A POI-Based Machine Learning Method for Predicting Residents' Health Status

Shicong Cao[1] and Hao Zheng[2](✉)

[1] School of Architecture, Clemson University, Clemson, USA
[2] Stuart Weitzman School of Design, University of Pennsylvania, Philadelphia, USA
zhhao@design.upenn.edu

Abstract. Health environment is a key factor in public health. Since people's health depends largely on their lifestyle, the built environment which supports a healthy living style is becoming more important. With the right urban planning decisions, it's possible to encourage healthier living and save healthcare expenditures for the society. However, there is not yet a quantitative relationship established between urban planning decisions and the health status of the residents. With the abundance of data and computing resources, this research aims to explore this relationship with a machine learning method. The data source is from both the OpenStreetMap and American Center for Decease Control and Prevention (CDC). By modeling the Point of Interest data and the geographic distribution of health-related outcome, the research explores the key factors in urban planning that could influence the health status of the residents quantitatively. It informs how to create a built environment that supports health and opens up possibilities for other data-driven methods in this field.

Keywords: Health environment · Point of interest · Local data · Health-related outcome · Machine learning

1 Introduction

1.1 Health Environment

Healthy living is a goal of many 21st century cities. Since 1999, the WHO Healthy city project has brought up the principles of urban planning supporting health and example cities whose development can be learned from (Duhl and Sanchez 1999) The approach is mainly narrative and case study, which doesn't provide enough ground and guidance for decision making in a quantitative way.

Previous studies suggest three domains where urban planning can most effectively focus support for health and well-being – physical activity, community interaction and healthy eating, since these domains address some of the major risk factors for chronic diseases (Kent and Thompson 2014). A literature review methodology is used in this study.

With the ever-rapid changing of the world, it is difficult to understand how those principles work and how much they will actually influence the health status of the

© The Author(s) 2022
P. F. Yuan et al. (Eds.): CDRF 2021, *Proceedings of the 2021 DigitalFUTURES*, pp. 139–147, 2022.
https://doi.org/10.1007/978-981-16-5983-6_13

residents. Since building a healthier living environment is cooperation across the society, it would be beneficial for stakeholders to share a common ground. Data can be the common ground.

With the availability of health data and the abundance of computing resources, it is now possible to quantitively evaluate a planning decision's outcome on residents' health. The research showcases the possibility that using open-source city point of interest data to predict the health status of the residents using machine learning methods.

1.2 Problem Statement

In order to estimate the obesity rate, one of the machine learning model, Convolutional Neural Network (CNN), has been used to analyze the satellite image (Newton et al. 2020). Analysis of the convolutional layers gives a suggestion of which visual features are more important for a low obesity rate. The limitation of the imagery method is the amount of computing it requires and the obscurity of the conclusion due to the limitation of dataset as well as 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). The advantage of this approach is that it considers human perception of the built environment. However, the amount of data processing and redundancy is a problem of this method.

The use of 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 the possibility that POI data can be a good indicator of urban conditions and activities. 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).

1.3 Objectives

This research aims to use machine learning to analyze the relation between POI data and residents' health status. By looking into the pattern behind the data, the objective is to testify the existing healthy city planning principles as well as discovering new relations between built environment and health. Compared with imagery methods, using POI data from OpenStreetMap provides a more quantifiable result and requires less computing resources. Also, the varieties of features of OSM makes it possible to search for the most important factors among many aspects of built environment.

2 Methodology

The workflow of this research follows five steps. In the first step, the number of the POI data from OpenStreetMap for the California state was collected and some initial data exploration was conducted. Second, the health-related outcomes data was collected and spatially joined with the census tract boundary and the POI data count. Third, a principal component analysis (PCA) analysis was conducted, and the features that best capture the variance were selected. Fourth, the selected features were used to train the supervised machine learning model. Finally, the models were used to predict the health-related outcomes in the test set. The results were validated and mapped accordingly.

2.1 Data Source

The data source constitutes of three data sets, the POI data from OpenStreetMap, the local health data from CDC, and the place boundary file from the Census Bureau. The test region is within the state of California, United States as a compromise between data availability, handling capacity, and statistical accuracy.

2.1.1 Local Health Data from CDC

Local Data for Better Health is a project that reports county-, place-, census tract-, and ZCTA-level data and uses small area estimation methods to obtain 27 chronic disease measures for the entire United States. 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 27 measures include 5 unhealthy behaviors, 13 health outcomes, and 9 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 (*Places: Local Data for Better Health*, no date). For the research, the specific dataset used has 18 health outcomes available in place level. The population size of each place is also included as a column in the dataset.

2.1.2 POI Data from OpenStreetMap

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). The research uses Overpass API to query the database by the tags to get the geographic location of certain features. To begin with, 54 features among all the primary features were queried. The selection was based on the number of data points available and the relation with the physical activity, community interaction, and healthy eating of the residents as mentioned in the literature review. The 54 features can be categorized into food, healthcare, transportation, community service, leisure, tourism, building, and nature. The data points of each feature were spatially joint with the place boundary data and the count was calculated.

2.1.3 TIGER/Line Shapefiles Place Boundary Data

Since the local health data provided by CDC can be spatially joined with the TIGER/Line mapping system, TIGER/Line Shapefile is also used to count the number of POIs in each place. Each shape generates one row of data with the number of POIs and local health outcomes. There are in total 1468 rows and so is the sample size. The area of each place is calculated from the dataset and added as a column. Figure 1 is a map of the 1468 places in California. A place is defined by the United States Census Bureau as a concentration of population that has a name, and it typically has a residential nucleus and a closely spaced street pattern, and it frequently includes commercial property and other urban land uses (*Census Bureau Definition*, no date). It is a geographic level data covering most of the population with relatively condensed areas. It can be seen that there is a concentration of places in the metropolitan area, near Los Angeles, Bay Area and Sacramento.

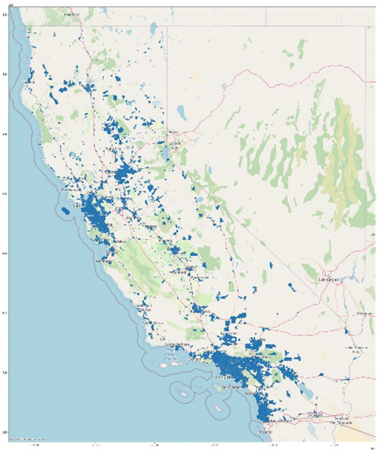


Fig. 1. Sample dataset of places in California

2.2 Data Analysis

After joining the three datasets mentioned above together, we get a dataset with 1468 rows and 74 columns. The first 54 columns are the POI data counts, then the population and area of the place, and the last 18 columns are health-related outcomes.

In the first step of the data exploration, the population and area of the places are plotted to get an initial idea of the sample selected. As in Fig. 2, most of the places are comparable in size and population, with a few outliers which are possibly denser areas within the metropolitan. The one sample with a large population and area is Los Angeles and the second largest in area and population is San Diego.

The test plot of POI data as in Fig. 3 shows some correlation with the area and population. The test plot of a test plot of the health outcome shows that most of the sample is within a certain range with no explicit pattern.

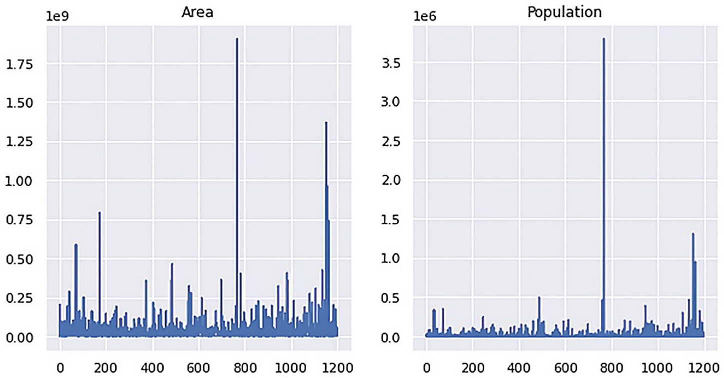


Fig. 2. Area and population of the sample places

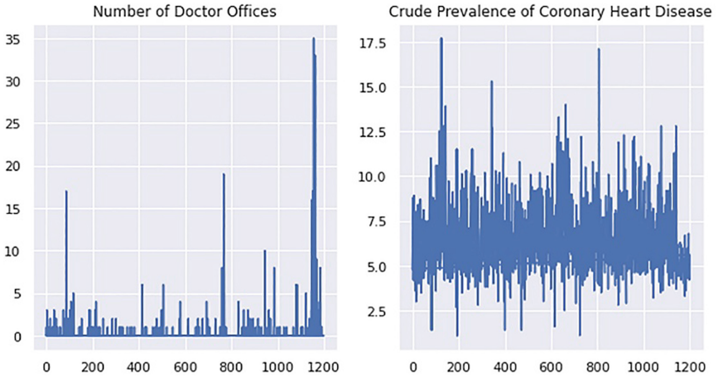


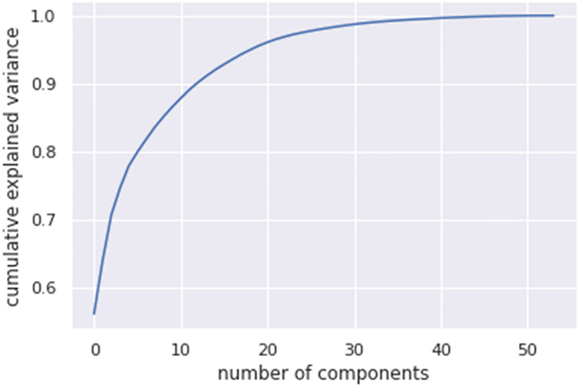
Fig. 3. Number of doctor offices and the crude prevalence of coronary heart disease

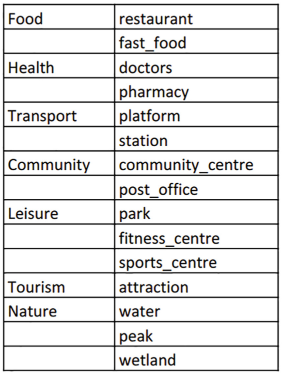
2.3 Principal Component Analysis (PCA)

The next step is to conduct a principal component analysis (PCA) with the POI data to decide which features best captures the variance. This is achieved with the sklearn package of Python. As the cumulative explained variance chart (Fig. 4 left) shows, 10 components can explain about 88% of the variance, and 20 components explain about 96% of the variance. After calculating the first principal component (PC1), 15 features with the highest scores were selected. Figure 4 right shows the features selected among the original 54 features. Note that the category of building has no feature selected.

2.4 Machine Learning

With the 15 POI features, the area and the population of the place as the input data, several machine learning models were trained with randomized training dataset to predict the 18 health outcomes. The median accuracy were calculated for each model with the test dataset, then an average accuracy rate is calculated for the 18 predictions.

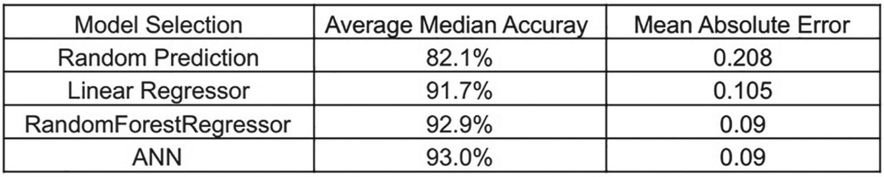




Food	restaurant
	fast_food
Health	doctors
	pharmacy
Transport	platform
	station
Community	community_centre
	post_office
Leisure	park
	fitness_centre
	sports_centre
Tourism	attraction
Nature	water
	peak
	wetland

Fig. 4. Left: number of clinics and the crude prevalence of coronary heart disease. Right: features selected with highest first principal components

We implemented four machine learning models (Fig. 5). A Random Prediction model is implemented with the DummyClassifier of sklearn package. Random values within the test data range are generated. Then a Linear Regression is conducted as a basic statistic prediction. It's implemented with the LinearRegression function of sklearn. Random Forest Regression is preferred since it's a machine learning algorithm based on decision tree and relatively fast to train. It's implemented with the RandomForestRegression function of sklearn. Artificial Neural Network (ANN) is a deep learning method that digitally mimicry the human brain to predict values. A 5-layer neural network is used in the research and a training step of 10000 achieves the best accuracy. The model is implemented with Tensorflow.



Model Selection	Average Median Accuray	Mean Absolute Error
Random Prediction	82.1%	0.208
Linear Regressor	91.7%	0.105
RandomForestRegressor	92.9%	0.09
ANN	93.0%	0.09

Fig. 5. Prediction accuracy for different machine learning models

3 Results

After comparing the Average Median Accuray and Mean Absolute Error, Random Forest and ANN have similar accuracy in the test set prediction while the Random Forest is much faster to implement and calculate. A feature importance analysis is conducted with the sklearn built-in function. As shown in Fig. 6, population and area are the two most important features. For the POI data, the most important features are water, park, platform, and restaurant.

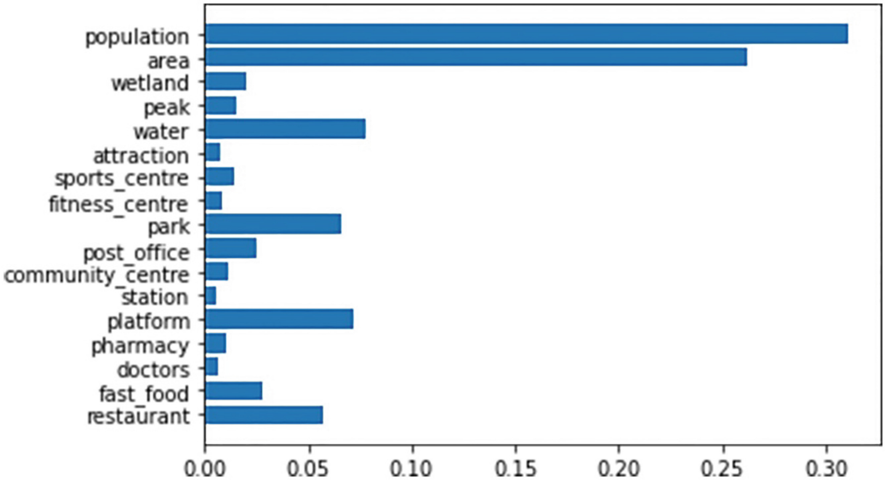


Fig. 6. Feature importance of random forest model

The prediction result on the test set is mapped as in Fig. 7. The Mean Error is calculated as the average of the 18 health outcomes with the prediction minus ground truth. The overall range of the Mean Error is between −0.18 to 0.18. Since the test data is already normalized within 0 and 1, the Mean Error could represent the average accuracy of the prediction by percentage.

There is a tendency to underestimate the health outcomes for larger and denser places for example Los Angeles and several places on the southeast. For medium and smaller-sized towns and rural areas, the prediction is closer to the ground truth. Note that the prediction accuracy with the Random Forest model is much better than Linear Regression on outliers in Metropolitan areas, the latter has a Mean Error of –1.4 in Los Angeles.

4 Discussion

4.1 Prediction and Theory Testimony

This research uses POI data from OpenStreetMap to predict the residents' health status. From the hundreds of features on the OSM, the research selected 54 features that have the higher density and fall into a category that could relate to health city planning. The PCA method reduces the features to 15 items, which could best represent the variance in the data.

With the permutation feature analysis on the test data, except for the area and population of the places, water, park, platform, and restaurant are the most important POI features. This corresponds to three domains: physical activity, community interaction, and healthy eating that could best support health and wellbeing. (Kent and Thompson 2014) Water and park usually correlate with public space for activity, platforms relate to transport accessibility and the abundance of restaurants (in this case formal eating

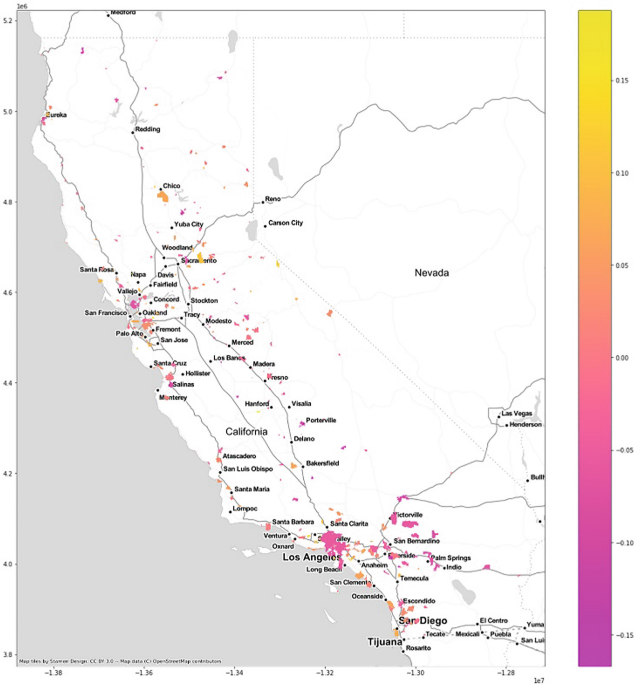


Fig. 7. Mean error for 18 health outcomes with random forest model

places) could represent the food availability of a certain area. The data-driven research supports previous qualitative research and testifies existing healthy city planning theory.

For model selection, compared with Linear Regression and Random Prediction, the Random Forest model and Artificial Neural Network model have better accuracy, while the Random Forest is faster to train and use. Also, for outliers, Random Forest and ANN work better than Linear Regression because of the non-linearity.

4.2 Challenges and Future Research

Since the research proves that it's possible to predict the health outcomes with POI data from OpenStreetMap, the next step could be to enlarge the test area and see if this methodology can be used for other areas.

There is a difference between metropolitan areas and suburban areas in terms of prediction accuracy. For places of a metropolitan for example Los Angeles, there is tendency of underestimating the health outcomes while for suburban areas there are many places overestimated. Further research could be conducted to see if the deciding factors are different for different urban contexts.

Another challenge is that the health outcome data is still from a model-generated source, although according to CDC, the model is based on survey and local data. If more direct health data is available in the future, probably more linkage between the built environment and health could be realized and testified.

5 Conclusion

The research explores a method to predict the health outcomes of the residents by using the POI data available from OSM, thus exploring the linkage between the built environment and the health of the residents. Different machine learning methods were used and evaluated. The result shows that the Random Forest model has the best balance point between prediction accuracy and ease of implementation.

With the abundance of data and computing resources, the research proposes a way of using data to support urban planning ideas. In order to improve the health status of society, decisions have to be made and data can be the common ground. This approach shows a vast potential in the future, that data could assist decision making towards a healthier built environment.

References

Census Bureau Definition. https://www.census.gov/programs-surveys/geography/about/glossary.html#par_textimage_14

Cichosz, P.: Urban crime risk prediction using point of interest data. ISPRS Int.J. Geo-Inf. 9(7) (2020). https://doi.org/10.3390/ijgi9070459

Feldmeyer, D., et al.: Using OpenStreetMap data and machine learning to generate socio-economic indicators. ISPRS Int. J. Geo Inf. **9**(9), 1–16 (2020). https://doi.org/10.3390/ijgi9090498

Kent, J.L., Thompson, S.: The three domains of urban planning for health and well-being. J. Plan. Lit. **29**(3), 239–256 (2014). https://doi.org/10.1177/0885412214520712

Newton, D., et al.: Deep Learning Methods for Urban Analysis and Health Estimation of Obesity (2020)

Places: Local Data for Better Health. https://www.cdc.gov/places/methodology/index.html

Duhl, L.J., Sanchez, A.K.: Healthy cities and the city planning process. WHO **62**(4), 91–103 (1999)

Tang, Z., et al.: A data-informed analytical approach to human-scale greenway planning: integrating multi-sourced urban data with machine learning algorithms. Urban For. Urban Green. 56(Jan) (2020). https://doi.org/10.1016/j.ufug.2020.126871

Zhou, H., et al.: Social inequalities in neighborhood visual walkability: Using street view imagery and deep learning technologies to facilitate healthy city planning. Sustain. Cities Soc. **50**(129), 101605 (2019). https://doi.org/10.1016/j.scs.2019.101605

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.