Exploring Life Quality Index via Education Level and Population Density

Wenqi Che[1, †] , Meiyi Du[2 †] , Qiteng Feng[2,*, †]

[1]Department of economics, University of California, San Diego, La Jolla, United State

[2]Department of economics and Department of statistics, University of Toronto, Ontario, Canada

*Corresponding author: qiteng.feng@mail.utoronto.ca

[†]These authors contributed equally.

Abstract. The subjective nature of quality of life, makes it complex to measure the quality of life. Among them, finding a few key criteria that affect the quality of life in cities will be the main purpose. In this study, a linear regression model was used to predict the numerical relationship between population density, bachelor's degree rate and education Gini index, and further Akaike information criterion (AIC) and Bayesian information criterion (BIC) were selected to validate and adjust the model, and finally F-test was used to compare the initial model with the adjusted model. In addition, the random forest algorithm was used to classify the quality of life of each city by population density as well as education level to identify the common characteristics of cities with high physical quality of life index. Three classification decision trees created using different combinations of population density, bachelor's degree rate and education Gini index were used, followed by additional multiple decision trees generated by the random forest algorithm. The final results of the receiver operator characteristic curve (ROC) curves and confusion matrix of the two sets of decision trees were observed and evaluated to stratify the cities. The experimental results show that population density has a negative effect on quality of life, while bachelor's degree rate has a positive effect. Although the three indicators of population density, bachelor's degree rate and education Gini index cannot effectively distinguish the four levels, they can effectively distinguish which cities are not the first level. it shows that cities that achieve high education level and low population density are not necessarily the cities with the highest quality of life, but those that do not are definitely not.

Keywords: Life quality evaluation, random forest, machine learning.

1. Introduction

Quality of life index is a measure of well-being of a society, which is defined as a person's or a community's standard of happiness, comfort, and health. Millions of people in the United States relocate to another part of the country every year in search of a better quality of life. However, life quality is a subjective notion that each individual defines differently. In the past decades, substantive progress has been made in learning how measures of various aspects of life quality interrelate [1]. Employment, income level, housing, GDP, inflation, and other characteristics that can be evaluated in numbers are utilized in the study area to assess a person's or a society's life quality. Cities have a wide range of living qualities, and few key criteria may have a big impact on the measurement of quality index.

In the early field of life quality measurement, there were number of factors used as an evaluation of quality-of-life index. They included purchasing power, homicide rate, fulfillment of basic physical needs, education rate [2], population density [3] and so on. As for the education rate, the previous studies provided evidence that education have a positive effect on human's well-being. Previous research showed that high education input was in hopes of having greater wealth and higher consumption power. Yet it has a negative relationship with life satisfaction [4].

On the other variable was choose, population has more mixed type of research outcomes. Cramer et al. [5] and Başkan et al. [6] were the results about how does population density effect people's life. Cramer et al. [5] demonstrated that having a high population density will lead to a high chance have mental illness, which can reduce the quality of life. Whereas Başkan et al. [6] indicated that people who live in an area with high population density were highly scored in four aspects: physiological,

psychosocial, society and environment. There was a confliction of results between the two researches. One of them applied the data from developed country -- Australia, and the other applied the data from developing country -- Turkey. Therefore, this paper hopes to further study how population density affects life quality in several aspects based on the data from America. However, the study was only centered on the education effect on income and mental health, but ignored some other impact on physical level, such as environment, income level, etc. The reason for this is that people think that the most important factor in a high quality of life is happiness, but they forget that the basis for a high quality of life is the material aspect.

In this regard, further research is needed in finding the common characteristics of cities with a high quality of life index and providing some advice to government to improve living quality of cities. A linear regression model that contains several characteristics was built to find the significant level of these factors on the quality-of-life index. Moreover, this study is going to explore more on the side effects that the significant factors like education level and population density may have on the aspects such as income, environment, mental health, and lead to a result of high life quality.

2. Datasets and Methods

2.1 Data sources and selection

To quantify the life standard for each city, a quality-of-life index form NUMBEO website was introduced into this paper [7], which comprehensively includes physical influential factors, such as purchasing power index X_1, property price to income ratio X_2, cost of living index X_3, safety index X_4, health care index X_5, traffic commute time index X_6, pollution index X_7, and climate index X_8. The formular can be found in equation (1), which indicates purchasing power, safety, health care and environment are positive related with quality-of-life index and property price to income ratio, cost of living, traffic commute time and pollution are the reverse.

$$Quality\ of\ life\ Index = 100 + \frac{x_1}{2.5} - x_2 - \frac{x_3}{10} + \frac{x_4}{2.0} + \frac{x_5}{2.5} - \frac{x_6}{2.0} - \frac{2.0*x_7}{3.0} + \frac{x_8}{3.0} \quad (1)$$

Purchasing power index and cost of living index are ratio referring to New York city as base, which considering the money purchasing power and life expenses containing groceries, transportation, and utilities. Property price to income ratio presents the level of average income, house price and mortgage. Other five money unrelated index, safety, health care, traffic commute time, pollution, and environment, are constructed based on surveys result from NUMBEO website visitors.

Population density is an objective statistic and collected from World Population Review website with unit people counted per mile square [8]. Unlike population density, education level doesn't have official index to measure. Thus, this article used bachelor's degree rate for over 25 years old people and education Gini index from City-Data website to instead [9]. The bachelor's degree rate can state general adult citizens' education reception, the major class related with quantify the life studying object, and education Gini index quantifies the inequality in education.

2.2 Set cities quality-of-life categories

In city screening process, a restriction of minimum population (over 100,000 resident) is established because the analysis target are cities. Small towns or villages with less population more inclined to be outliers due to tiny sample size more susceptible to extreme observation. Missing index values form NUMBEO website cannot replace with simple default values since it will affect the outcome of the models. The data set omitted all cities with blank terms and remained 117 observations divided into four categories, level one to level four, according to the four quantiles of the quality-of-life index, with level one representing the highest class and level four in converse.

2.3 Methods

In this study, the association between education, population density and physical quality of life was analyzed via two main methods: linear regression model and random forest algorithm. The linear

regression model primarily focuses on the numerical relationship between response variable and predicting variable population density, bachelor's degree rate and education Gini index. To validate models, the four assumptions of linear regression model, linearity, homoscedasticity, independence, and normality are diagnosed, and Akaike information criterion (AIC) and Bayesian information criterion (BIC) was implemented in stepwise selection for choosing variable process. AIC is an estimator of prediction error and reflect relative quality of statistical models for a given set of data. Bayesian information criterion BIC is a criterion for model selection which is related to AIC. And R-squared is used to measure the representation of proportion of the variance for a dependent variable that's explained by predictors in the model. And the model with smaller AIC, BIC and larger R squared is better. Finally, F-test was applied to compare the initial model and the adjusted model.

In addition to the linear relationships between responded variable and predicting variables was examined, this study also attempted to classify cities based on life quality through education level and population density. Firstly, three classification decision trees were created based on the different combination of population density, bachelor's degree rate and education Gini index. Then, using random forest algorithm [10] to generate multiple decision trees, and evaluating three decision trees and random forest based on their results of receiver operating characteristic (ROC) curves and confusion matrixes.

3. Results and discussions

3.1 The linear regression model

In the research, an initial linear regression model is built to predict the quality-of-life index and test to obtain final model. The initial linear regression model consists of all features. There are assumptions that our variables are random, independent and identically distributed, and linearity relationship existed. Firstly, model 1 contains three parameters: population density, bachelor's degree rate and education Gini index.

Table.1. Comparison of P-value of predictors (Model 1).

The name of predictor	P-value
Intercept	< 2 *10-16
Population Density	1.88 * 10-5
Bachelor Degree Rate	4.16 8 10-10
Education Gini Index	0.282

As the Table 1 shown, the p-value of population density and bachelor degree rate are less than 0.05, which indicates that there is strong evidence that population density and bachelor degree rate are significant in predicting living quality index. The p-value of education Gini index is 0.282, which represent that it does not significantly affect living quality index. So, model 2 is built containing only population density and bachelor degree rate.

To further determine which model is the more suitable, AIC and BIC are applied in model comparison. The Table 2 below shows the AIC and BIC of both model 1 and model 2. It is obvious to see that both AIC and BIC of the final model is smaller keeping the adjusted R squared same. Therefore, the model is better without the education Gini index parameter.

Table.2. Linear model with all features.

Model	AIC	BIC	R.squared
Model 1	766.09	778.80	0.48
Model 2	765.30	775.48	0.48

Last but not least, assumptions are checked in Figure 1.

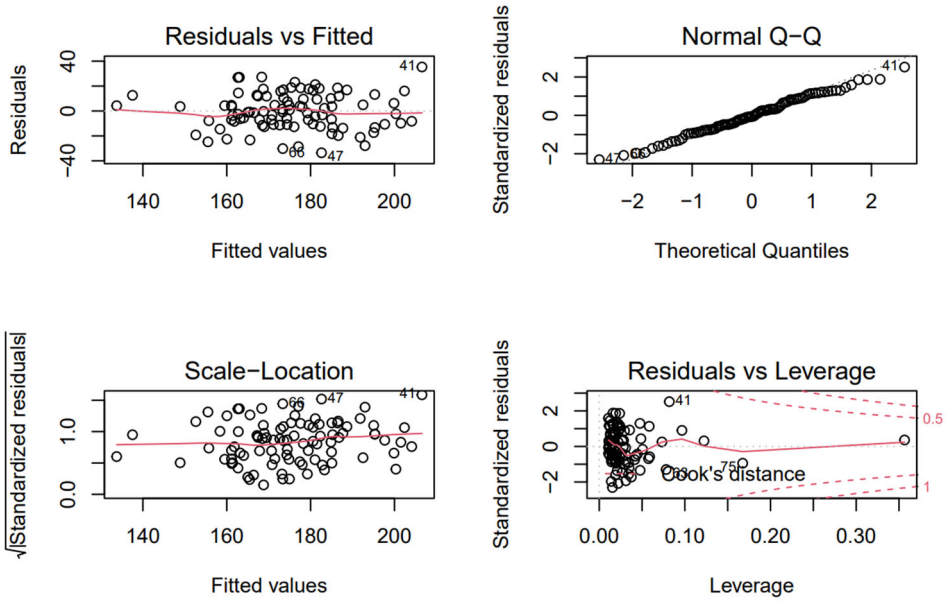

Figure 1. Regression model assumptions.

The Residual vs Fitted Figure observes equally spread residuals around the horizontal line without distinct pattern. There is no evidence show our assumption violation for linearity. As for the normal Q-Q graphic, the residuals follow a horizontal line. There is no evidence show violation of assumption for normality. Then, it finds that there is a straight line with randomly spread points in Scale-location graphic. So there is no evidence show assumption violation for homoscedasticity. Any of assumption violation does not occur. Model 2 is used as the final model for estimate quality of life index.

Table.3. The Estimate Coefficient and P-value of predictors

The name of predictor	Estimate Coefficient	P-value
Intercept	149.4	< 2 * 10-16
Population density	-2.089 * 10-3	5.76 * 10-8
Bachelor degree rate	1.079	8.84 * 10-13

The final model shows when the population density and bachelor degree rate is zero, the quality of life index is 149.4. However, it does not have real meaning since the population density of cities cannot be zero. Moreover, the population density has a negative impact on the living quality while the bachelor degree rate has a positive impact.

3.2 Random forest for city classification

To further determine the common characteristics of cities with high physical life quality index, this study applied a random forest algorithm to classify the cities via population density, bachelor's degree rate, and education Gini index. However, since the primary purpose was to discuss the difference between high physical life quality cities and others in the above three proposed indicators, the overfitting problem of random forests is not essential in this study. In fact, the high accuracy of the model indicates there are standard features for high physical life quality cities on these three indicators.

The initial model simulated 500 trees with 1 variable tried at each split classified, and the result did not meet speculative expectations, that the error rate for each level all greater than 0.58 showed in confusion matrix (Table 4). Thus, population density, bachelor's degree rate, and education Gini index cannot effectively distinguish cities alone. However, after combining other three levels cities into one category and remaining only level 1, the accuracy of model is improved due to dramatically decline in false positive rate, percentage of cities in others category perceived as level 1 cities. The confusion matrix of after combined model (Table 5) stated that, although the three indicators are invalid to distinguish the specific level for each city, they can sufficiently identify which city is not with high physically quality of life according to the specificity in Table 5, ability to correctly classify cities in

others category. Hence, it can be inferred that the failure of any of three indicators meeting the standard imply this city's physically quality of life index doesn't reach level 1.

Table.4. Confusion matrix for Initial model.

	Level 1	Level 2	Level 3	Level 4	Class.error
Level 1	11	12	6	1	0.6333333
Level 2	5	11	5	11	0.6562500
Level 3	3	6	9	6	0.6250000
Level 4	0	10	7	14	0.5483871

Table.5. Confusion matrix for after combined model.

	Level 1	others	Class.error
Level 1	11	19	0.6333333
Others	8	79	0.0919540

Noteworthily, unlike in regression model with no significant linear relationship, education Gini index is important in decision tree models based on the result in Mean decrease Accuracy and Mean Decrease Gini (Figure 2). This result suggested that differences in educational resources are not linearly related to urban quality of life but a city with huge gap in education equality probably is not with a high physical quality of life.

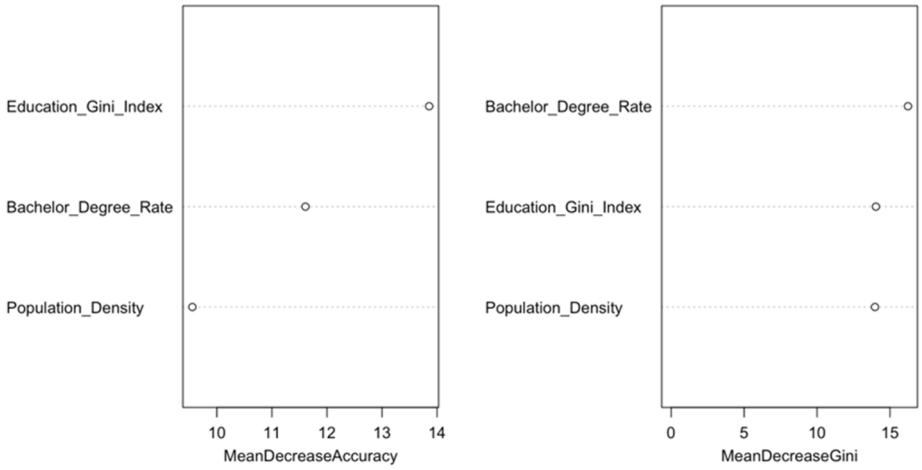

Figure 2. Mean Decrease Accuracy and Mean Decrease Gini

4. Conclusions

In this study, investigating the impact of two elements on high material quality of life: population density and educational attainment (bachelor's degree rate and educational Gini index) was proposed. This paper applied a linear regression model and a random forest algorithm to determine the common characteristics of cities with high material quality of life index. Through model comparison and random forest classification, the experimental results show that population density has a negative impact on quality of life, while the bachelor's degree rate has a positive impact. And the results show that the actual quality of life in a city with a large gap in educational equality may not be high. The limitations of this study are that this study chose two objective coefficients to represent education levels and the variables studied are not comprehensive enough to fully represent the absolute impact on quality of life. In the future, further study plans to obtain more data from outside the United States to comprehensively study the factors affecting quality of life worldwide.

References

[1] Abbey, Antonia, and Frank M. Andrews. Modeling the Psychological Determinants of Life Quality [J]. Social Indicators Research, vol. 16, no. 1, D. Reidel Publishing Company, 1985, pp. 1–34, https://doi.org/10.1007/BF00317657.

[2] Diener, E. A value based index for measuring national quality of life [J]. Soc Indic Res 36, 107–127 (1995). https://doi.org/10.1007/BF01079721

[3] Douglas Carnahan, Walter Gove, Omer R. Galle, Urbanization, Population Density, and Overcrowding: Trends in the Quality of Life in Urban America [J], Social Forces, Volume 53, Issue 1, September 1974, Pages 62–72, https://doi.org/10.1093/sf/53.1.62

[4] Powdthavee, Nattavudh, et al. What's the Good of Education on Our Overall Quality of Life? A Simultaneous Equation Model of Education and Life Satisfaction for Australia [J]. Journal of Behavioral and Experimental Economics, vol. 54, Elsevier Inc, 2015, pp. 10–21, https://doi.org/10.1016/j.socec.2014.11.002.

[5] Cramer, Victoria, et al. Quality of Life in a City: The Effect of Population Density [J]. Social Indicators Research, vol. 69, no. 1, Kluwer Academic Publishers, 2004, pp. 103–16, https://doi.org/10.1023/B:SOCI.0000032663.59079.0b.

[6] Başkan, Aytekin Hamdi, et al. Impact of the Population Density on Quality of Life. International [J], Journal of Human Sciences, vol. 14, no. 1, J-Human Sciences, 2017, pp. 506–18, https://doi.org/10.14687/jhs.v14i1.4416.

[7] Quality of Life. Quality of Life Index [R], www.numbeo.com/quality-of-life/rankings_current.jsp.

[8] World Population Review. The 200 Largest Cities in the United States by Population 2022 [R], https://worldpopulationreview.com. Accessed 20 April 2022.

[9] City-data. Advanced U.S. City Search, www.city-data.com//advanced/search.php.

[10] Biau, G., & Scornet, E. A random forest guided tour [J]. Test, 25(2), 197-227, 2016.