Predicting and mapping neighborhood-scale health outcomes: machine learning approaches

Chen Feng[1], Junfeng Jiao[1], Yefu Chen[1]
[1]Urban Information Lab, SOA

GOOD SYSTEMS
A UT Grand Challenge

URBAN INFO LAB
The University of Texas at Austin

TEXAS
The University of Texas at Austin
School of Architecture

Introduction

Over half of the world's population now live in urban areas, and as cities are becoming the predominant mode of living, there is an increasing need to monitor and advance the health of city residents. While previous studies provide accurate local level estimates of health outcomes, health risks, and health behaviors, they are costly, time-consuming, and thus hard to be kept up to date.

Research on how neighborhood and community-level factors affect the health of residents have received attention (Diez Roux & Mair, 2010; Giles-Corti et al., 2016). The two broad domains of neighborhood attribute frequently studied are features of the physical environment and features of the social environment of neighborhoods. The physical environment includes aspects of the built environment, and most evidence points to the link between the level of social cohesion and mental health problems, especially the level of depressive symptoms (Aneshensel & Sucoff, 1996; Mair et al., 2009).

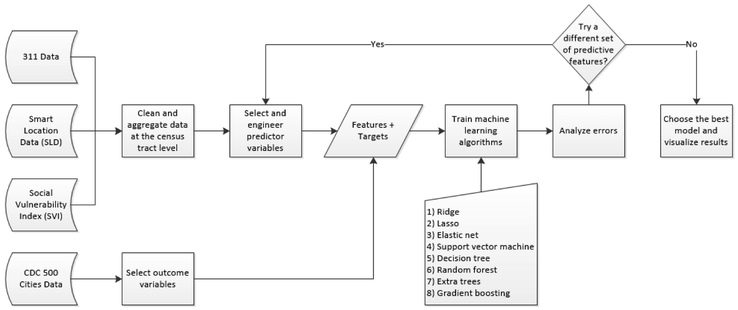
Research Goal

This study aims to conduct a machine-learning method analyzing the impacts of physical environmental and social cohesion on health status. The method and findings aim to help planners and city officials predict health outcomes, identify potential health issues, and evaluate neighborhood-level interventions to improve the health of residents at a relatively low cost and in a timely manner.

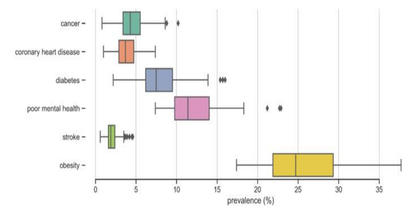
Methods

The workflow consists of five major parts: (a) data preparation, (b) feature engineering and selection, (c) training and tuning, (d) model comparison, and (e) visualization. The outcome variables are the prevalence of different health outcomes or health behaviors, and the predictor variables are quantitative descriptions of the social, physical, or perceived neighborhood environments.

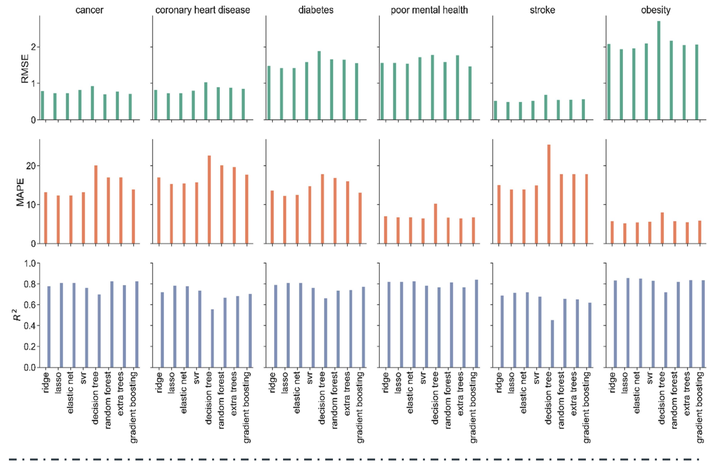
A variety of machine learning algorithms are applied and compared, including Ridge Regression, Lasso Regression, Elastic Net, Support Vector Machine, Decision Tree, Random Forest, Extra Trees, and Gradient Boosting, and we use root mean squared error (RMSE), MAPE, and R-square as the primary indicators of predictive accuracy.

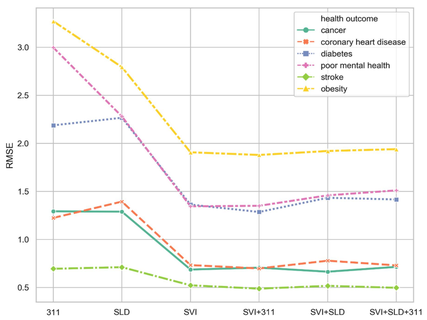


Figures and Results



As shown in the left figure, the tract-level prevalence varied greatly for different health outcomes. The following figure shows the performances of the eight different algorithms (trained with the complete set of features) were generally comparable.





As shown in the left figure, including features extracted from SVI data in our models always led to a significant drop in RMSE, which confirms the previous findings that sociodemographic factors are key determinants of population health (Luo et al., 2015). Adding more features did not necessarily improve the predictive accuracy.

Also, the 311 data may have captured some aspects of the neighborhood environments that could affect the health outcomes under study. By contrast, the measures of the neighborhood built environment only had a minor (and sometimes even adverse) effect on improving the models' predictive power.

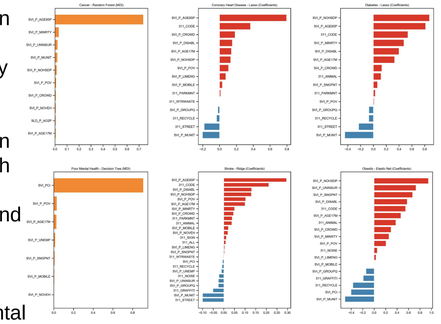
Health outcome	Algorithm	Datasets
Cancer	Random Forest	SVI + SLD
Coronary heart disease	Lasso Regression	SVI + 311
Diabetes	Support Vector Machine	SVI + 311
Poor mental health	Decision Tree	SVI
Stroke	Ridge Regression	SVI + 311
Obesity	Elastic Net	SVI + 311

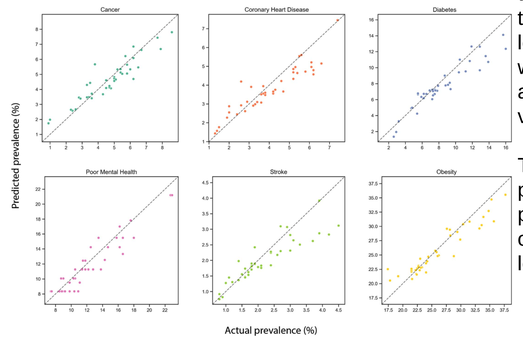
Figures and Results

The proportion of the population aged 65 and above was a key determinant of cancer, coronary heart disease, and stroke.

The proportion of the population aged 25 and above with no high school diploma had much predictive power for diabetes and obesity.

Per capita income was the strongest predictor of poor mental health, and the frequency of 311 service requests related can be an important predictor of the prevalence of coronary heart disease, diabetes, stroke, and obesity.





In predicting the prevalence of cancer, coronary heart disease, and obesity, the predictive models tended to overestimate the prevalence at the lower range of values while underestimating it at a higher range of values.

The model for predicting the prevalence of stroke only worked well at the lower range of values.

Conclusion and Reference

The study yielded several interesting findings. Our analysis showed that, even at the census tract level, the sociodemographic and socioeconomic factors were the most powerful predictors of the prevalence of the health outcomes; however, we caution against overinterpreting this finding and rushing to conclude that the built environment has little or no impact on population health because our models did not explicitly address the causal pathways linking the different factors to health outcomes. Also, in several cases, the additional predictive power contributed by the 311 data is particularly encouraging. Finally, no single algorithm always outperforms other algorithms in predicting health outcomes—a variety of algorithms need to be tried in order to achieve the best results.

- A.V. Diez Roux and C. Mair. Neighborhoods and health. Annals of the New York Academy of Sciences, 1186 (1) (2010).
- C. Feng and J. Jiao. Predicting and mapping neighborhood-scale health outcomes: A machine learning approach. Computers, Environment and Urban Systems, 85 (2021).
- W. Luo et al. Is demography destiny? Application of machine learning techniques to accurately predict population health outcomes from a minimal demographic dataset. PLoS One, 10 (5) (2015),

The University of Texas at Austin
Bridging Barriers
Office of the Vice President for Research