RMSE represents how well the model can predict the response variable absolutely by measuring the standard deviation of the residuals. R^2 indicates the amount of variability in the dependent variable that is explained by the model [8]. With these validation metrics, we can effectively choose the best model for our data from the regression models.

We performed an 80-20 train-test split on the processed training data. Then for each of the regression models, we used 10-fold cross-validation to determine validation metrics, averaging the validation metrics across the 10 training folds. We then compared these metrics to select the best model for predicting each of poor mental health days and suicide rates. Finally, we trained using the full size of our training set and obtained metrics for these models on our unseen test set.

3.2 Matching

In this section, we describe the methodology we applied for matching analysis to isolate the independent effects of the socioeconomic and climatic variables we are interested in.

3.2.1 Matching to mitigate selection effects

To maintain validity, we must also control for other correlated factors to mitigate potential confounding effects. We are interested in the effects of the variables described in Section 2.1. Many of the socioeconomic and climatic variables are very likely to be correlated, meaning that trying to directly isolate individual effects would be confounded by other variables. For example, the median household income of the counties in our dataset is positively correlated with some college attendance rate (Pearson $r = 0.625$) and is negatively correlated with the RUCC value (Pearson $r = -0.476$). If we were to compare high and low household income groups of counties without accounting for other such correlated factors, the distribution of these factors across the high and low-income groups would be uneven. Therefore, it is important to consider potentially correlated factors and adequately control for them when analyzing effects on mental health outcomes.

To disentangle the individual effects of important variables, we employ a matching-based method designed to create a comparable set of groups with similar covariate distributions. The goal behind the matching analysis is to identify pairs of treated and untreated individuals, in this case, counties, that are very similar to each other besides the treatment. These identified pairs provide the counterfactual for each other, which allows us to estimate individual treatment effects. Due to the high dimensional nature of socioeconomic status, correlated socioeconomic variables, and the inclusion of additional climatic factors, finding a unique match for each treated county can be difficult. For example, for a county with a high median household income and high college attendance rate, it is difficult to find a unique county with a high median household income and low college attendance rate due to the correlation observed above. Therefore, we employed one-to-one matching with replacement to achieve better matches for the treated units. We employ the nearest neighbor matching method and a generalized linear model (glm) distance measure to perform the matching using the MatchIt package [7].

3.2.2 Determining treatment and control groups

For each of the socioeconomic and climatic variables, we split all available counties into treatment and control groups using a threshold. For the factors that have similar mean and median values, we used a value close to the median to split the population into two groups. In other cases where the distribution of the variables across counties is skewed, we consulted other literature for a suitable threshold. Specifically, we found that the ratio of a county's population to mental health providers was significantly skewed with a mean of 5262 and a median of 2092, indicating that counties with large ratios of population to mental health providers were significantly larger than those with smaller ratios. In this case, we followed the U.S. Department of Health and Human Services (USDHHS) [18] threshold ratio to meet the mental health needs of a standard population – 9000 to 1.

We consistently defined the treatment group as the group more likely to have a lower suicide rate and a lower number of poor mental health days. Per this, we defined which side of the threshold for each variable would be defined as the treated group with the other side of the threshold as the control group. For education attainment (high school graduation rate and some college attendance rate) and median household income, we defined counties above the threshold as being in the treated group, and for the unemployment rate, we defined counties below the threshold as being in the treated group, supported by the NCES' definition of socioeconomic status [12]. For climatic factors, prior work suggests that higher temperatures [1] and increased precipitation [14] can be associated with worsened mental health. Therefore, we defined counties with lower temperatures and lower precipitation than the threshold to be in the treated group. Prior work also suggests that more rural counties can be associated with higher suicide rates [10]. Therefore, we defined counties with lower RUCC values (more urban) than the threshold to be in the treated group. We also defined counties with a lower ratio of population to mental health providers (i.e. more mental health providers per person) than the USDHHS threshold [18] to be in the treated group.

4 Results

4.1 Regression models

4.1.1 Model comparison and final model selection

Table 1 outlines the validation metrics (average MAE, average RMSE, and average R^2) for training each of the models.

Table 1: Regression Training Results

	Poor Mental Health Days			Suicide Rate		
Model Class	Avg. MAE	Avg. RMSE	Avg. R^2	Avg. MAE	Avg. RMSE	Avg. R^2
Lasso Regression	0.439	0.575	0.246	5.118	7.121	0.108
Ridge Regression	0.402	0.538	0.339	4.201	5.731	0.420
LinearGAM Regression	0.376	0.508	0.411	3.868	5.287	0.506
Gradient Boosting Regression	0.370	0.501	0.426	**3.844**	**5.220**	**0.518**
Random Forest Regression	**0.365**	**0.497**	**0.437**	3.865	5.377	0.491
Null Regression	0.509	0.663	0	5.544	7.583	0

Table 1 shows that the Random Forest model, when predicting poor mental health days, has the best R^2 value and the lowest MAE and RMSE values. This indicates that of the models that we trained, it likely has the best predictive power for poor mental health days. We also found that the Gradient Boosting model, when predicting suicide rates, has the best R^2 value and the lowest MAE and RMSE values. This indicates that out of the models that we trained, it likely has the best predictive power for suicide rates. We found that all chosen models perform better than the benchmark null model.

Table 2: Final Model Regression on Held-Out Test Set

Model Class	Outcome	Average MAE	Average RMSE	Average R^2
Random Forest	Poor Mental Health Days	0.359	0.477	0.437
Gradient Boosting	Suicide Rate	3.827	5.238	0.446

We trained these models on our entire training set to predict their respective mental health outcomes, and evaluated them on the held-out test set, with results shown in Table 2.

4.1.2 Key predictors and ranking

For each of our final models, we determine which of our features are the largest contributors to the predictions. We do this by generating partial dependence plots for the top 3 features for each model, determined by feature importance ranking computed using the scikit-learn package [15]. Larger magnitudes of feature importance indicate higher importance in predicting the outcome. Each partial dependence plot shows the marginal effect of a feature on the predicted outcome of each model.

Our results for the Random Forest model trained for predicting poor mental health days are shown in Table 3. We found that median household income and some college attendance rate are the two most significant features for predicting the number of poor mental health days.

Table 3: Top 3 Feature Importances – Random Forest for Poor Mental Health Days

Feature	Feature Importance
Median household income	0.140449
Some college attendance rate	0.117291
Unemployment rate	0.036349

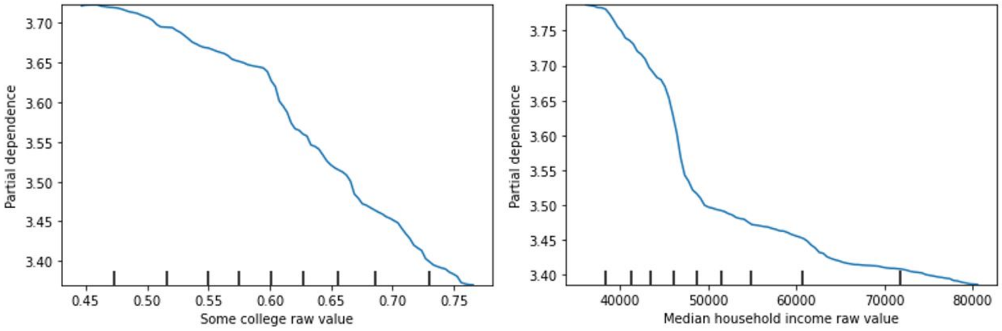

Figure 2: Partial dependence plots of most important features for Random Forest regression

The partial dependence plots for the two most important features are shown in Figure 2. We found that increasing both median household income and some college attendance rate is associated with a decrease in the number of poor mental health days.

Table 4: Top 3 Feature Importances – Gradient Boosting for Suicide Rates

Feature	Feature Importance
RUCC	0.561435
May Average Temperature	0.054525
Some college & Median household income	0.048594

Our results for the Gradient Boosting model trained for predicting suicide rates are shown in Table 4. We find that the RUCC value is the most significant feature for predicting a county's suicide rate. Since the feature importance of the RUCC value is at least an order of magnitude higher than all other features, we only consider this feature in our discussion.

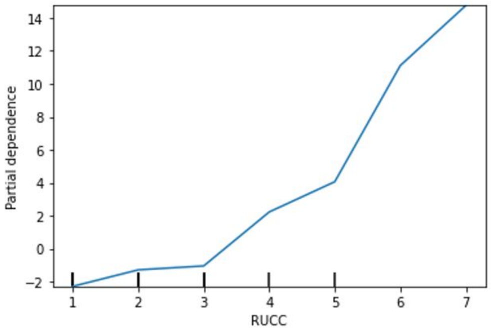

Figure 3: Partial dependence plots of most important features for Gradient Boosting regression

Figure 3 shows the partial dependence plots for the RUCC value. We found that larger RUCC values (i.e. more rural counties) are associated with an increase in suicide rates.

4.2 Matching

4.2.1 Evaluating the matching of counties

Table 5: Summary of pre-matching balance assessment across all covariates

Treatment	Treated	Split Threshold	Treated Count	Control Count	Mean Covariate |SMD|
High school graduation rate	Higher	0.814	2354	2087	0.292
Some college attendance rate	Higher	0.602	2213	2228	0.631
Median household income	Higher	$49,786	2026	2415	0.631
Unemployment rate	Lower	0.0798	2353	2088	0.386
Ratio of pop. to MH providers	Lower	9000	3783	658	0.489
Avg. monthly temperature	Lower	56.1°F	2281	2160	0.318
Avg. monthly precipitation	Lower	3.51"	2240	2201	0.185
RUCC	Lower	2.26	2544	1897	0.330

As the idea behind matching is to find a close match from the control group for each unit in the treated group, we leverage the Standardized Mean Difference (SMD) for each factor across the county groups. The SMD is a measure used to quantify the degree to which two groups are different. It is computed by dividing the difference in means of a variable across two groups by the standard deviation of the treated group [16]. Table 5 outlines the results of splitting the data into treatment and control groups based on the threshold values to compute the pre-matching mean SMD across all other covariates. By using a threshold of $|\text{SMD}| < 0.25$ [16] to determine if two groups are comparable, as is common practice, we can determine the quality of the matching while minimizing potential confounding effects of other covariates that may be correlated with the treatment variable. For example, when we split our counties into the treated and control group along the some college attendance rate factor, all factors except average monthly precipitation fail to meet the $|\text{SMD}| < 0.25$ criterion before matching. This indicates that counties with higher college attendance rates are more likely to have less unemployment ($|\text{SMD}| = -0.927$), higher median household income ($|\text{SMD}| = 0.930$), lower RUCC values (i.e. more urban; $|\text{SMD}| = -0.809$) and lower ratio of population to mental health providers ($|\text{SMD}| = -0.808$). This indicates that these variables are either positively or negatively correlated with the treatment and may have confounding effects that threaten the validity of this analysis. To minimize confounds, we perform pair matching and tune the caliper value for the matching to match each treated unit to a control unit such that the $|\text{SMD}| < 0.25$ criterion is met across all covariates for the two comparison groups. The caliper sets a threshold on how close matches must be to be considered acceptable.

Table 6: Summary of post-matching balance assessment across all covariates. (* = We were unable to find a match such that the max absolute SMD of the RUCC value when splitting on median household income and the max absolute SMD of college attendance when splitting on ratio of population to mental health providers would be < 0.25)

Treatment	Treated	Max |SMD|	Mean |SMD|	Treated Matched	Treated Unmatched	Control Matched	Control Unmatched
High school grad.	Higher	0.106	0.0605	2353	1	1056	1031
Some college	Higher	0.214	0.132	2213	0	742	1486
Household income	Higher	**0.297***	0.123	2026	0	732	1683
Unemployment	Lower	0.0990	0.0459	2353	0	967	1121
Pop. to MH providers	Lower	**0.3590***	0.180	3659	124	527	131
Temperature	Lower	0.1048	0.0508	2281	0	1092	1068
Precipitation	Lower	0.0848	0.0250	2240	0	1146	1055
RUCC	Lower	0.1914	0.139	2544	0	778	1119

Table 6 outlines the max and mean SMDs across all other covariates when splitting on each treatment variable, along with the treatment and control group retention rates after matching.

4.2.2 Estimations of effects

Table 7: Estimated treatment effects on mental health outcomes (* = negative effects indicate that the treatment reduces the number of poor mental health days/suicide rate; # = the threshold that the p-value must be lower than to be considered statistically significant is $\alpha = 0.05$ [11])

Treatment	Treated	Effect on Poor Mental Health Days*	p-value$^\#$	Effect on Suicide Rate*	p-value$^\#$
High school grad.	Higher	0.0196	**0.345**	-0.306	**0.237**
Some college	Higher	-0.246	6.34e-17	-1.63	1.03e-5
Household income	Higher	-0.293	1.03e-28	-0.530	**0.0546**
Unemployment	Lower	0.00125	**0.956**	1.065	0.000127
Pop. to MH providers	Lower	0.128	9.41e-5	-0.698	0.0362
Temperature	Lower	-0.103	1.85e-6	0.276	**0.249**
Precipitation	Lower	-0.0903	7.45e-6	0.511	0.0345
RUCC	Lower	0.148	4.77e-6	-2.47	8.82e-8

After identifying the treated and controlled county groups that have comparable distributions of all other covariates, we compared the mental health outcomes of interest across the matched county groups. Due to correlations and inequalities in the distributions of variables, it is very difficult to estimate the Average Treatment Effect (ATE). Instead, this matching process estimates the Average Treatment Effect on the Treated (ATT) population. For example, if we consider the median household income treatment, we specifically estimate the effects of having high income on poor mental health days and suicide rates while removing potential contributions and effects from all other observed covariates. The ATT estimates provide quantitative insights into the effects of being at lower risk of poor mental health (i.e. our definition of the treatment group for each treatment) that can be used to suggest actions or inform policies to mitigate mental health issues. Table 7 outlines the estimated effects of each treatment on the mental health outcomes along with a measure of the statistical significance of the estimate to ensure that we can reject the null hypothesis. For example, we estimate that counties that have a higher college attendance rate, on average, have 0.246 fewer poor mental health days (in a month) and 1.63 fewer suicides per 100,000 people.

5 Discussion

5.1 Discussion of Regression Results

From our regression analysis, we found that the most important factors for predicting poor mental health days are median household income and some college. In our first hypothesis in Section 1.2, we discussed that our data exploration revealed strong relationships between median household income and some college with poor mental health days. Thus, our first hypothesis is supported by our regression analysis.

We also found that the most important factor for predicting suicide rates is the RUCC value, where more rural counties are associated with higher suicide rates. Our second hypothesis in Section 1.2 discussed that high temperatures and more rural counties would be associated with higher suicide rates. Thus, our regression analysis supports the urbanization feature of our second hypothesis, but cannot confidently do so for the temperature feature.

5.2 Discussion of Matching Results

For both RUCC values and the ratio of population to mental health providers, we observe that the treated group corresponds to more poor mental health days, but lower suicide rates. In this case, more urban areas correspond to more poor mental health days, yet lower suicide rates. On the other hand, the better ratio of population to mental health providers, the higher the poor mental health days, but lower suicide rates. This was surprising to us because our initial data exploration did not reveal any relationship between the ratio of population to mental health providers and either of the mental health outcomes. The relationship was only revealed when we matched counties solely on this factor.

For the climate factors, we see that lower average monthly temperature and precipitation correspond to fewer poor mental health days, implying that people may feel more relaxed living in colder, dryer regions. However, we observe that lower precipitation increases suicide rates, perhaps due to stresses from drought or chronic dry climate.

Finally, we observe counties with better post-secondary education and earnings help reduce individuals' tendency of poor mental health conditions. For education specifically, it also corresponds with lower suicide rates. However, we see a lower unemployment rate corresponding to higher suicide rates.

5.3 Interpretation of Results

Both our regression analysis and causal inference analysis indicate that higher college attendance and income are associated with lower levels of poor mental health days. We also see from our causal inference analysis that higher college attendance leads to lower suicide rates. We believe that this is the case because secondary education is more area-focused and intensive, contributing to individuals' abilities to find jobs and cultivating their sense of meaning in life. Notably, we see a lower unemployment rate corresponding to higher suicide rates in our causal inference. This suggests that stresses engendered by jobs could exceed anxiety resulting from unemployment, but this requires further research.

In the case of urbanization, our regression analysis finds that rural communities are more likely to have higher suicide rates. This is consistent with our causal inference analysis which indicates that urban communities have lower suicide rates but higher levels of poor mental health days. A possible explanation might be that the fast-paced lifestyle in urban counties causes their residents to be more anxious, but also lessens individuals' thoughts about suicide due to the business nature.

In the case of mental health resources, we see from the causal inference analysis that areas with more mental health providers have lower suicide rates and higher levels of poor mental health days. Such a relationship can either be a direct effect of government interventions to establish more mental health resources in areas of poor mental health or that the ease of accessing providers tends to increase individuals' incentive to seek such help, resulting in reporting more poor mental health days. However, such resources effectively control people's mental health conditions and thus reduce their likelihood and tendency of committing suicide. Further research into this is required to conclude anything definitively.

5.4 Limitations

We note several limitations in both our data and analysis results. The CDC Wonder tool does not provide data for counties with less than 10 suicide in a given year to protect anonymity. Thus, our results may only be generalizable for a smaller population of counties that exhibit at least 10 suicides in a given year. Our study incorporates the ratio of population to mental health providers, but we do not have data for the utilization rate of mental health resources. So, it is difficult to conclusively say if provided resources are useful. Finally, the number of poor mental health days is a self-reported measure recorded from telephone surveys. This measurement could be biased by the tendency for different populations to have a higher likelihood of participating in such surveys.

In our matching analysis, we also note that when splitting treatment and control groups, there were two treatment variables (median household income and ratio of population to mental health providers) where we were unable to find a match such that the $|\text{SMD}| < 0.25$ criterion was met for exactly one covariate (see Table 6). Due to this, we cannot confidently say that any estimated effects for those treatments were free of potential confounding. Additionally, for some of the estimated effects, the p-values are larger than 0.05, indicating that we cannot confidently reject the null hypothesis for those treatment effects. This may be due to there not being a relationship between the treatment and outcome in question, or due to a lack of data. Further research is required to answer this question.

5.5 Implications for Further Research

Future research can explore new sets of measurements that capture the usage rate of mental health resources. This variable could give more insight into how an increase in usage rates impacts mental health and suicide rates. In addition, exploring particular aspects of urban-ness increases the feasibility of an intervention to stimulate better mental health, whether it means building a transportation system or establishing more recreational facilities. We believe such research should aim to leverage similar matching-based approaches to control for confounders, which allows us to disentangle individual impacts of factors on specific mental health outcomes. New ways to measure less extreme mental health outcomes are encouraged in place of, or in addition to, poor mental health days.

6 Related work

Mukherjee and Wei [10] explored the effect of socio-environmental factors on the suicide rates between urban and suburban regions. Our dataset has significant overlap with theirs, but they used a longer span from 2000-2017 and a smaller set of counties with a higher suicide trend. In short, they also retrieved counties' suicide data from CDC Wonder and examine them monthly. They also obtained climate data from NOAA on a monthly basis, yet they analyzed days of a certain level of temperature and precipitation, such as but not limited to, seasonal cooling days and the number of days with temperature $\geq$ 90°F. They also grouped counties into urban and rural, yet used the NCHS scheme of six-level urban-rural classification instead. For analyses, they separated counties into large central metropolitan counties and medium/small metropolitan counties. For each group, they ran a library of regression models, including generalized linear model, four variants of regression splices), and all models we employed. The features included demographics, education, unemployment, income, mild/high-temperature days, and extreme daily precipitation to predict suicide rates. Hence, they identified relevant factors of suicide rates for each group separately. In the end, they discovered that suicide rates were higher in non-urban areas. Higher suicide rates correlated with higher temperature and precipitation in urban counties and an increase in seasonal cooling degree days in suburban counties.

Burke et al. [1] employ ordinary least squares (OLS) regression to estimate the monthly suicide rate. In contrast to our study of using static temperature values, the authors use binned models where suicide rates are modeled as a function of accumulated exposure to different daily temperatures. They found that a 1°C increase in average monthly temperature increases the monthly suicide rate by 0.68% in the United States over the years 1968–2004. They also used a second suicide rate dataset from the CDC to find that a 1°C increase in average annual temperature increases the annual suicide rate by 1.3%. These results were in contrast to prior studies conducted regarding temperature in the United States, which showed varied responses according to the authors.

7 Conclusion

Our work holistically investigates the relationships between socioeconomic factors, climate factors, and mental health outcomes. Where previous work only investigated suicides, an extreme manifestation of mental distress, our work includes poor mental health days. We see from our analyses that this is an important outcome to investigate, as it interacts with the features differently than suicide rates. Our regression analysis indicates that income and education are important predictors of poor mental health days, and this is substantiated by the matching analysis. Our regression analysis also indicates that more rural counties are associated with higher suicide rates, which is also supported by our matching analysis. Our matching analysis also indicates that urban areas have higher levels of poor mental health days. By combining our regression and matching analyses, we develop a more holistic understanding of mental health in U.S. counties. Our findings demonstrate quantitative and actionable insights that could inform policy actions to mitigate the exacerbation of mental health issues in the United States. Additionally, we believe that our analysis could provide direction for further research into specific interventions that can help combat poor mental health.

Acknowledgments

We thank Tim Althoff, Jina Suh, and Mike Merrill at the Paul G. Allen School of Computer Science & Engineering, University of Washington for their guidance, support, and comments.

References

[1] M. Burke, F. González, P. Baylis, S. Heft-Neal, C. Baysan, S. Basu, and S. Hsiang. Higher temperatures increase suicide rates in the United States and Mexico. *Nature Climate Change*, 8: 723–729, Aug 2018.

[2] Centers for Disease Control and Prevention (CDC). CDC WONDER, Dec 2021. URL `https://wonder.cdc.gov/`. Last accessed Dec 9, 2021.

[3] Centers for Disease Control and Prevention (CDC). Facts About Suicide, Aug 2021. URL `https://www.cdc.gov/suicide/facts/index.html`. Last accessed Dec 9, 2021.

[4] County Health Rankings. National Data & Documentation: 2010-2019, 2019. URL `https://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation/national-data-documentation-2010-2019`. Last accessed Dec 9, 2021.

[5] County Health Rankings. Mental health providers, 2021. URL `https://www.countyhealthrankings.org/explore-health-rankings/measures-data-sources/county-health-rankings-model/health-factors/clinical-care/access-to-care/mental-health-providers`. Last accessed Dec 12, 2021.

[6] County Health Rankings. Some College, 2021. URL `https://www.countyhealthrankings.org/explore-health-rankings/measures-data-sources/county-health-rankings-model/health-factors/social-and-economic-factors/education/some-college`. Last accessed Dec 12, 2021.

[7] D. E. Ho, K. Imai, G. King, and E. A. Stuart. MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 42(8):1–28, 2011.

[8] IBM. Examining the model, 2021. URL `https://www.ibm.com/docs/en/spss-modeler/18.3.0?topic=node-examining-model`. Last accessed Dec 12, 2021.

[9] Mental Health America (MHA). The State of Mental Health in America, 2021. URL `https://www.mhanational.org/issues/state-mental-health-america`. Last accessed Dec 12, 2021.

[10] S. Mukherjee and Z. Wei. Suicide disparities across metropolitan areas in the US: A comparative assessment of socio-environmental factors using a data-driven predictive approach. *PLOS One*, 16(11), Nov 2021.

[11] F. S. Nahm. What the *P* values really tell us. *The Korean Journal of Pain*, 30(4):241–242, Oct 2017.

[12] National Center for Education Statistics (NCES). Improving the Measurement of Socioeconomic Status for the National Assessment of Educational Progress: A Theoretical Foundation, Nov 2012. URL `https://nces.ed.gov/nationsreportcard/pdf/researchcenter/Socioeconomic_Factors.pdf`. Last accessed Dec 9, 2021.

[13] National Oceanic and Atmospheric Administration (NOAA) National Centers for Environment Information. Climate at a Glance: County Time Series, Dec 2021. URL `https://www.ncdc.noaa.gov/cag/county/time-series`. Last accessed Dec 9, 2021.

[14] N. Obradovich, R. Migliorini, M. P. Paulus, and I. Rahwan. Empirical evidence of mental health risks posed by climate change. *Proceedings of the National Academy of Sciences*, 115(43): 10953–10958, Oct 2018.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[16] E. A. Stuart, B. K. Lee, and F. P. Leacy. Prognostic score–based balance measures for propensity score methods in comparative effectiveness research. *Journal of Clinical Epidemiology*, 66(8): S84–S90, 2013.

[17] C. Tannenbaum, J. Lexchin, R. Tamblyn, and S. Romans. Indicators for Measuring Mental Health: Towards Better Surveillance. *Healthcare Policy = Politiques de Santé*, 5(2):e177—-e186, Nov 2009.

[18] U.S. Deparment of Health and Human Services (USDHHS) Health Resources and Services Administration (HRSA). Health Professional Shortage Areas: Designated HPSA Quarterly Summary, Oct 2021. URL `https://data.hrsa.gov/topics/health-workforce/shortage-areas`. Last accessed Dec 10, 2021.

[19] U.S. Department of Agriculture (USDA) Economic Research Service (ERS). Rural-Urban Continuum Codes, Dec 2020. URL `https://www.ers.usda.gov/data-products/rural-urban-continuum-codes.aspx`. Last accessed Dec 9, 2021.

[20] L. Wasserman and C. Shalizi. Causal Inference, 2017. URL `http://www.stat.cmu.edu/~larry/=stat401/Causal.pdf`. Last accessed Dec 12, 2021.