Supplementary Information for

A Pilot Study Exploring the Relationship between Urban Greenspace Accessibility and Mental Health Prevalence in the City of San Diego in the Context of Socioeconomic and Demographic Factors

Alexis Wilderman[1], Marcus Lam[2], Zhi-Yong Yin[1*]

[1] Environmental & Ocean Sciences, University of San Diego, San Diego, CA 92110, USA

[2] School of Leadership and Education Sciences, University of San Diego, San Diego, CA 92110, USA

Author emails:

Alexis Wilderman: awilderman@sandiego.edu

Marcus Lam: mlam@sandiego.edu

Zhi-Yong Yin*: zyin@sandiego.edu

* Corresponding author

Content

Part 1. Notes on Datasets

1. CDC 500 Cities Project Dataset

The CDC 500 Cities Project used model-based small area estimates (SAE) to represent the geographic distribution of 27 estimates, including health outcomes (e.g., diabetes, high blood pressure, etc.), clinical preventive services (e.g., health insurance coverage, routine check-ups, etc.), and chronic disease risk factors (e.g., physical activity, current smoking, etc.) across the entire US [22,57]. In this study, we used the SAE of various measures of public health conditions, including the metal health prevalence (MHP), at the census tract-level within this dataset. The SAE dataset was created by combining data from the CDC's Behavioral Risk Factor Surveillance System (BRFSS) responses in a given year, with Census Bureau and American Community Survey (ACS) population estimates within the tract [22,23]. The BRFSS is a national surveillance system that uses random digit dialing to conduct telephone (cellphone and landline) surveys of the US population aged ≥18 years to collect demographic and health data [22,23]. BRFSS is the gold standard in telephone-based surveys and completes nearly 500,000 interviews annually with a response rate approaching 50% [23,57].

To produce the SAE dataset, multilevel logistic models were fit to BRFSS data to determine an individual's probability of having a health behavior or outcome as a function of compositional effects [26]. The modeling process uses individual-level responses from the surveys to get compositional effects, including age, race/ethnicity, sex, and education, with county-level rates of poverty from ACS and county- and state-level contextual effects [25,26]. In this case, compositional effects are the socio-demographic variables, or variables that differ in characteristics of individuals, as stated previously. Contextual effects are the random, unmeasured factors that may affect an individual's probability of having a health outcome [26]. Finally, to estimate city- and tract-level prevalence of health conditions, "the models were post-stratified to their respective area group composition population estimates using 2010 Census demographic data" [22,23].

2. GIS Data of Urban Greenspace and Parks

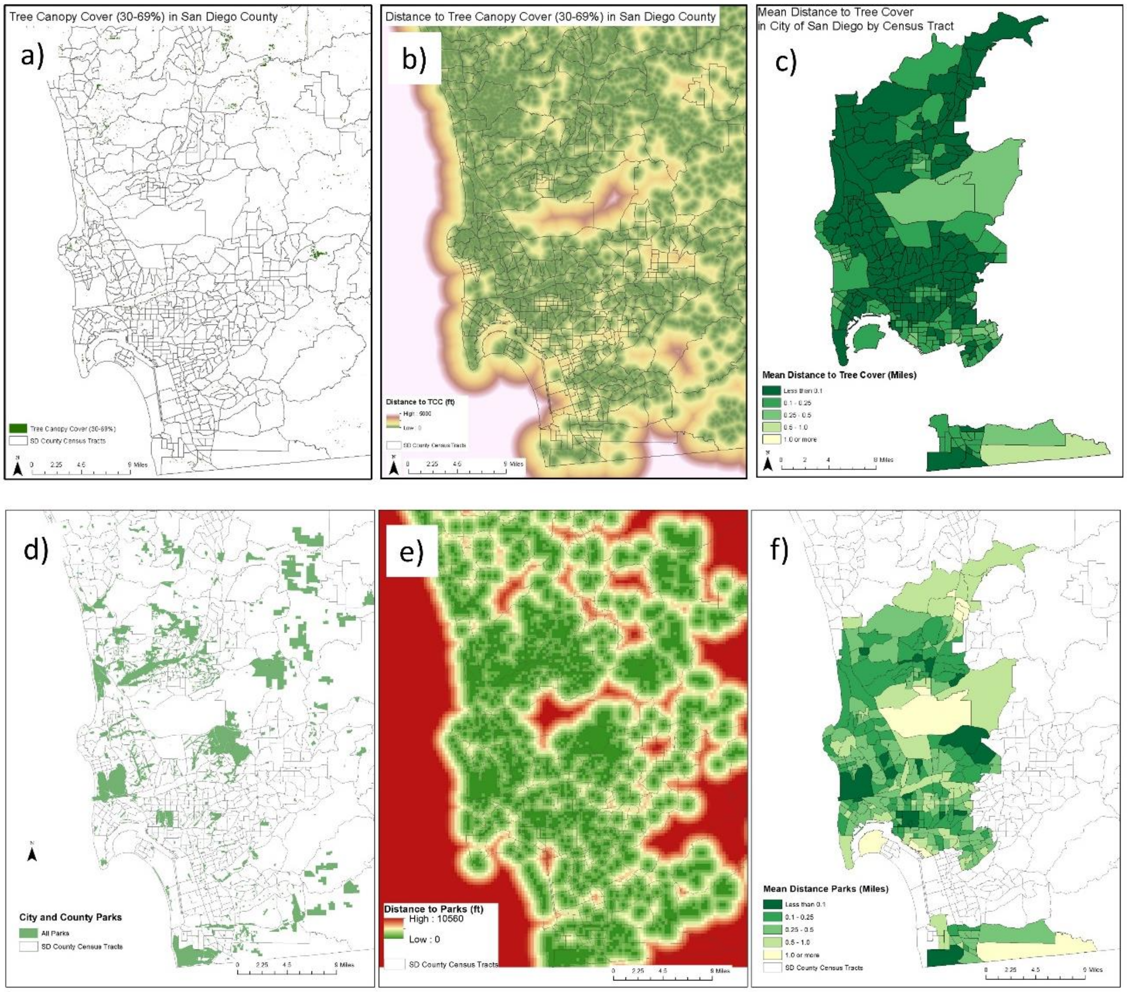
The urban greenspace proximity was based on the San Diego County vegetation GIS data layer (ECO_VEGETATION_CN) obtained from SANDAG/SANGIS (https://rdw.sandag.org/Account/gisdtview?dir=Ecology). The latest update of this dataset occurred in 2018 by the San Diego County Department of Planning and Landuse using aerial imagery and georeferenced bio-maps [28] and, therefore, it has a better spatial resolution than the vegetation or tree canopy cover data based on satellite images [7].

Among the major vegetation types, coastal sage scrub and chaparral are one of the most widespread vegetation communities in San Diego and several well-known San Diego hiking destinations, including Torrey Pines State Natural Reserve and Tecolote Canyon, are comprised of chaparral and scrub [61]. Disturbed or Developed Areas, which was the largest category of land use type in the dataset, was excluded from this measure.

Boundary polygons of parks were also obtained from SANDAG/SANGIS (https://rdw.sandag.org/Account/gisdtview?dir=Park), which is a consolidated dataset with parks maintained by the County of San Diego, sixteen incorporated cities, San Diego Port District, SANGIS, and California State Parks [28]. It was last updated in January 2021 [28]. The types of parks included in this study are Historic, Local, National, Open Space, Reserves, Regional, State, and Other. The "Other" type includes two designated open space parks: one is located in Los Peñasquitos Canyon and the second is located in Switzer Canyon Open Space.

Tree canopy cover data was obtained from the United States Department of Agriculture (USDA) United States Forest Service (USFS) [29]. This dataset is based on Landsat images of 30-meter resolution (https://landsat.gsfc.nasa.gov/) and is available for the Continental United States, although only data for the City of San Diego was used. The data values represent the proportion covered by tree canopy within a pixel, ranging from 0% to 100% [29]. For the City of San Diego, however, the maximum tree cover percentage is 69%.

Supplementary Figure 1 shows the processes of creating the mean distances to locations of tree canopy cover (30-69%) and parks summarized by the census tract polygons, as described in the Data and Methods section of the main text.



Supplementary Figure 1: Creation of *distance to tree canopy cover* in ArcGIS using (a) tree canopy cover data obtained from USDA/FS [29], (b) Euclidean distance tool of ArcGIS to calculate distance to locations with tree canopy coverage > 30%, and (c) the Zonal Statistics tool to calculate mean distances for census tracts in the City of San Diego. The same process was repeated for the distance to the park polygons (*all parks*) summarized by census tracts (d–f).

3. Race and Ethnicity

Throughout the paper, the phrase "race and ethnicity” or “race-ethnicity" is used as one term instead of two separate terms. In the City of San Diego, non-Hispanic White residents make up 44.4% of the total population according to ACS 2011-2015 5-year estimates [20]. To represent the two dominant underrepresented racial/ethnic groups, *percent Hispanic* represents the Hispanic population of all races (Figure 4d), and p*ercent Black* represents the non-Hispanic Black population to avoid any double-counting (Figure 4c). While many of the Hispanic residents are from Mexico or of Mexican descent, which is generally characterized as "Latinx", this group also includes people with heritages from other countries of Central and South America [20].

4. Crime Data

The crime incident data were downloaded from the SANDAG Automated Regional Justice Information System (ARJIS) (https://www.sandag.org/index.asp?subclassid=21&fuseaction=home.subclasshome). The crime incidents for the period of 180 days prior to October 2019, including 61,942 of all types of crimes, were registered to the approximate locations of crime occurrences at the nearest hundredth of street numbers. These incidents were first geocoded using ArcGIS and then exported as a point data layer. Although the dates do not match up with the years of the other data included in the study, *crime density* is used as a proxy of crime rate variation in the City of San Diego. Since a reader may have difficulty finding the specific 180-day crime data prior to October 2019, we will attach the shapefile of the geocoded crime occurrence data (compressed) with this Supplementary Information or it can be provided upon request by the corresponding author.

Part 2. Collinearity among Independent Variables

Before we constructed the regression model of MHP using the variables identified in Equation 1 of the main text, we first examined the collinearity among the potential independent variables. We assumed that these factors of MHP are influenced by other factors such as crime occurrence, accessibility to greenspace and parks, race-ethnicity factors (specifically the percentages of Hispanic and non-Hispanic Black population), and economic conditions (employment). In this process, we tried to include the causal factors of a given variable in the pool of independent variables. For example, we decided to use median housing value as a potential causal factor of percent uninsured population, percent of population with regular checkups, educational attainment, and income, since it is a measure of household wealth; while housing affordability (housing value/income) is considered a factor of percent of population below poverty line. In the meantime, median household income is considered one of the causal factors of poverty, but not *vice versa*, although it can be argued that both poverty and income may be influenced by the same set of other factors. Finally, since housing affordability is calculated using housing value and income, we selected median housing value as the causal factor of housing affordability, but included factors of income in the model. The following are the collinearity models of the major independent variables of MHP, with the terms in parentheses at the end of each equation representing "factors" of the model rather than the variable names (e.g. percent degree is a proxy of preparation for good employment and housing value is a measure of family wealth):

1) Percent Uninsured = f [Median Income, Degree, Poverty, Housing Value, Hours Worked, Race-Ethnicity, Crime Density, Greenspace]; (employment + all factors of wealth)
2) Regular Checkup = f [Uninsured, Median Income, Poverty, Degree, Housing Value, Hours Worked, Race-Ethnicity, Crime Density, Greenspace]; (insurance + wealth)
3) Income = f [Degree, Housing Value, Hours Worked, Race-Ethnicity, Crime Density, Greenspace]; (factors of preparation for high-income jobs + employment)
4) Educational Attainment = f [Median Income, Housing Value, Poverty, Hours Worked, Race-Ethnicity, Crime Density, Greenspace]; (factors to prepare for higher education)
5) (Lack of) Housing Affordability = f [Degree, Housing Value, Poverty, Hours Worked, Race-Ethnicity, Crime Density, Greenspace]; (housing value + factors of income + poverty)

6) Poverty = f [Median Income, Degree, Hours Worked, Affordability, Race-Ethnicity, Crime Density, Greenspace]; (factors of wealth + employment + housing affordability).

Similar to the general model of MHP (Eq.1 in the main text), the race-ethnicity variables are included in all the models above because we want to see whether the race-ethnicity factors have both direct and indirect effects on these variables. Supplementary Table 1 contains the results of stepwise regression with all variables that entered the models statistically significant at the 0.095 level. For simplicity, we only included the names of variables and their signs of regression coefficients, and the R^2 values. The common logarithmic transformation (with 10 as the base and noted as lg_VariableName) was performed for all variables used in regression analysis.

Supplementary Table 1: Regression of the major independent variables in Eq. 1 of the main text against other independent variables. The lists of the independent variables follow the sequences they entered the models (with signs of the regression coefficients). All variables are log-transformed (N = 257 after removal of census tracts with population less than 200 and missing values).

Factors of MHP	**Variables entered the model**	$\mathbf{R^2}$	$\mathbf{R_{adj}^2}$
Percent Uninsured	lg_Degree (-), lg_Hispanic (+), lg_Poverty (+), lg_Housing Value (-), lg_Hours Worked (-), lg_Black (+)	0.956	0.955
Regular Checkup	lg_Crime Density (-), lg_Greenspace (+), lg_Black (+), lg_Uninsured (-), lg_Degree (-)	0.510	0.500
Median Income	lg_Hispanic (-), lg_Crime Density (-), lg_Hours Worked (+), lg_Degree (+), lg_Housing Value (+), lg_Parks (-)	0.827	0.822
Percent Degree	lg_Hispanic (-), lg_Poverty (-), lg_Crime Density (+), lg_Housing Value (+), lg_Greenspace (-), lg_Median Income (+)	0.806	0.802

Housing Affordability	lg_Poverty (+), lg_Housing Value (+), lg_Crime Density (+), lg_Hispanic (+), lg_Degree (-)	0.763	0.758
Poverty	lg_Median Income (-), lg_Crime Density (+), lg_Hours Worked (-)	0.746	0.743

Based on this analysis, we can conclude that the percent of uninsured population is clearly determined by socioeconomic and race-ethnicity factors with more than 95% of the variance explained using the variables obtained from the US Census ACS dataset. On the other hand, regular checkup has a relatively low explained variance (R^2 = 0.510). It is surprising to see that the higher the educational attainment in a census tract, the lower the rate of regular checkups and that the only race-ethnicity variable that entered this model (percentage of Black) has a positive impact. It is also surprising that race-ethnicity variables do not enter the model of poverty. As expected, educational attainment and household income are mutually influential and the percentage of Hispanic population has negative impacts on both. However, the inclusions and effects of the other variables (such as crime occurrence, housing value, and accessibility to greenspace) are different among the models, but the signs of most regression coefficients match the relationships specified in the correlation analysis earlier (Table 3 in the main text). Generally speaking, those variables with very high R^2 values should be noted for potential issues associated with high collinearity in regression analysis below [42], such as percent uninsured (R^2 = 0.956), median income (R^2 = 0.827), and percent degree (R^2 = 0.806), which also means that the effects of these variables on MHP may be replaced by the other variables.

Part 3. Regression Models of MHP

1. Stepwise Regression

Supplementary Table 2 contains the regression models for MHP using stepwise regression based on the independent variable pool specified in Eq. 1 in the main text. All variables are transformed using the logarithmic function with 10 as the base (noted as lg_VariableName). In the first run of stepwise regression (Supplementary Table 2, Part a), the model explained 94% of the variance in MHP ($R_{adj}^2 = 0.938$) with F = 488.39 (p-value < 0.001). Percent of uninsured population (Uninsured) was the first one to enter and also contributed most to the variation in MHP, based the standardized regression coefficient (Beta = 0.375). However, it also has the highest variance inflation factor (VIF = 17.840) among the independent variables, suggesting that its effects can be replaced by other independent variables. After removal of Uninsured, the VIF values of all independent variables are below 10.0 (Supplementary Table 2, Part b) while R^2 remained high at 0.937 ($R_{adj}^2 = 0.936$) with F = 458.02 (p-value < 0.001). Educational attainment (Degree) now makes the largest contribution to the variation of MHP (Beta = -0.385), followed by regular checkup, hours worked, and percent of Hispanic population. We further simplified the model by removing median income from the independent variable pool, which has a VIF of 6.905 and its removal does not influence the model performance (presented in Table 4b of the main text) and further reduces the VIF values of all remaining independent variables to below 5.0, suggesting that the potential errors associated with the collinearity should be inconsequential [39].

Supplementary Table 2: Results of stepwise regression of log-transformed MHP (lg_MHP) (N = 257 after removal of census tracts with population fewer than 200 and missing values due to log-transformation).

Part a: Stepwise regression model with all variables significant at the 0.05 level. Column "B" is the regression coefficients and "Beta" is the standardized regression coefficients.

Part a.	**B**	**Std. Error**	**Beta**	**t**	**Sig.**	**Tolerance**	**VIF**
(Constant)	3.893	0.269		14.467	0.000		
lg_Uninsured	0.167	0.029	0.375	5.722	0.000	0.056	17.840
lg_Checkup	-1.105	0.099	-0.229	-11.182	0.000	0.575	1.740
lg_Median Income	-0.043	0.022	-0.081	-1.980	0.049	0.145	6.903
lg_Hours Worked	-0.473	0.070	-0.127	-6.737	0.000	0.679	1.473

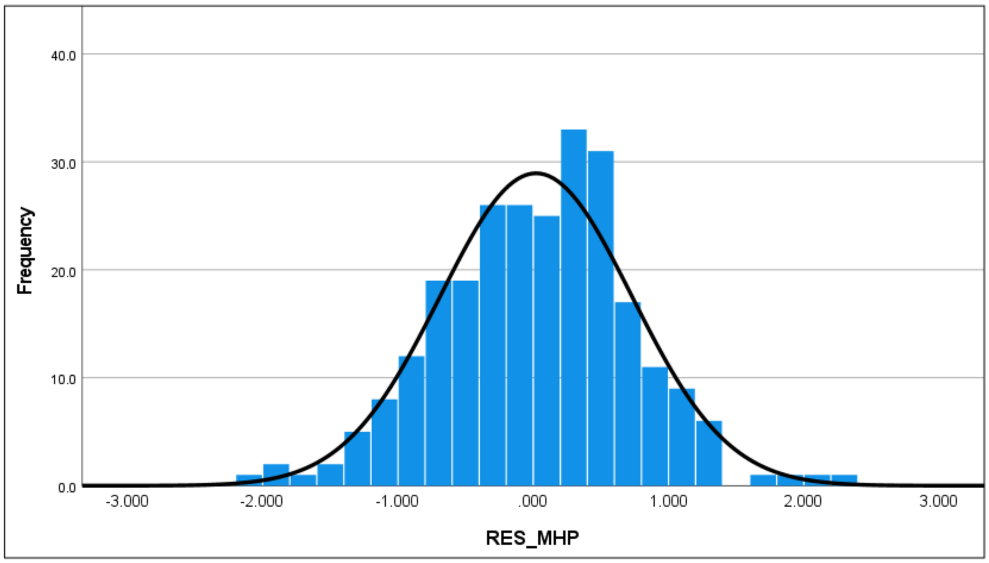
lg_Affordability	0.073	0.015	0.101	4.885	0.000	0.567	1.765
lg_Black	0.018	0.004	0.094	4.974	0.000	0.678	1.475
lg_Degree	-0.083	0.020	-0.217	-4.102	0.000	0.086	11.664
Lg_Poverty	0.023	0.009	0.083	2.627	0.009	0.238	4.194

Part b: Stepwise regression model after removal of lg_Uninsured

Part b.	**B**	**Std. Error**	**Beta**	**t**	**Sig.**	**Tolerance**	**VIF**
(Constant)	4.470	0.236		18.950	0.000		
lg_Median Income	-0.051	0.022	-0.096	-2.285	0.023	0.145	6.905
lg_Hispanic	0.045	0.011	0.149	4.041	0.000	0.189	5.293
lg_Checkup	-1.219	0.098	-0.252	-12.499	0.000	0.627	1.596
lg_Percent Degree	-0.147	0.014	-0.385	-10.736	0.000	0.199	5.029
lg_Hours Worked	-0.548	0.071	-0.147	-7.702	0.000	0.702	1.424
lg_Black	0.022	0.004	0.110	5.775	0.000	0.699	1.431
Lg_Poverty	0.035	0.009	0.124	3.878	0.000	0.250	4.007
Lg_Affordability	0.058	0.015	0.079	3.864	0.000	0.606	1.650

2. Residual Analysis

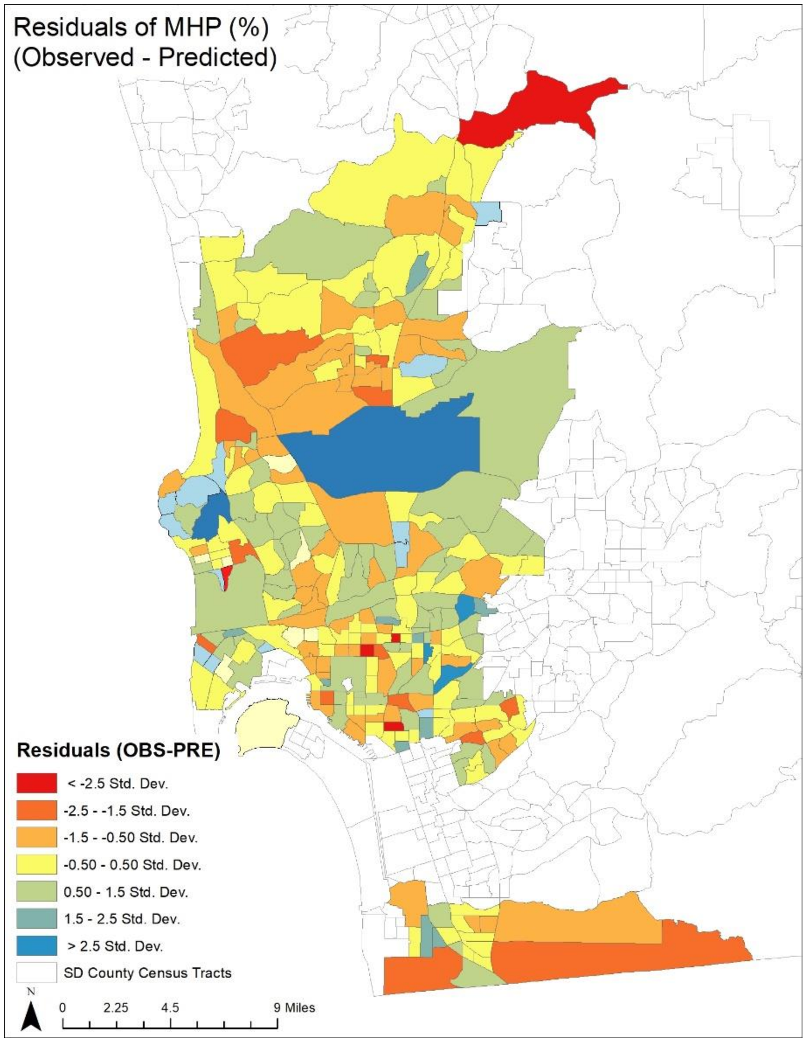
Supplementary Figure 2 shows the histogram of the residual of MHP and Supplementary Table 3 contains the descriptive statistics of the residual calculated as the difference between the observed MHP and predicted values for 257 census tracts. The residual has a mean value close to zero and relatively low skewness and kurtosis values. The non-parametric one-sample Kolmogorov-Smirnov test [60] on the 257 valid residual values suggests that the hypothesis that the residual of MHP has a normal distribution could not be rejected (p-value > 0.200). Therefore, we can conclude that the final regression model of MHP can produce reasonably accurate estimates. Supplementary Figure 3 shows the spatial distribution of residuals (N = 257), or the observed mental health outcomes minus the predicted mental health outcomes, of the final model to see the spatial pattern in overestimating or underestimating the MHP values in the City of San Diego. A negative residual value represents that the prediction was too high, or that the mental health condition was overestimated by our model, and a positive residual represents that the prediction was too low or underestimated. Generally, the residuals seemed to vary randomly throughout the City of San Diego. However, there are areas where the residual values tend to be consistently positive or negative, which may reflect that the MHP as well as its factors vary with a spatial scale somewhat greater than the census tracts, making the adjacent tracts tend to have similar values. Overall, there does not seem to be significant violations of the assumptions of regression analysis [42] and, therefore, the model coefficient estimates should be reasonably accurate.



Supplementary Figure 2: Histogram of residual of MHP (RES_MHP).

Supplementary Table 3: Descriptive statistics of the residual of MHP.

	N	Minimum	Maximum	Mean	Std. Dev.	Skewness	Kurtosis
Residual of MHP	257	-2.042	2.315	0.021	0.708	-0.071	0.393



Supplementary Figure 3: Variation in residuals for MHP by census tracts based on the final linear regression model (Table 4b in the main text). Negative values represent overestimation by the model and positive values represent underestimation. Note that the legend displays the residuals as the departure from the mean in the unit of standard deviation (Supplementary Table 3).

3. Regression Analysis Using Greenspace Proximity and the "Dissimilar Variables" from the Partial Correlation Analysis

Based on the partial correlation results (Table 5 of main text), we expanded the model presented in Table 4a (main text) by including the group of Dissimilar Variables in Table 5 (main text) using stepwise regression. The final model explained 73% of the variance in lg_MHP ($R_{adj}^2 = 0.724$) with F value of 112.357 ($p < 0.001$), and all variables entered the model are statistically significant with the signs of their regression coefficients representing the same relationships to MHP as in the rank correlation analysis (Table 3 of the main text). Although this model (Supplementary Table 4) does not provide the same high level of predictive power as the model in Table 4b of the main text, it demonstrates how greenspace proximity variables can be used to estimate the impact of urban greenspace accessibility on MHP together with the SES and race-ethnicity variables. This outcome is worthy of further investigation in future studies.

Supplementary Table 4: Results of stepwise regression with lg_MHP (log-transformed MHP) as the dependent variable and both greenspace proximity and SES and race-ethnicity variables as the independent variables. All variables are log-transformed (N = 257 after removal of census tracts with population fewer than 200 and missing values due to log-transformation).

	B	**Std. Error**	**Beta**	**t**	**Sig.**	**Tolerance**	**VIF**
(Constant)	5.899	0.361		16.357	0.000		
lg_Hours Worked	-1.376	0.131	-0.368	-10.467	0.000	0.875	1.143
lg_Black	0.061	0.007	0.312	8.767	0.000	0.853	1.172
lg_Checkup	-1.673	0.169	-0.347	-9.892	0.000	0.882	1.134
lg_Greenspace	0.055	0.011	0.188	5.067	0.000	0.786	1.272
lg_Affordability	0.122	0.025	0.167	4.832	0.000	0.908	1.102
lg_Tree Cover	0.029	0.013	0.077	2.156	0.032	0.848	1.180

Part 4. General Characteristics of the Subsections of City of San Diego

The three subsections of the City have very different socioeconomic and demographic characteristics. In general, there is a north to south gradient in terms of socioeconomic status, educational attainment, and race-ethnicity composition, as well as the mental health prevalence. The southern subsection (n = 66) has the greatest mean MHP (14.3%), greatest Black population (9.9%), greatest Hispanic population (61.5%), lowest White population (13.4%), lowest mean median income ($49,033), and lowest mean educational attainment (23.3%) (Supplementary Table 5a). The middle subsection (n = 73) had a mean MHP of 12.5% and a mix of racial and ethnic groups (52.5% White, 8.2% Black, and 29.0% Hispanic) (Supplementary Table 5b). The northern subsection (n = 141) had the lowest mean MHP (9.9%), lowest Black population (3.0%), lowest Hispanic population (14.2%), greatest White population (59.0%), highest mean median income ($89,746), and the highest educational attainment (62.5%) (Supplementary Table 5c). The southern census tracts had the highest mean distance to greenspace (2623.8 ft or 0.497 miles) and the northern census tracts had the lowest (1211.5 ft or 0.229 miles).

Supplementary Table 5: Descriptive Statistics of the Variables Used for the South, Middle, and North Sections of City of San Diego

a) South (N = 66)	Minimum	Maximum	Mean	Std. Deviation
Mental Health Prevalence (%)	8.7	19.1	14.3	2.7
Median Household Income ($)	$21,657	$96,319	$49,033	$18,469
Percent Degree (%)	5.5	82.3	23.3	13.8
Hours Worked (hrs/wk)	33.5	43.0	36.8	1.6
Housing Value ($) (N = 275)	$134,700	$771,100	$298,729	$92,509
Affordability (years) (N = 275)	3.6	16.5	6.8	2.9
Percent Below Poverty (%)	5.7	50.4	23.8	12.4
Percent White (%)	0.8	70.5	13.4	14.8
Percent Black (%)	0.2	38.2	9.9	8.4
Percent Hispanic (%)	11.5	98.3	61.5	24.1
Regular Checkup (%)	61.2	72.2	66.5	2.9
Percent Uninsured (%)	6.1	32.5	20.7	6.7
Crime Density (#/mi^2)	12.3	1384.3	230.3	294.7
Distance to Greenspace (ft)	417.0	5352.7	2623.8	1261.2
Distance to Tree Cover (ft)	251.9	2759.9	865.7	462.4
Distance to Parks (ft)	160.2	8048.5	2215.6	1262.3
Total Population (# people)	1847	16606	5094	2050

b) Middle (N = 73)	Minimum	Maximum	Mean	Std. Deviation
Mental Health Prevalence (%)	7.7	18.5	12.5	2.9
Median Household Income ($)	$22,713	$125,370	$58,858	$23,992
Percent Degree (%)	11.6	76.7	47.8	19.6
Hours Worked (hrs/wk)	27.3	43.4	37.8	2.9
Housing Value ($) (N = 275)	$180,900	$1,038,200	$454,236	$195,496
Affordability (years) (N = 275)	4.6	16.3	8.0	2.3
Percent Below Poverty (%)	2.1	52.6	18.1	12.5
Percent White (%)	4.5	97.9	52.5	27.7
Percent Black (%)	0.0	28.7	8.2	6.9
Percent Hispanic (%)	5.0	72.2	29.0	17.2

Regular Checkup (%)	60.1	72.8	65.6	2.8
Percent Uninsured (%)	4.3	27.4	12.1	6.4
Crime Density (#/mi^2)	39.3	538.4	251.2	131.2
Distance to Greenspace (ft)	0.0	4621.7	1866.8	1041.1
Distance to Tree Cover (ft)	73.3	1424.5	449.9	297.7
Distance to Parks (ft)	0.0	4386.1	1879.2	998.6
Total Population (# people)	1723	7157	4359	1375

c) North (N = 141)	**Minimum**	**Maximum**	**Mean**	**Std. Deviation**
Mental Health Prevalence (%)	6.5	15.9	9.9	1.7
Median Household Income ($)	$36,927	$186,738	$89,746	$32,399
Percent Degree (%)	19.8	88.2	62.5	15.0
Hours Worked (hrs/wk)	27.2	52.6	38.6	2.4
Housing Value ($) (N = 275)	$124,500	$1,720,900	$583,318	$268,505
Affordability (years) (N = 275)	2.2	17.7	6.6	2.3
Percent Below Poverty (%)	1.0	41.1	9.9	8.0
Percent White (%)	15.4	94.0	59.0	18.4
Percent Black (%)	0.0	15.5	3.0	3.1
Percent Hispanic (%)	3.8	50.8	14.2	7.7
Regular Checkup (%)	55.5	84.6	67.7	4.2
Percent Uninsured (%)	4.0	19.4	7.6	2.4
Crime Density (#/mi^2)	2.0	1356.5	99.5	178.5
Distance to Greenspace (ft)	26.3	4534.2	1211.5	888.6
Distance to Tree Cover (ft)	96.7	2125.8	470.7	367.8
Distance to Parks (ft)	162.7	7685.5	1834.1	1375.1
Total Population (# people)	1572	19414	5014	2420