STEM Fellowship Big Data Challenge 2023

The Influence of a House's Proximity to Amenities on Educational Attainment and Demographics in the U.S.

Adam Omarali[1], Haoyang (Leo) Liu[1], Yiyi Xu[1], Navid Farkhondehpay[1], and Zicheng (James) Tian[1]

[1]Crescent School

February 22, 2023

Abstract

As worldwide urbanization continues with more cities being constructed, identifying important factors for a high-functioning city is becoming increasingly relevant. A house's proximity to amenities has been extensively studied in relation to housing prices, but not usually with other success factors due to granular data limitations. We present a novel method to collect and analyze granular, localized proximity data vis-à-vis educational, demographic, and socioeconomic metrics. To investigate the relationships, a multi-stage simple random sample of houses ($n = 1630$) was selected For each house, and amenity data was compiled via the Geoapify API to be used to calculate a proximity score. We concluded that proximity to amenities has a positive correlation with bachelor's degree attainment ($r = 0.6$), and less affordable areas have more access to amenities ($r = 0.41$). These findings could inspire changes in equitable housing and city planning to fairly distribute amenities and ensure adequate educational opportunity.

Keywords
Proximity Data, City Design, Urban Planning, Education, Proximity to Amenities

1 Introduction

Cities are designed to make everyday life simple while aiding human development. UN Sustainability Goal 11 focuses on designing safe, resilient, sustainable, and inclusive cities [1]. Determining if cities are designed to promote educational development is an important aspect of designing a successful city [2, 3]. This paper focuses on an individual house's access to amenities as an indicator of city design. Many papers have explored amenity scores for a variety of amenity types [4], and others have already examined the relationship between amenity proximity and housing prices, but research is sparse for other indicators [5, 6]. Proximity's need for granular data at a dissemination block level to produce accurate conclusions [7] may explain the limited amount of research available on this subject. To grasp a balance of necessities and wants for a city, we will explore transit, healthcare, parks, financial services, schools, grocery stores, restaurants and fitness as amenity types.

Successful education is necessary for a well-functioning society, as indicated by UN Sustainability Goal 3 [8]. Educational attainment has increased over time in the US, but can differ according to race and age [9]. While education and demographics have been explored extensively, this paper aims to explore each in the context of amenity proximity to offer a more holistic analysis of the necessary attributes of a well-functioning society. Exploring city designs that boast high educational attainment rates at various levels (high school, bachelor) can pose insight into where high-functioning individuals migrate. Demographic data must also be explored to ensure an equitable distribution of the educational and economic benefits that arise from amenity proximity. This paper aims to identify potential issues in demographic equity through analysis of home value, affordability index, and diversity index variables, which have historically been good indicators of demographic data.

2 Materials & Methods

2.1 Programming Libraries

In the process of conducting this research, many different technologies were used. Python 3.9 was used for data manipulation and analyses together with data science packages Pandas, NumPy, Matplotlib, Seaborn, SciPy, and Folium. Requests were used to make Rest API calls. Git version 2.30.1 was used as version control, and Jupyter 4.8.1 was used as the environment for our programs.

2.2 Sampling Methods

The main sampling method we used in our study to form a list of addresses was multi-stage random sampling. We first used simple random sampling to sample 163 counties (a geographic area including rural areas and urban cities) out of the 1936 available counties that OpenAddresses [10] provided. Afterwards, we used simple random sampling again within each selected county to sample 10 addresses, forming a total of 1630 addresses.

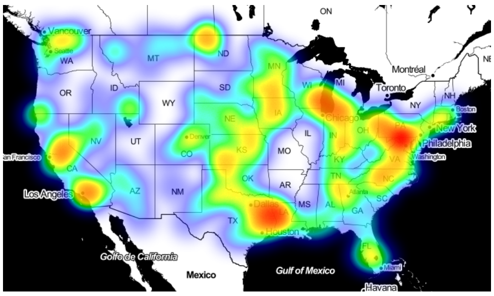

Figure 1: This map displays the places that were selected as samples. A warmer colour indicates more samples taken in the area.

2.3 Data Collection

For this paper, we did not use any existing datasets to query amenity proximity as they did not suit our needs for granular, quantitative, and representative data. To achieve granular data at the household level, map-based application programming interfaces (APIs) were used.

With 1630 randomly selected addresses as origins, we used the Geoapify Places API [11] to query the 15 nearest amenities for each of the 8 categories of amenities: healthcare, park, financial, school, transit, grocery, fitness, and restaurant. Using the results of this query, we calculated the amenity proximity from the origin to these types of amenities. In this calculation, we first evaluated the weighted distance to destinations. Since an individual is more likely to visit a closer destination rather than a further one, each subsequent destination beyond the first is weighted down by 0.5. This function mimics exponential decay.

$$\text{dist} = \frac{\text{result}_1 + 0.5 \cdot \text{result}_2 + 0.25 \cdot \text{result}_3 + \cdots}{1 + 0.5 + 0.25 + \cdots}$$

If there are fewer than 15 destinations found, less than 15 results are used. Then, to calculate the proximity, the weighted distance and the number of destinations found are taken into consideration. A coefficient of 1000 is applied to scale ratings to a normal range.

$$\text{prox} = 1000 \cdot \frac{\#\text{ destinations found}}{\text{dist}}$$

With every one of the 1630 addresses, we searched for amenities nearby and generated 8 proximity scores for 8 types of amenities as well as an overall amenity variable as the sum of the individual amenity types. The distribution of the proximity score takes a right-skewed shape, where a higher proximity score indicates higher amenity accessibility. To verify that the proximity function is a valid metric, we grouped houses into rural (county population within 1 mile of a house < 2500) and urban, and plotted the proximity distribution. From Figure 2, we see effectively that rural houses have significantly lower proximity scores than urban ones.

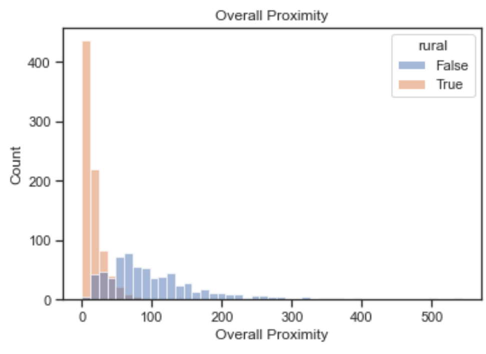

Figure 2: This graphic shows the distribution of proximity coloured by rural and urban houses.

In addition to proximity data, our research also compiled educational and demographic data. These data are acquired from the ArcGIS/Esri API [12]. Using its API, we queried educational and demographic data in a 1-mile radius around sampled houses. For fields that may be confounded by population, we divided those fields by the population in that area. These are highlighted by the "Percentage" suffix in our data fields.

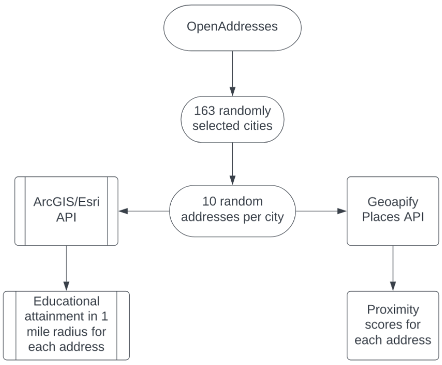

Figure 3: This graphic shows the sampling methodology along with the data collection pipeline. Every address' longitude and latitude coordinates were taken as inputs for the Geoapify Places API and ArcGIS/Esri API.

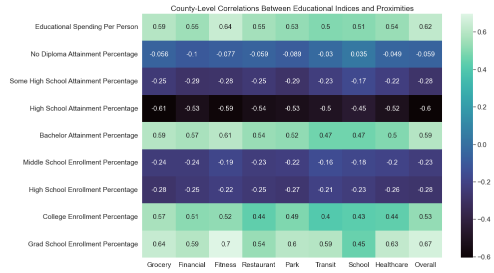

Figure 5: This heatmap displays the correlation values for a variety of city mean educational attainments (as a percentage of the population), and educational enrollments (as a percentage of the population aged 25 and under), alongside individual mean proximity scores for each amenity type.

3 Results

3.1 Educational Development

Educational Development explores the effect of proximity on degree attainment and educational spending. Both variables take population levels into account, recognizing them as confounding variables. A city's mean level of educational attainment is represented as a percentage of the 1-mile radius population to a house while educational spending is calculated per person.

3.2 Demographic Features

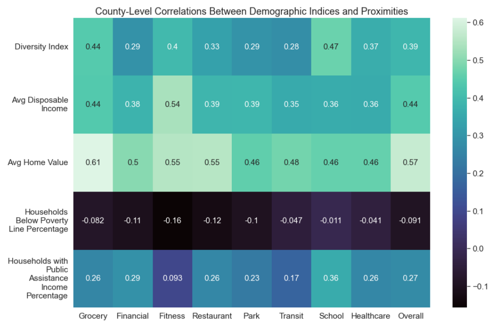

Figure 6: This heatmap displays correlation values for a variety of demographic features related to poverty, housing affordability, and diversity compared to individual mean proximity scores for each amenity type.

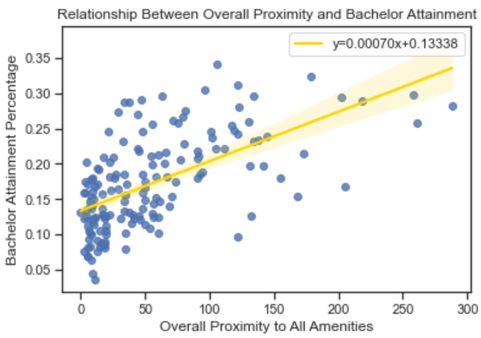

Figure 4: This scatter plot explores the relationship between a county's overall proximity to amenities and the percent of the population aged 25+ who have a bachelor's degree. Each data point represents 1 county. $r = 0.600$

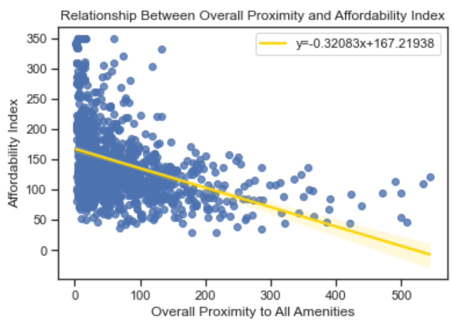

Figure 7: This scatter plot displays the relationship between a house's affordability index value and a house's overall proximity value. Housing affordability has a differing correlation value when compared as a mean at the county level compared to each individual house. $r = -0.410$

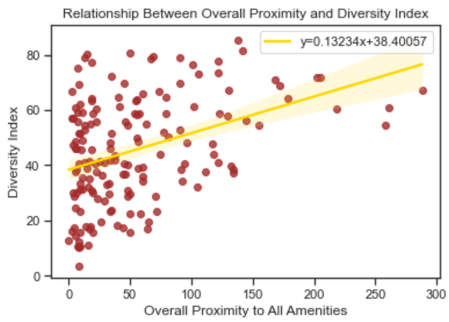

Figure 8: This scatter plot displays the relationship between a county's mean diversity index and the mean overall proximity. $r = 0.386$

3.3 Proximity Analysis

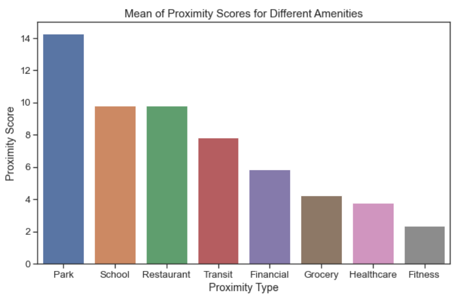

Figure 9: This bar chart shows the proximity scores calculated by the proximity function for each amenity type.

4 Discussion

Through the data that was collected, there seems to be a moderately strong trend in that higher amenity proximity leads to more "desirable" living places. A higher overall proximity is related to more academically successful citizens and a stronger diversity index.

4.1 Educational Development

Figure 2 shows a positive, moderately strong Pearson coefficient of correlation ($r = 0.6$) indicating a relationship between bachelor's degree attainment and overall proximity. The p-value for the linear relationship is $3.294 \cdot 10^{-17}$. Since the p-value is below the significance level ($\alpha = 0.05$), we have convincing evidence that the linear relationship is significantly different that 0. Further connections are demonstrated in Figure 3. For example, it seems that an address's proximity to fitness centers has the greatest correlation to educational spending per person as well as bachelor's degree attainment percentage. Fitness is a luxury amenity and therefore found close to fewer houses as indicated by Figure 9. Surprisingly, transit and school amenities have a lower correlation value to educational development factors. In contrast with other education variables, it's interesting that high school diploma attainment percentage is negatively correlated with all proximity scores. Perhaps high schools that are close to many amenities cause students to be distracted or have limited physical space which causes schools to have a lower diploma attainment rate. More research into this result is needed. Educational spending per person as well as Bachelor's degree attainment rate both are positively correlated to all of the proximity scores. Areas of high accessibility to amenities seem to place more importance on education and are more successful at obtaining a bachelor's educational success level.

4.2 Demographic Features

Figure 4 shows the Pearson coefficient of correlation value (r) of each of the specific amenity proximities (as well as the sum) compared to different demographic indices. Amenity proximity positively impacts average home values the most, especially proximity to grocery stores. Amenity proximity has little correlation with the percentage of households below the poverty line, indicating that core amenities have been fairly distributed independent of the poverty level. Figure 5 further reinforces the conclusion drawn from Figure 4, which is that proximity to amenities causes houses to be priced higher, therefore being less affordable. The p-value for the linear relationship is $2.018 \cdot 10^{-50}$. Since the p-value is below the significance level ($\alpha = 0.05$), we have convincing evidence that the linear relationship is significantly different than 0. Proximity to amenities is positively correlated to a county's diversity index. In other words, it seems that the more diverse counties are also the more amenity-rich ones.

4.3 Limitations

In the process of conducting the research, there were limiting factors that hindered the gathering of data and analysis of results. A major roadblock we encountered was computational limitations: each request of one address for one type of amenity requires around 10 seconds, and the Geoapify API only allows around 3000 requests per day. Therefore, due to both time and usage constraints, we could not sample any more addresses than we currently have. The impact of this limitation is crucial to our research since the county-level data are obtained by taking the mean of 10 sampled houses in that county. This mean can easily be pulled as the number of samples is low and the mean is not a statistic that is resistant to extreme outliers, which are present in our data. In future work, more addresses must be sampled from each county in order to stabilize the mean and accurately represent it.

4.4 Future Work

In addition to increasing sample sizes and taking more samples, this data can be used to form predictive models and conduct feature analysis. Identifying which amenities have the largest effect on education and demographic factors will be beneficial. Creating an interface where people can design cities and receive proximity scores for their design will make our findings readily usable. From designs, predictive analysis can be used to estimate educational and demographic statistics. Finally, the beauty of our novel map-based data pipeline is that data can be collected across the world so long as secondary indicators can be found. This analysis will allow for claims to be made around worldwide specific city development by analyzing how proximity design varies by city.

5 Conclusions

The conclusions made in this paper do not imply anything in regard to cause and effect. According to the data analyzed in this research, a house's accessibility to facilities affects demographics and educational attainment in the U.S. A home's resident's levels of education and diversity demographic are positively correlated to the proximity of the home to services like grocery stores, fitness centers, and schools. Specifically, amenity proximity positively impacts the bachelor's degree attainment percentage in an area with a correlation coefficient of 0.600, as well as increasing educational spending per person with a correlation coefficient of 0.590. Conflicting results were identified in the relationship between high school diploma attainment and proximity. In terms of demographics, a house that is more accessible to amenities generally comes with a higher price tag (correlation coefficient of 0.410). This emphasizes how crucial it is to take amenities into account while planning and building housing complexes, especially if the developer wants to maximize the value of the home. Although affordability decreases as proximity increases, houses below the poverty line maintain reasonable access to amenities. The precise methods through which accessibility to amenities affects demographics and educational attainment can benefit from further study.

6 Acknowledgements

We appreciate the opportunity provided by STEM Fellowship for us to engage in this research journey and all the resources they provided. During our research, we consulted mentors several times to align our approach to the project and received very valuable guidance. We wholeheartedly highlight that this work would not have been possible without their help. In addition, we also acknowledge and thank Mr. Michael Stephens for the feedback and advice he provided. The project is made possible through the Crescent School chapter of STEM Fellowship that he oversees.

References

[1] Goal 11 | department of economic and social affairs.

[2] Joeke Kuyvenhoven and Willem R Boterman. Neighbourhood and school effects on educational inequalities in the transition from primary to secondary education in amsterdam. *Urban Studies*, 58(13):2660–2682, 2021.

[3] Timo M. Kauppinen. Neighborhood effects in a european city: Secondary education of young people in helsinki. *Social Science Research*, 36(1):421–444, 2007.

[4] Yan Song, Jiang Zhou, Yingjie Zhang, Dingxin Wu, and Honghai Xu. How much are amenities worth? an empirical study on urban land and housing price differentials across chinese cities. *Land*, 11(6), 2022.

[5] Christopher Hewitt and W. E. Hewitt. The effect of proximity to urban rail on housing prices in ottawa. *Journal of Public Transportation*, 15:43–65, 12 2012.

[6] Liv Osland, John Östh, and Viggo Nordvik. House price valuation of environmental amenities: An application of gis-derived data. *Regional Science Policy and Practice*, 2020.

[7] Alessandro Alasia, Nick Newstead, Joseph Kuchar, and Marian Radulescu. Measuring proximity to services and amenities: An experimental set of indicators for neighbourhoods and localities, Feb 2021.

[8] Goal 4 | department of economic and social affairs.

[9] Ryan c, bauman k. educational attainment in the united states: 2015 population characteristics.

[10] Openaddresses - the free and open global address collection, 2022.

[11] Places api - search places and amenities by criteria, Nov 2022.

[12] Arcgis online.