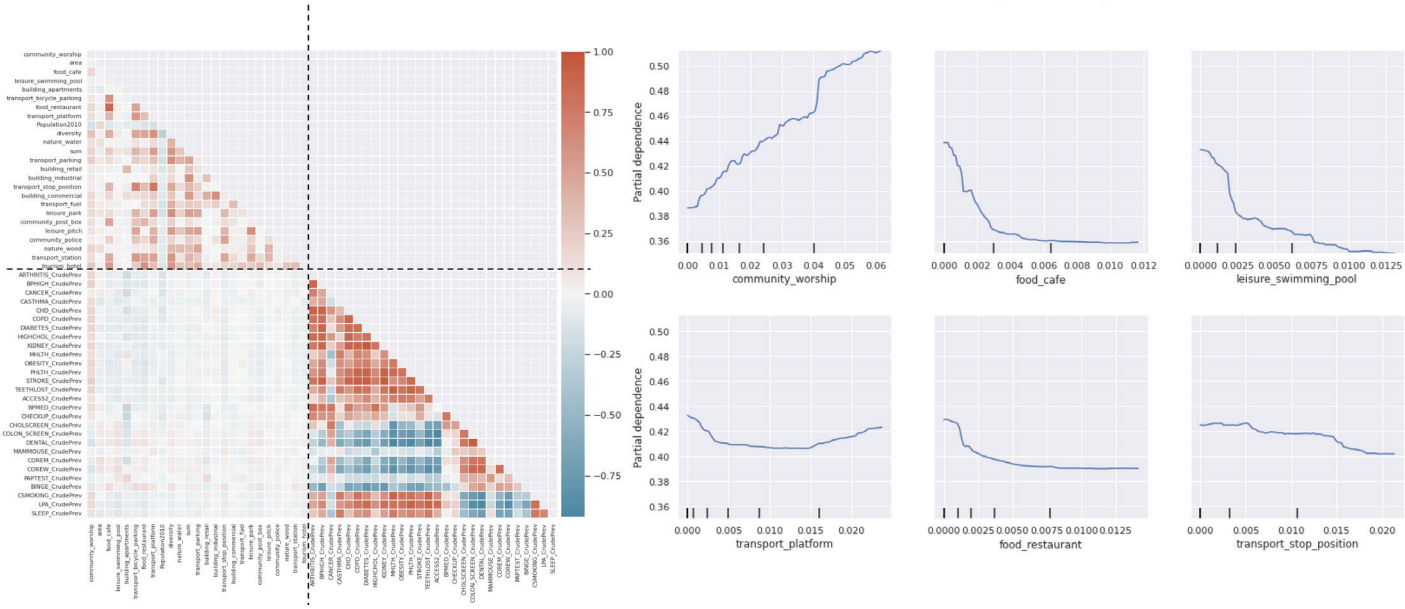
A POI-Based Machine Learning Method in Predicting Health

Predicting Residents' Health Status and Implications for Healthy City Planning

Shicong Cao
Heinle, Wischer und Partner Freie Architekten

Hao Zheng
University of Pennsylvania Stuart Weitzman School of Design



ABSTRACT

This research aims to explore the quantitative relationship between urban planning decisions and the health status of residents. By modeling the Point of Interest (POI) data and the geographic distribution of health-related outcomes, the research explores the critical factors in urban planning that could influence the health status of residents. It also informs decision-making regarding a healthier built environment and opens up possibilities for other data-driven methods. The data source constitutes two data sets, the POI data from OpenStreetMap, and the CDC dataset PLACES: Local Data for Better Health. After the data is collected and joined spatially, a machine learning method is used to select the most critical urban features in predicting the health outcomes of residents. Several machine learning models are trained and compared. With the chosen model, the prediction is evaluated on the test dataset and mapped geographically. The relations between factors are explored and interpreted. Finally, to understand the implications for urban design, the impact of modified POI data on the prediction of residents' health status is calculated and compared. This research proves the possibility of predicting residents' health from urban conditions with machine learning methods. The result verifies existing healthy urban design theories from a different perspective. This approach shows vast potential that data could in future assist decision-making to achieve a healthier built environment.

Above left: Correlation between selected features and output variables

Above right: Partial dependence plot of major input variables and prevalence of obesity

INTRODUCTION

Health and POI Data

Healthy living is a goal of many 21st century cities. The World Health Organization's Healthy Cities project has identified urban planning principles supporting health and example cities whose development can be learned from (Duhl and Sanchez 1999). The definition of health in the constitution of the WHO is a state of complete physical, mental and social wellbeing and not merely the absence of disease or infirmity (Kelley 2008). Health is, therefore, a social issue that needs to be addressed systemically and not only from a medical care point of view.

Studies show that about 60% of people's health depends on their lifestyle and the environment (Schroeder 2007). In contrast, only 3% of United States health expenditure goes into public health activities. The U.S. has a higher healthcare expenditure as a percentage of GDP than any other developed country but the lowest healthcare performance (Battisto and Wilhelm 2019). This makes us question the effectiveness of the policy to shape our healthcare system through the lens of medical care rather than through that of more encompassing public health.

Big data analytics, artificial intelligence, and other emerging technologies make it possible to understand and evaluate the effect of the built environment on residents' health quantitatively. The available data source, data processing procedures, and interaction technologies are poised to revolutionize urban management (Engin et al. 2020). By learning from data, solutions that benefit the community in the long term could be discovered and communicated. Data-driven process empowers the local community with the evidence they need to make better decisions for their city and neighborhoods.

Research Background

The link between the built environment and health has long been acknowledged. Research shows that there is interdependence between environments and individual behavior (Macintyre, Ellaway, and Cummins 2002). Primarily there are three domains where urban planning can most effectively focus support for health and wellbeing—physical activity, community interaction, and healthy eating—since these domains address some of the significant risk factors for chronic diseases (Kent and Thompson 2014). But the limitation of the traditional urban research method is relying on qualitative methods, sometimes without hard evidence.

To estimate the obesity rate, one of the machine learning models, Convolutional Neural Network (CNN), has been used to analyze the satellite image (Newton et al. 2020). Analysis of the convolutional layers suggests which visual features are more critical for a low obesity rate. The limitations of the imagery method are the amount of computing it requires and the obscurity of the conclusion due to the restriction of the dataset and the black box effect of the algorithm. Street view imagery is also used as a source of data to measure visual walkability (Zhou et al. 2019). This approach considers the human perception of the built environment. The amount of data processing and redundancy could be the problem.

The use of OpenStreetMap (OSM) data to generate socio-economic indicators and urban crime risk has been studied and testified (Feldmeyer et al. 2020; Cichosz 2020). The data processing method can be used for reference, and it showcases that POI data can be a good indicator of urban conditions and activities. Urban POI data analysis can also be integrated with other methods of data collection. POI data, location-based service positioning data, and street view images are used in conjunction to measure greenway suitability and give suggestions on greenway networks planning (Tang et al. 2020).

Objectives

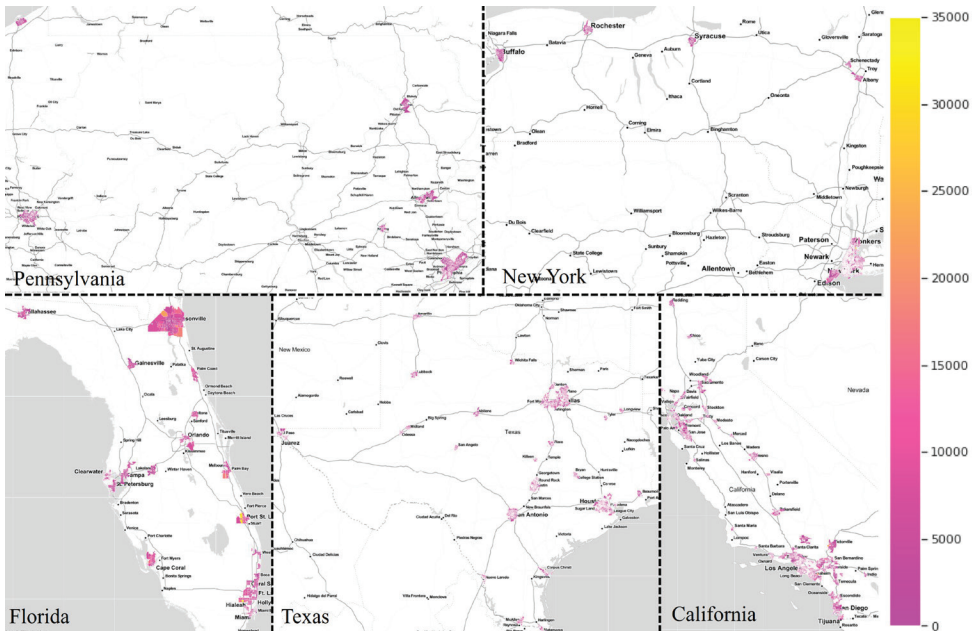
This research aims to explore using open-source city Point of Interest (POI) data to predict the health status of the residents using machine learning methods. By modeling the Point of Interest (POI) data and the geographic distribution of health-related outcomes, the research evaluates the key factors in urban planning that support the residents' health and wellbeing. The data processing and modeling methods inform decision-making to support a healthier built environment and open up possibilities for other data-driven methods.

METHODOLOGY

The workflow of this research follows five steps. In the first step, POI data from OpenStreetMap for the study area were collected and spatially joined with the health-related outcomes data within the census tract boundary. Second, Principal Component Analysis (PCA) and correlation analysis were conducted after some initial data cleaning. Third, a set of machine learning models were trained, and feature importance was calculated for feature selection. Fourth, the selected features were used to train machine learning models, and the results evaluated. Finally, the effect of modification of input variables on output variables was interpreted.

Data Source

The data source constitutes two data sets, the POI data



1 Sample area and spatial distribution of population

1

from OpenStreetMap, and 500 Cities Project dataset from the CDC. The test regions are within the five most populous states of the United States—California, Texas, Florida, New York, and Pennsylvania—and are a compromise between data availability, handling capacity, and statistical accuracy.

- 500 Cities: Local Data for Better Health. 500 Cities is a project that provides city and census tract-level small area estimates for chronic disease risk factors, health outcomes, and clinical preventive services use for the largest five hundred cities in the United States (CDC n.d. "500 Cities Project"). The dataset is generated with an innovative peer-reviewed multilevel regression and poststratification (MRP) approach that links geocoded health surveys and high spatial resolution population demographic and socioeconomic data. The twenty-eight measures include four unhealthy behaviors, fourteen health outcomes, and ten prevention practices. The measures include major risk behaviors that lead to illness, suffering and early death related to chronic diseases and conditions, as well as the conditions and diseases that are the most common, costly, and preventable of all health problems (CDC n.d., "Places"). The population size of each census tract is also included as a column in the dataset as well as the census tract boundaries. These small area estimates allowed cities and local health departments to understand better the geographic distribution of health-related variables in their jurisdictions and assisted them in planning public health interventions.
- OpenStreetMap is an open-source database with volunteers mapping geographic elements of the world. It represents physical features on the ground using tags attached to its basic data structures (its nodes, ways, and relations) (Feldmeyer et al. 2020). The research uses Overpass API to query the database by tags to get the geographic location of certain features. Fifty features were initially selected—based on the relation with physical wellbeing and the abundance of data points—and can be categorized into food, healthcare, transportation, community service, leisure, tourism, building, nature, and shop. Since the OpenStreetMap is user-generated data, there are various levels of completeness and accuracy of features. Overall, the data quality in the U.S. is good enough for POI analysis (Idham Muttaqien 2017; Barrington-Leigh and Millard-Ball 2017).

The two datasets are spatially joined within the boundary of each census tract. There are in total 12588 rows; namely, 12588 census tracts were sampled. As Figure 1 shows, the sample area consists of the major cities within each state. Most census tracts have a moderate population while some have a substantial population.

Two columns are added—a column describing the total count of POI and another column describing the total number of POI categories available within each census tract—to get a better understanding of the density and diversity of POI points.

Feature Extraction

POI Data Exploration and Preparation

Some initial data exploration showed that there are rows

with zero or minimal data points. The lower quartile of rows with a total number of POI less than 12 were dropped from the dataset since it provided little information. The reason for lacking data points could be due to the completeness of OSM or fewer activities within the area. The model performance improved after dropping this part of the data.

A boxplot (Figure 2, upper left) shows that most census tracts have a moderate number of POI while there are also many outliers with a large number of POI, representing larger and denser census tracts.

Considering the different sizes and densities of the census tracts, we divide the POI count by population and area and get the POI counts per capita and per unit area. A principal component analysis (PCA) shows different patterns of cumulative explained variance for each data manipulation method (Figure 2, upper right, lower left and right). For the original total POI count data, a few features can explain most of the variance. The curve is flatter for the per-unit-area data, which means that variance is more spread out. The per-capita data has the best balance between redundancy and bias. It is more suitable for feature selection. It also makes more sense to evaluate the number of POIs serving the designated population.

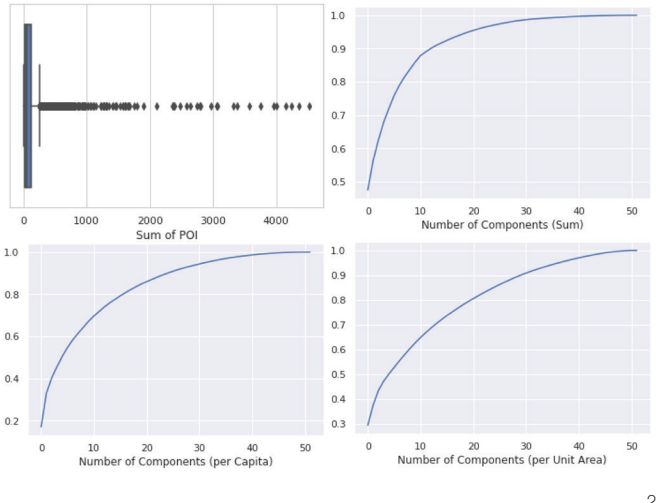
Health Data Exploration and Preparation

A boxplot (Figure 3) of the test data shows that the health data have different ranges, including some outliers. Since there are no explicit patterns, we use Tukey's rule to remove outliers and set the outer range to three-interquartile range (IQR). The outliers are assigned either to the upper fence value or lower fence value.

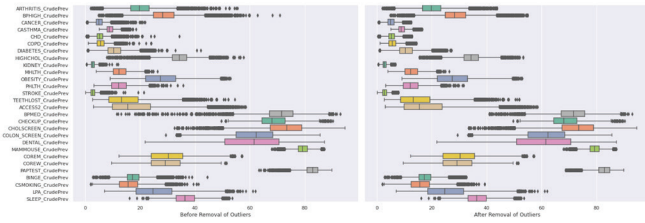
A mapping of the health outcome shows that there are some spatial patterns. As in Figure 4, the prevalence of coronary heart disease is higher in Florida, where there is a higher percentage of older population. The prevalence is lower in New York City, where there are more young people (United States Census Bureau n.d.).

Correlation Analysis

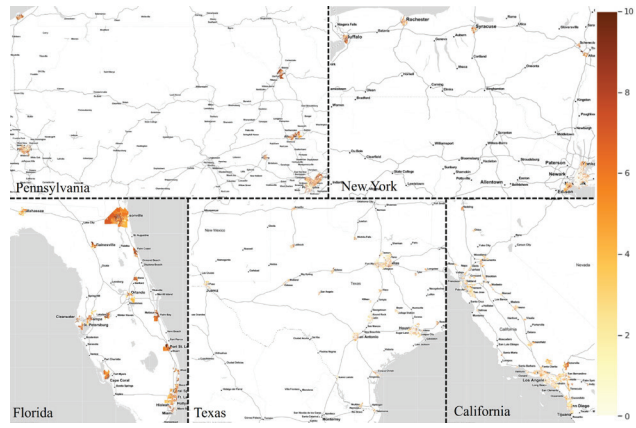
A Pearson correlation coefficient heat map, as on Figure 5, showed correlations among input variables both within and among categories. The category of food has a strong correlation within itself. It also correlates with public transportation, bike parking, hotels, and shops. The healthcare category does not connect with other features. The transportation category has some correlation within itself and with leisure and tourism categories. Leisure has some correlation within itself except for the swimming pool criteria; as a category, it correlates with



2



3



4

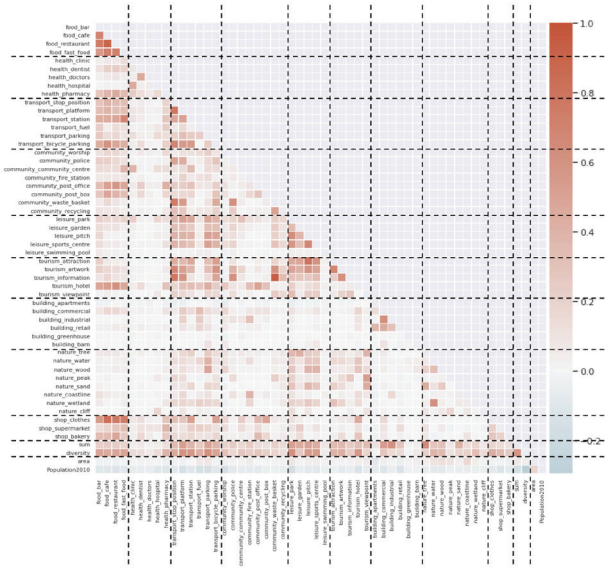
2 Upper left: Distribution of total POI counts; Upper right: Cumulative explained variance for POI counts; Lower left: Cumulative explained variance for POI counts divided by population; Lower right: Cumulative explained variance for POI counts divided by area.

3 Distribution of health data before and after removal of outliers

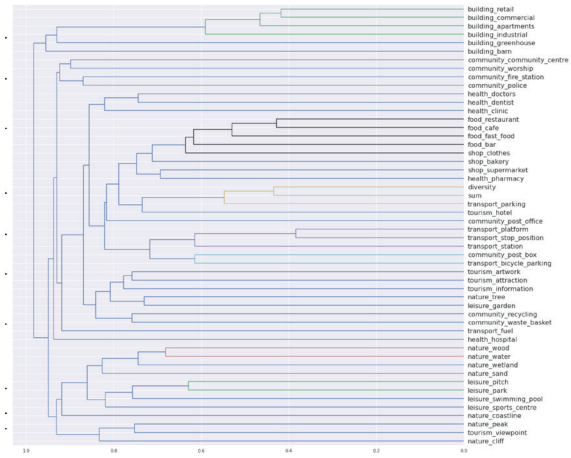
4 Spatial distribution of prevalence of CHD

transportation, tourism, and moderately with nature. There is some correlation between commercial, industrial, retail, and apartment buildings, and the building categtory does not correlate with other categories. Water correlates with wetland within the nature category. Shops correlate with food services and moderately with transportation.

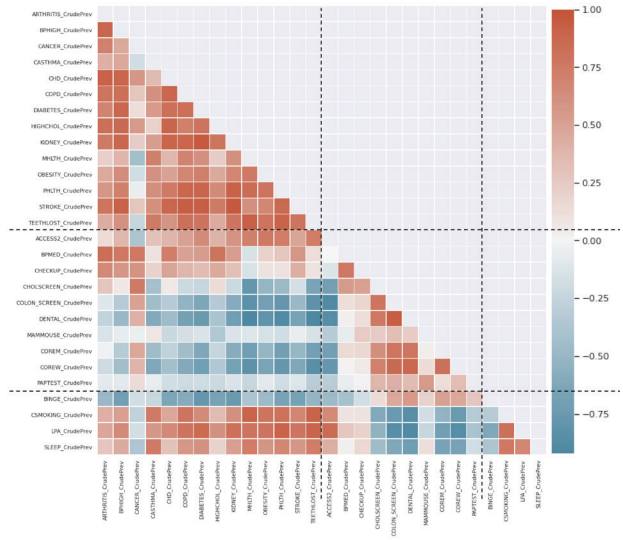
Hierarchical clustering groups similar objects into clusters and helps us to understand the relationships among



5



6



7

5 Correlation heatmap of input variables

6 Hierarchical clustering of input variables

7 Correlation analysis of output variables

features. As shown in Figure 6, there is a similarity between building_retail and building_commercial. They also belong to the same cluster with building_apartment and building_industrial, indicating the building density of an area. Food_restaurant and food_cafe have similarities, and they also belong to the same cluster with other food services and shop_clothes, which might indicate the prosperity of retail in the area. Diversity, sum, and transportation_parking belong to the same cluster, indicating the overall abundance of POI. Transport_platform, transport_stop position, and transport_station are within the same cluster, indicating the accessibility by public transportation. There is also a similarity between community_post _box and transport_bicycle_parking. Nature_wood and nature_water belong to the same cluster, which might indicate natural resources. Leisure_pitch and leisure_park have similarities that might indicate the recreation and outdoor activities within an area.

Among the output variables, there is overall a strong correlation, as shown in Figure 7. Bad health outcomes correlate positively within themselves and unhealthy behaviors and negatively with prevention measures. However, cancer seems to have a negative correlation with some bad health outcomes while positive with preventions. Within the prevention group, taking blood pressure medicine and routine checkup in previous years positively correlates with harmful health outcomes, while others correlate negatively. Within the unhealthy behaviors group, binge drinking correlates negatively with bad health outcomes and positively with preventions, which might indicate a correlation of better health and socioeconomic status (CDC 2012).

Machine Learning

An initial model training was conducted for model selection. We implemented six machine learning models, as follows. A Random Prediction model is implemented with random values within the test data range are generated. Then a Linear Regression is conducted as a basic statistical prediction. A Decision Tree model is a simple non-linear machine learning algorithm, where the data is continuously split according to a specific parameter. The Random Forest model is an ensemble learning method that operates by constructing many decision trees at training time. K-nearest Neighbors algorithm is a non-parametric classification method that returns the average values of k-nearest neighbors. Artificial Neural Network (ANN) is a deep learning method that digitally mimics the human brain to predict values. A 5-layer neural network is used in the research, and a training step of 4000 achieves the best accuracy.

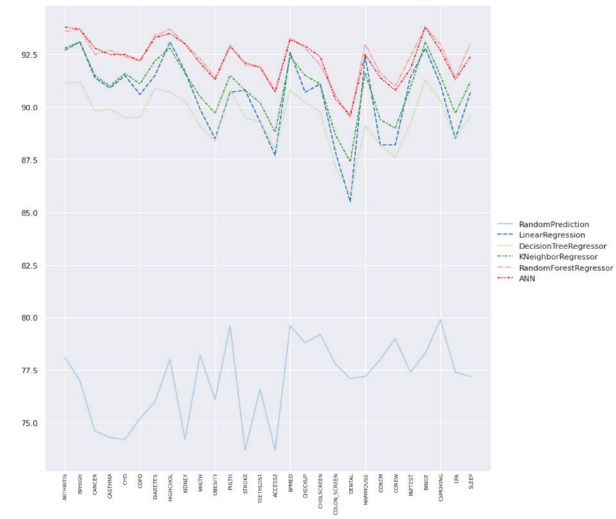
Model	Mean Absolute Error (Before/After)	Median Accuracy (%) (Before/After)
Random Prediction	0.2567/0.2572	77.20/77.09
Linear Regressor	0.1154/0.1151	90.62/90.53
Decision Tree Regressor	0.1359/0.1368	89.51/89.425
Random Forest Regressor	0.0961/0.0961	92.33/92.34
K-Neighbors Regressor	0.1129/0.1111	90.85/90.99
Artificial Neural Network	0.1003/0.0987	91.98/92.26

Table 1 Model Performance before and after feature selection

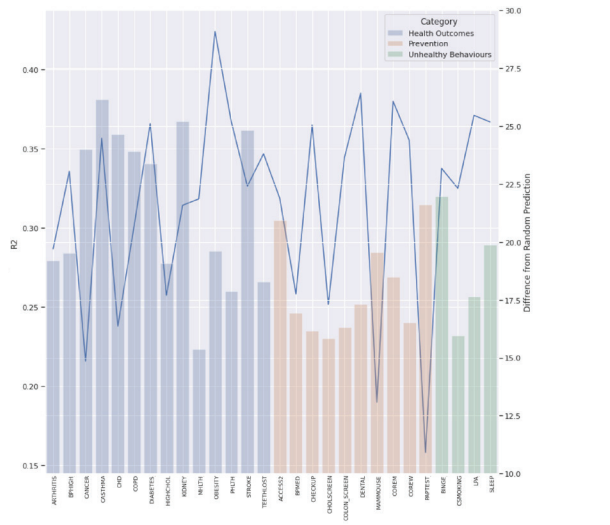
Since the data were standardized into the range of 0 to 1, the mean absolute error could represent the performance of the model. The median accuracy was calculated for each model with the test dataset. An average accuracy rate for the 28 predictions is calculated as the output.

Furthermore, as there are correlations among features, a permutation method is used for calculating the feature importance of the Random Forest model instead of the built-in impurity method of sklearn. The idea is to permute the values of each feature and measure how much the permutation decreases the accuracy of the model. Therefore, the respective importance of correlated features will not be shared with one another, and scores will not be reduced (Strobl et al. 2007).

We selected the features that had an importance score over 0.005 and retrained the model from the result of the feature importance calculation. The performance of the model improved slightly. We also did an R-squared score analysis for each of the output variables to evaluate to what extent the model's variance can be explained. Moreover, a partial dependence plot was used to understand the relation between deciding features and the output variable. The plot was implemented with sklearn plot_partial_dependence. We also mapped the prediction accuracy to see if there is any spatial pattern for prediction accuracy. Lastly, input variables were changed by a certain percentage, and the effect on the output variable was measured and compared.



8



9

8 Model Performance on All the Output Variables

9 Outperformance over Random Prediction and R-squared Score of Random Forest Model on All the Output Variables

RESULT AND DISCUSSION

Model Evaluation

As can be seen from Table 1, before feature selection, the Random Forest model has the lowest mean absolute error and highest median accuracy.

After feature selection, the ANN model's performance has increased the most because of removing irrelevant features and avoiding overfitting. The performance of K-Neighbors Regressor increased by a little. The version of Random Forest Regressor stays the same and still has the best performance among all the models. The performance of Decision Tree Regressor, Linear Regressor, and Random Prediction decreased by a little, probably because of fewer input variables.

A plot of different models' performance on each output variable (Figure 8) shows that machine learning models

perform better than random prediction for all health outcomes. The Random Forest model has the best performance on all the output variables.

The overall R-squared score for Random Forest Model is 0.3088, which means that the model could explain about one-third of the variance of the health outcomes. This result is in line with the perception that part of people's health depends on their lifestyle behaviors and environmental exposure. As on Figure 9, a plot of the R-squared score on each output variable shows that the model could best explain the variance for the prevalence of obesity, which has an R-squared score of 0.4239.

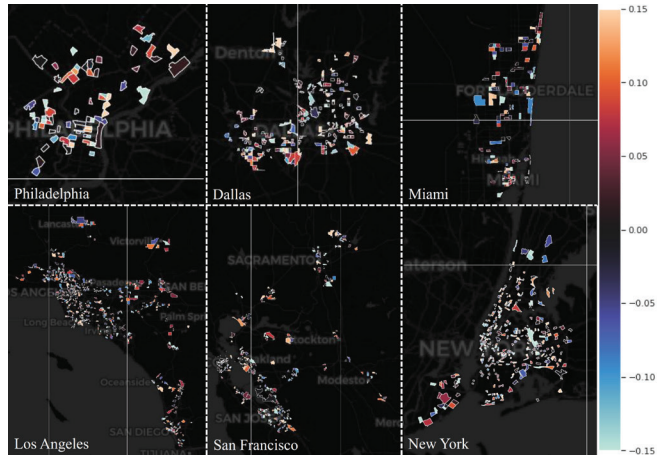
As in the bar plot part of the right image of Figure 9, we calculated the percentage of the increased model performance of the Random Forest model compared to random prediction. The model has better prediction accuracy for health outcomes than prevention and unhealthy behaviors, especially for the prevalence of cancer, asthma, coronary heart disease (CHD), chronic obstructive pulmonary disease (COPD), diabetes, chronic kidney disease, and stroke.

A mapping of the difference of prediction for CHD from ground truth shows that most of the prediction is close to the ground truth, as is shown in Figure 10. There is no clear pattern of overestimation or underestimation spatially.

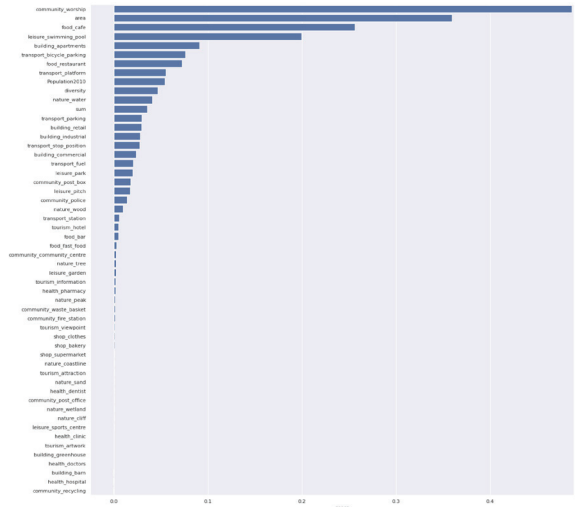
Feature Importance

As shown in Figure 11, the permutation feature importance of the Random Forest model shows that community_worship has by far the highest importance, followed by area, food_cafe, leisure_swimming pool. The next tier includes building_apartments, transport_bicycle_parking, food_restaurant, transport_platform, population, POI diversity, natural_water, and POI sum. The other features that influence the model accuracy are transport_parking, building_retail, building_industrial, transport_stop_position, building_commercial, transport_fuel, leisure_park, community_post_box, leisure pitch, community_police, and natural_wood.

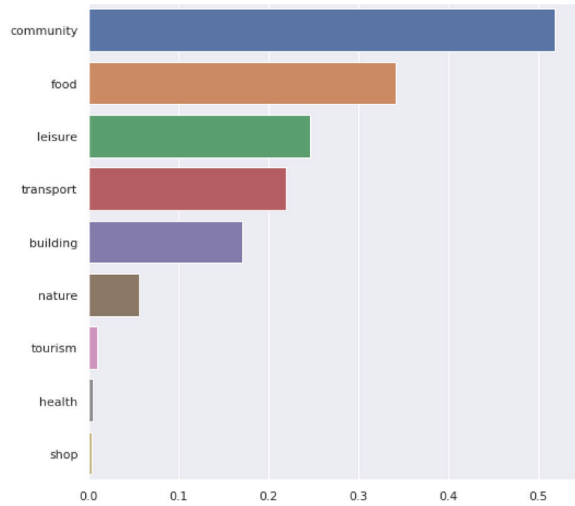
By adding the feature importance score of each category, as shown in Figure 12, we can see that community service has the highest cumulative feature importance, followed by food, leisure, transport, building, and nature. In contrast, tourism, health, and shop categories have little impact on the model's accuracy. Also, apart from places of worship, the rest of the community category does not have a lot of impact on the model's accuracy.



10



11

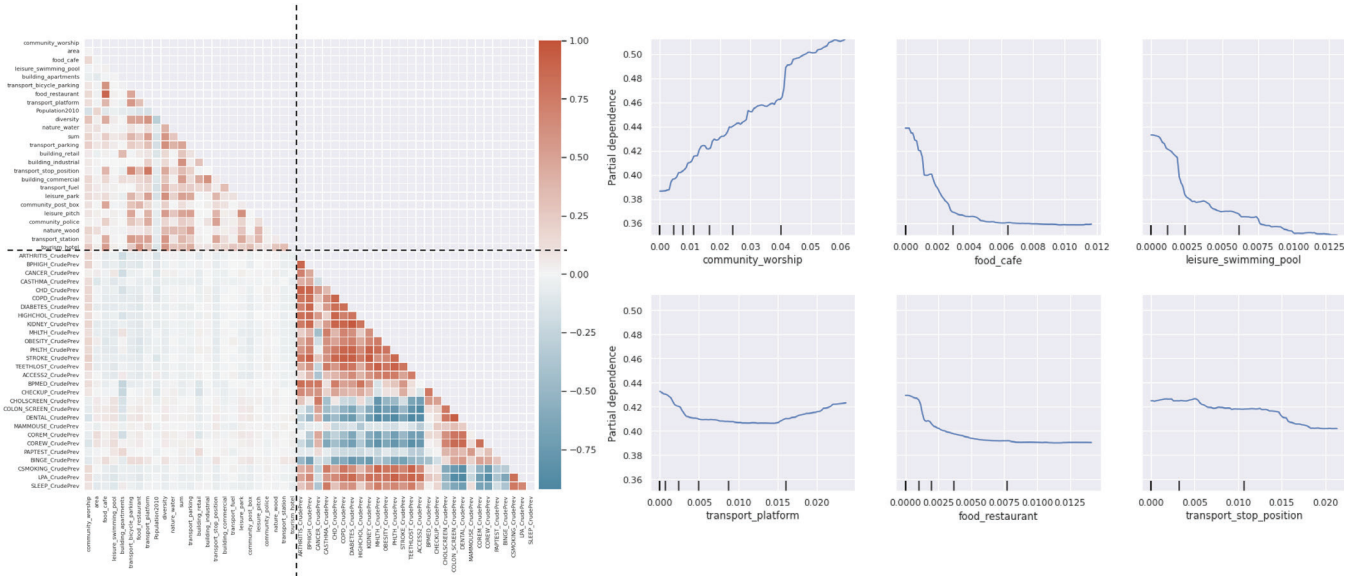


12

10 Spatial distribution of prediction difference from ground truth for CHD

11 Feature importance of random forest regressor for CHD

12 Accumulated feature importance for eight POI categories



13 Left: Correlation between selected features and output variables; Right: Partial dependence plot of major input variables and prevalence of obesity

The four most important categories support the theory that physical activity, community interaction, and healthy eating could best support health and wellbeing (Kent and Thompson 2014). Building density and use have some influence on health outcomes. Because the sample area is primarily within the urban context, nature's influence is relatively small. It is counterintuitive to see that the health category has little prediction power on residents' health.

Partial Dependence Analysis

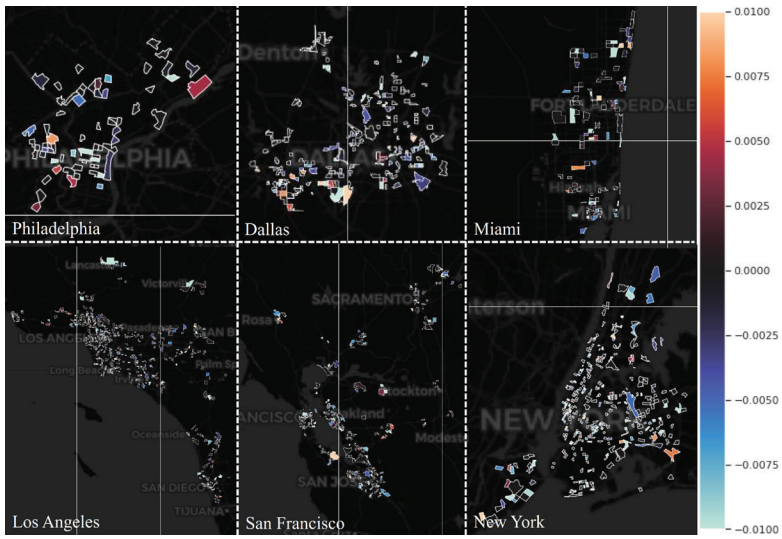
The heatmap of selected input data and output data (Figure 13, left) shows no apparent correlation between the input and output data. A slight positive correlation between community worship per capita and bad health outcomes compared to many other features has a slight negative correlation with bad health outcomes. Previous research shows that religions are relevant for poverty. More churches in the community might indicate lower socioeconomic status, which is relevant for worse health outcomes, especially as a source of normativity and motivation. The twofold influence is particularly exercised in the three areas of business and finance, politics and culture, education, and health (Sedmak 2019).

A partial dependence plot (Figure 13, right) depicts how the predictions partially depend on values of the input variables of interest (Wright 2018). In this case, we plot the relationship of the prediction of obesity and the six features with the highest feature importance scores. The prediction prevalence of obesity increases as places of worship per capita rise and decreases as the number of cafés, swimming pools, and restaurants increases. However, not many data points with large values are available, so the reliability of the estimates decreases in those areas. The prediction decreases as the number of stop positions increases, implying that public transportation has a health benefit. The prediction decreases and increases as the number of platforms increases, indicating a difference in urban landscape related to railways.

Modification of POI data and the Change of Prediction

Then we modified several POI features and made the prediction again. By decreasing the number of places of worship, we got a new prediction for the prevalence of obesity on the test set. As shown in Figure 14, most predictions go down while some predictions go up, which reflects the non-linearity of the model. On average, the prediction of the prevalence of obesity decreased by 0.29%. By increasing the number of cafés by 10 percent, the average prediction decreased by 0.05%, and by increasing the number of swimming pools by 10%, the average prediction dropped by 0.09%. The effect of increasing the amount of stop positions was to move the prediction down by 0.03%.

The result is in line with that of the partial dependence plot. The influence of the number of places of worship is still greatest, followed by the number of swimming pools, cafés, and stop positions. We think that it is a result of both the socioeconomic status and the lifestyle that these features indicate (Gelormino et al. 2015). Although there is no direct causality established, it is noticeable that these factors are not related to the medical care system as usually expected. Urban planners could use this machine learning prediction as a tool to evaluate quantitatively which measures potentially have the best return on investment. It is also an effective way to create a consensus among participants on what aspects of the built environment could be improved to support residents' health.



14 Change of prediction after decreasing the number of worship places per capita by 10%

14

CONCLUSION AND DISCUSSION

This research explores a method to predict the health outcomes of residents by using the POI data available from OSM, thus exploring the linkage between the built environment and the health of residents. Different machine learning methods were used and evaluated. The result shows that the Random Forest model strikes the best balance between prediction accuracy and ease of implementation. The interpretation of the machine learning model suggests some critical features that potentially influence residents' health. The result supports as well as complements the existing healthy urban design theory.

This research showcased the possibility that big data analysis could benefit the urban design process. Urban planners could use this method to discover and support ideas that improve the health status of residents. With the increasing availability of data and computational resources, there is considerable potential that data could change how a healthy city is designed and become common ground for decision-making.

For the next step of research, a more comprehensive analysis of different urban characteristics would be helpful, for instance, the building code, density, vehicular and pedestrian traffic environments, and medical state benefits. Comparing other countries would also give a more complete picture of the most important factors for residents' health within different contexts.

One of the challenges of this research is the data source. Although the quality of OpenStreetMap data in the U.S. is relatively good, it is not complete everywhere (Idham Muttaqien 2017). The reliability of POI data can affect the final result. In the future, if we have more comprehensive and real-time data, predictions could be more accurate. Another challenge is to understand the correlation and causality between built environment and health. The built environment results from socioeconomic factors, and it also shapes the way people live (Williams et al. 2009). It is essential to have a case-by-case understanding of how a specific feature influences the health status of residents. As more relationships between the built environment and health may be established through data, there is a challenge and an opportunity for traditional urban planning practice. By making data more explicable, the decision-making process could become more transparent. The community would be empowered to make healthier decisions for themselves in the long run.

ACKNOWLEDGEMENTS

The code of the research is available at: https://github.com/shicong0720/Finding-Healthy-Community-Design-Criteria-with-AI

REFERENCES

Barrington-Leigh, Christopher, and Adam Millard-Ball. 2017. "The World's User-Generated Road Map Is More than 80% Complete." *PLoS ONE* 12 (8): 1–20. https://doi.org/10.1371/journal.pone.0180698.

Battisto, Dina, and Jacob J. Wilhelm. 2019. *Architecture and Health: Guiding Principles for Practice*. London: Routledge.

Centers for Disease Control and Prevention (CDC). 2012. "Vital Signs: Binge Drinking Prevalence, Frequency, and Intensity among Adults - United States, 2010." *Morbidity and Mortality Weekly Report* 61 (1): 14–19. http://www.ncbi.nlm.nih.gov/pubmed/22237031.

Centers for Disease Control and Prevention (CDC). n.d. "Places: Local Data for Better Health." Accessed 15 September 2021. https://www.cdc.gov/places/methodology/index.html.

Centers for Disease Control and Prevention (CDC). n.d. "500 Cities Project: 2016 to 2019." Accessed 15 September 2021. https://www.cdc.gov/places/about/500-cities-2016-2019/index.html.

Cichosz, Paweł. 2020. "Urban Crime Risk Prediction Using Point of Interest Data." *ISPRS International Journal of Geo-Information* 9 (7). https://doi.org/10.3390/ijgi9070459.

Duhl, L.J, A.K. Sanchez, and World Health Organization. Regional Office for Europe. 1999. *Healthy Cities and the City Planning Process: A Background Document on Links between Health and Urban Planning*. Copenhagen: WHO Regional Office for Europe.

Engin, Zeynep, Justin van Dijk, Tian Lan, Paul A. Longley, Philip Treleaven, Michael Batty, and Alan Penn. 2020. "Data-Driven Urban Management: Mapping the Landscape." *Journal of Urban Management* 9 (2): 140–50. https://doi.org/10.1016/j.jum.2019.12.001.

Feldmeyer, Daniel, Claude Meisch, Holger Sauter, and Joern Birkmann. 2020. "Using OpenStreetMap Data and Machine Learning to Generate Socio-Economic Indicators." *ISPRS International Journal of Geo-Information* 9 (9): 1–16. https://doi.org/10.3390/ijgi9090498.

Gelormino, Elena, Giulia Melis, Cristina Marietta, and Giuseppe Costa. 2015. "From Built Environment to Health Inequalities : An Explanatory Framework Based on Evidence." *PMEDR* 2: 737–45. https://doi.org/10.1016/j.pmedr.2015.08.019.

Idham Muttaqien, Bani. 2017. "Assessing the Credibility of Volunteered Geographic Information: The Case of OpenStreetMap." MS diss., University of Twente. https://webapps.itc.utwente.nl/librarywww/papers_2017/msc/gfm/muttaqien.pdf.

Kelley, Lee. 2008. "The World Health Organization (WHO)." The World Health Organization (WHO), no. July 1994: 1–157. https://doi.org/10.4324/9780203029732.

Macintyre, Sally, Anne Ellaway, and Steven Cummins. 2002. "Place Effects on Health: How Can We Conceptualise, Operationalise and Measure Them?" *Social Science and Medicine* 55 (1): 125–39. https://doi.org/10.1016/S0277-9536(01)00214-3.

Newton, David, Dan Piatkowski, Wesley Marshall, and Atharva Tendle. 2020. "Deep Learning Methods for Urban Analysis and Health; Estimation of Obesity." In *ECAADe 2020: Health and Materials in Architecture and Cities*. Vol. 1. 297–304.

Schroeder, Steven A. 2007. "We Can Do Better — Improving the Health of the American People." *New England Journal of Medicine* 357 (12): 1221–28. https://doi.org/10.1056/NEJMsa073350.

Sedmak, Clemens. 2019. "Evidence-Based Dialogue: The Relationship between Religion and Poverty through the Lens of Randomized Controlled Trials." *Palgrave Communications* 5 (1): 1–7. https://doi.org/10.1057/s41599-019-0215-z.

Strobl, Carolin, Anne Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. 2007. "Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution." *BMC Bioinformatics* 8. https://doi.org/10.1186/1471-2105-8-25.

Tang, Ziyi, Yu Ye, Zhidian Jiang, Chaowei Fu, Rong Huang, and Dong Yao. 2020. "A Data-Informed Analytical Approach to Human-Scale Greenway Planning: Integrating Multi-Sourced Urban Data with Machine Learning Algorithms." *Urban Forestry and Urban Greening* 56 (January). https://doi.org/10.1016/j.ufug.2020.126871.

United States Census Bureau. n.d. "Quick Facts." n.d. Accessed 15 September 2021. https://www.census.gov/quickfacts/fact/table/US/PST045219.

Williams, Oli, Teresa Lavin, C. Higgins, Margaret Kelaher, Deborah J. Warr, Theonie Tacticos, Terrence D. Hill, et al. 2009. "Health Impacts of the Built Environment." *Institute of Public Health in Ireland*. Vol. 15.

Wright, Ray. 2018. "Interpreting Black-Box Machine Learning Models Using Partial Dependence and Individual Conditional Expectation Plots," 1950–2018. Accessed 15 September 2021. https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2018/1950-2018.pdf.

Zhou, Hao, Shenjing He, Yuyang Cai, Miao Wang, and Shiliang Su. 2019. "Social Inequalities in Neighborhood Visual Walkability: Using Street View Imagery and Deep Learning Technologies to Facilitate Healthy City Planning." *Sustainable Cities and Society* 50 (129): 101605. https://doi.org/10.1016/j.scs.2019.101605.

IMAGE CREDITS

All drawings and images by the authors.

Shicong Cao is a registered architect and researcher currently working at Heinle Wischer und Partner in Berlin. Her research interests focus on artificial intelligence and the healthy built environment. She holds a Master's degree in Architecture + Health from Clemson University, a Master's degree in Architecture from Politecnico di Milano and a bachelor's degree in Civil Engineering from Tongji University,

Hao Zheng is a PhD Candidate at the University of Pennsylvania, Stuart Weitzman School of Design, specializing in machine learning, digital fabrication, mixed reality, and generative design. He holds a Master of Architecture from the University of California, Berkeley, and Bachelor of Architecture and Arts degrees from Shanghai Jiao Tong University.