Department of Statistics 2022-23

# A Bayesian approach to valuing walkable neighborhoods in Washington DC

Jacob Eliason


ST451 - Bayesian Machine Learning

# Table of contents

# 1 Introduction

## 1.1 Background

In early 2023, the "15-minute city" concept, coined by French urbanist Carlos Morena (Moreno [2016]), became the subject of protests in the the United States, the United Kingdom, and Canada when it was misrepresented by internet conspiracy theorists as a secret ploy for a new authoritarian restriction on freedom of movement. The idea was originally promoted to create urban environments where people can access essential needs and services within a 15-minute walk or bike ride from their homes. It has recently gained traction in cities like Paris and Melbourne (Pozoukidou and Chatziyiannaki [2021]), as governments re-envision urban living post-pandemic and address climate change by reducing transportation emissions.

The misguided backlash against the concept of the 15-minute city highlights the need to develop and clearly communicate the evidence for the value of proposed urban development. Research that demonstrates the benefits of sustainable urban living can help policymakers better make the case for forward-thinking policies to their constituents. To that end, the body of evidence supporting the benefits of walkable neighborhoods has grown substantially in recent years. Walkability can be quantified in a number of different ways, such as calculating the density of destinations, the diversity of land uses, and the design of the street network. Walkable neighborhoods have been linked to various health, economic, and environmental benefits, including increased physical activity, reduced air pollution, and higher property values.

However, despite the substantial body of evidence supporting the benefits of walkable neighborhoods, there remains a need for a more nuanced understanding of the relationship between various walkability-related factors and the value of residential properties. The diverse range of amenities that contribute to a neighborhood's walkability can vary significantly in their impact on property values, depending on factors such as local context, population demographics, and the specific amenity types involved. Furthermore, the methodologies employed in previous research often rely on traditional regression models, which may not fully capture the complex, non-linear relationships between walkability factors and property values, nor adequately account for unobserved heterogeneity within the data.

## 1.2 Research Question

This paper seeks to contribute to the existing literature on walkable neighborhoods by employing a fully Bayesian approach to evaluate the relationships between proximity to specific categories of amenities and residential property listing prices. By using Bayesian methods, which allow for the incorporation of prior knowledge and the estimation of uncertainty in model parameters, this paper aims to provide a more robust and accurate assessment of the impact of walkability-related factors on property values. Furthermore, this paper aims to serve as an illustration of the benefits of using Bayesian methods in urban planning research. As shown here, the flexibility, interpretability, and the incorporation of uncertainty can make for better-informed decision-making and policy development.

# 2 Data

## 2.1 Zillow

The primary dataset used in this study consists of American and Canadian residential properties listed for sale on Zillow during the first quarter of 2022. Zillow is a popular online platform that provides extensive information on real estate listings, making it a valuable source of data for this analysis. The data was scraped and uploaded to Kaggle in the second half of 2022 (Data

Ranch [2022]). The full dataset contains a total of 127,014 observations, covering a wide range of property types, sizes, and locations. This dataset is useful for analysis because the listings come from a narrow window in time, which allows us to assume that the economic and policy conditions are relatively constant across the sample. However, many of the variables that influence property prices also vary across geography and usually require significant effort to account for. I therefore choose to trade generalizability for simplicity by focusing on a single city: Washington, D.C. Focusing on a single city allows us to better isolate the effects of walkability-related factors on property prices, as the properties under study will be subject to similar local regulations, zoning policies, and urban planning initiatives. Moreover, concentrating on one city reduces (but does not eliminate) the impact of differences in regional economic conditions, such as employment opportunities, income levels, and population growth, which can also affect property values. After filtering the full dataset to include only properties located within the city's boundaries, the dataset contains 791 observations. Along with latitude and longitude coordinates for the listing, the dataset includes variables for a small number of property characteristics such as price, square footage, number of bedrooms, and number of bathrooms, which are important determinants of property values.

## 2.2 Open Data DC

In order to measure the walkability of each property, I use data from the Open Data DC portal (Office of the Chief Technology Officer [2023]). Open Data DC is a public platform hosted by the Washington D.C. city government that provides access to a wide range of datasets related to the District. I select city-curated datasets containing the locations of amenities within the boundaries of the district that are commonly cited as contributing to walkability: Metro stations, grocery stores, pharmacies, schools (by combining public, private, and charter schools up to the high school level), and post offices. The files are provided in GeoJSON format, which allows for easy integration with the Zillow dataset.

## 2.3 Feature Engineering

To measure the proximity of each residential property to various amenities, this study tests several feature engineering techniques. Three types of proximity variables are calculated for each amenity category: distance (in meters) to the closest instance of the amenity, count of unique instances of the amenity within a half-mile radius, and a binary variable indicating the existence of at least one instance of the amenity within a half-mile radius.

These proximity variables are calculated using the `geopandas` package in Python, with the underlying assumption that the shortest path between two points is a straight line. This simplification may not always hold in practice due to the presence of obstacles and the layout of the street network.

Finally, I perform several adjustments to the dataset before proceeding with analysis. I remove property listings that show square footage less than 50 feet or which are listed for less than $10,000, as inspection of the data reveals that these are likely to be erroneous entries. I also remove listings that are missing data for any of the variables used in the analysis. This leaves a final dataset of 729 observations.

## 2.4 Exploratory data analysis

To explore the data, I show summary statistics for all variables below. Note that price is given in thousands of U.S. dollars here. All other variables are shown untransformed.

Table 1: Summary statistics for Washington DC residential property listings

| | Mean | SD | Min | Q1 | Median | Q3 | Max |
|---|---|---|---|---|---|---|---|
| price | 1122.5 | 1848.1 | 60.0 | 399.9 | 570.0 | 949.0 | 20000.0 |
| beds | 2.4 | 1.6 | 0.0 | 1.0 | 2.0 | 3.0 | 12.0 |
| baths | 2.5 | 1.9 | 0.0 | 1.0 | 2.0 | 3.0 | 17.0 |
| area | 1698.0 | 1836.8 | 316.0 | 716.0 | 1100.0 | 1870.0 | 17631.0 |
| nearest_grocery | 563.1 | 447.6 | 9.9 | 251.1 | 441.8 | 782.9 | 3134.4 |
| nearest_pharmacy | 416.5 | 278.4 | 11.1 | 210.4 | 354.5 | 571.0 | 1858.6 |
| nearest_schools | 338.7 | 175.2 | 35.8 | 203.4 | 315.8 | 435.0 | 1302.3 |
| nearest_post | 719.9 | 481.4 | 43.6 | 394.2 | 582.1 | 947.6 | 3281.3 |
| nearest_metro | 929.1 | 693.6 | 15.6 | 431.9 | 729.5 | 1247.1 | 3801.1 |
| count_nearby_grocery | 2.2 | 1.9 | 0.0 | 1.0 | 2.0 | 4.0 | 8.0 |
| count_nearby_pharmacy | 4.2 | 3.3 | 0.0 | 2.0 | 3.0 | 6.0 | 15.0 |
| count_nearby_schools | 5.2 | 3.3 | 0.0 | 3.0 | 5.0 | 7.0 | 17.0 |
| count_nearby_post | 1.1 | 1.1 | 0.0 | 0.0 | 1.0 | 2.0 | 7.0 |
| count_nearby_metro | 0.9 | 1.0 | 0.0 | 0.0 | 1.0 | 1.0 | 6.0 |
| nearby_grocery | 0.8 | 0.4 | 0.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| nearby_pharmacy | 0.9 | 0.3 | 0.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| nearby_schools | 1.0 | 0.1 | 0.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| nearby_post | 0.7 | 0.5 | 0.0 | 0.0 | 1.0 | 1.0 | 1.0 |
| nearby_metro | 0.6 | 0.5 | 0.0 | 0.0 | 1.0 | 1.0 | 1.0 |

As shown below in Figure 1, there is moderate spatial correlation in property prices across the city. The most expensive properties are largely located in the northwest quadrant of the city, while the least expensive properties are largely located in the southeast quadrant.

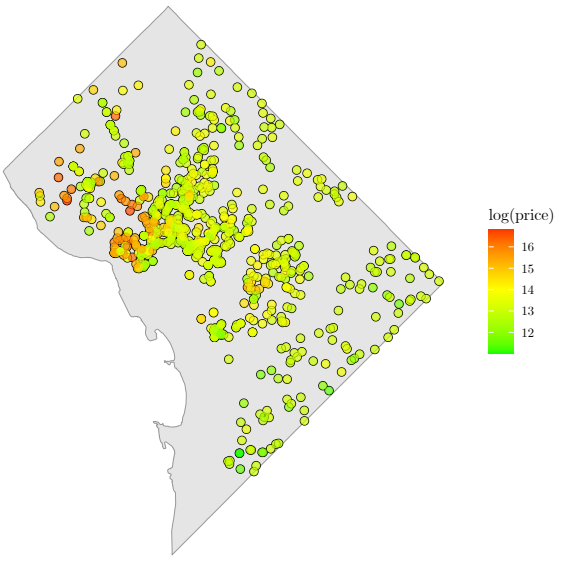


Figure 1: Residential property listing prices by location in Washington DC

# 3 Methods

For the following analyses, I standardize all data (except for binary variables) by subtracting the mean and dividing by the standard deviation. I split the standardized data into a training set and a test set for model estimation and evaluation respectively. I always use the log of price and the log of area when relevant as they each have strong right skew. In order to compare the performance of different types of models, I measure performance using the mean squared error (MSE) on the test set. The MSE is a commonly used loss function for regression problems, and penalizes large deviations from the true values, encouraging accurate predictions.

## 3.1 Bayesian Regression

Bayesian regression is an approach to linear modeling that combines prior knowledge with observed data to obtain a posterior distribution for the parameters of interest—in contrast with frequentist regression, which estimates a single set of parameters using maximum likelihood estimation. These posterior distributions can be used to make probabilistic statements about the parameters, such as the probability that a parameter is greater than zero.

This portion of the analysis first seeks to evaluate which of the three measurement types for proximity to amenities is most useful. This is assessed by comparing the test set MSE of models with distance variables to those with binary or count variables. A lower test loss indicates a better fit and more accurate predictions. After selecting a set of proximity variables, I'll review the coefficients of the chosen model and consider those a "naive" answer to the research question. This will assume linear relationships between price and distance variables and independence between observations, which may prove to be reasonable assumptions. This will nonetheless serve as a baseline for comparison with more sophisticated models presented later on.

In the implementation, I define four different Bayesian regression models, each with a unique set of proximity variables: Model A (base model without proximity variables), Model B (distance variables), Model C (count variables), and Model D (binary variables).

For all models, I set the priors for the beds, baths, and area variables as follows:

- beds: normal distribution with a mean of 0 and standard deviation of 1
- baths: normal distribution with a mean of 0 and standard deviation of 1
- area: normal distribution with a mean of 1 and standard deviation of 3 (to account for its expected larger effect)

For all proximity variables in Models B, C, and D, I set the normal distribution with a mean of 0 and a standard deviation of 1 as the prior.

These priors are reasonable, non-informative, and reflect the standardized nature of the respective variables. The modeling process I use then combines these assertions with the observed data to obtain the posterior distributions.

The analysis is conducted using the `brms` package in R. Bayesian regression models fit using `brms` use the Stan programming language to perform Markov chain Monte Carlo (MCMC) sampling using the No-U-Turn Sampler (NUTS) algorithm.

## 3.2 Dirichlet Process Mixture Model

A Dirichlet Process Mixture Model (DPMM) is a nonparametric clustering approach that can automatically infer the number of clusters in the data. Unlike traditional clustering algorithms, such as k-means, DPMM does not require the user to specify the number of clusters beforehand, "automatically [making] the trade-off between model complexity and fitting the data" (Bishop

[2006]). This is advantageous when the true number of underlying groups is unknown, as is the case in this context. Using a DPMM, I aim to uncover latent subgroups within the data that may correspond to distinct neighborhood characteristics, which in turn may impact the relationship between walkability and property prices. As the relationships between walkability-related factors and property values may not be linear or homogeneous across the entire dataset. By dividing the data into subgroups, I'll explore and model these relationships separately for each cluster, which may reveal complex or non-linear patterns that were not evident in the overall data.

I specify a DPMM using the `PyMC` library in Python. I set parameters intended to produce a small number of clusters as follows: I set a truncation level of 12 for the Dirichlet Process, which determines the maximum number of clusters the model can infer. The concentration parameter is set to 0.75 (where higher values would favor more clusters). For the base distribution priors, I assume a Gaussian distribution for the means with a mean of 0 and standard deviation of 10. The covariance matrix is modeled using a LKJ Cholesky covariance distribution with an eta parameter value of 2. To perform inference, I use the No-U-Turn Sampler (NUTS) MCMC algorithm with a target acceptance rate of 0.9. I draw 4,000 samples after an initial tuning phase of 1,000 samples, running the algorithm on 6 separate chains simultaneously.

### 3.3 Bayesian Regression Within Neighborhoods

I next build upon the previous methods by incorporating the information contained in the subgroups identified using the DPMM. I do this by fitting linear models of the same form as before within each cluster. By modeling the relationship between walkability-related factors and property values within each subgroup separately, I hope to capture the unique characteristics and dynamics of each neighborhood. This approach allows me to account for the heterogeneity in the data and avoid the issues of omitted variable bias that arise when the influence of unobserved location factors is mistakenly attributed to the distance variables.

I fit separate Bayesian regression models for each subgroup identified by the DPMM using the same priors and formula as in Model B above (which used the distance variables identified as having the best predictive power on the test set).

### 3.4 Gaussian Process Regression

To address some of the remaining limitations of the previous methods and account for non-linear relationships and spatial correlation in the data, I employ Gaussian Process Regression (GPR) as an alternative modeling approach. GPR is a non-parametric regression technique that can capture the complex dynamics between walkability-related factors and property values.

The GPR model is implemented using the `GPy` library in Python. As before, I standardize all the data by subtracting the mean and dividing by the standard deviation, also using the same train and test split for consistency.

I define a kernel function used in the Gaussian process that is a combination of the Matérn kernel, a white noise kernel, and a radial basis function (RBF) kernel. The Matérn kernel gives the distance between two points as a function of their Euclidean distance, which allows me to capture the spatial correlation in the data. The RBF kernel is used to capture the non-linear relationships between the walkability-related factors and property values, and the white noise kernel is used to account for the noise in the data.

To optimize the hyperparameters of the Gaussian process, I perform a grid search using functions from `sklearn` with a custom scoring function based on the MSE. The grid search explores various combinations of noise variance and length scale parameters. The best hyperparameters found in the grid search—0.5 for the length-scale parameter in the Matérn kernel and 0.25 for the variance parameter in the white noise kernel—are then used to fit the final GPR model. The Matérn kernel uses a smoothness parameter of 5/2. I specify the use of Automatic Relevance

Determination for the RBF kernel, which will allow me to evaluate the relative importance of each walkability-related factor in the model.

Finally, I compare several alternatively-specified GPR models using a kernel containing only the RBF and white noise components to compare how much of the variance in the data is explained by the spatial correlation captured by the Matérn kernel. I compare various combinations of variables using this kernel.

# 4 Results

In this section, I present and compare the results obtained from the methods described above.

## Bayesian Regression

I find that the model containing proximity variables defined in terms of distance in meters from the nearest amenity instance (B) has lower MSE on the test set than the models containing alternative definitions of proximity (C, D), or the model containing only the base variables (A). I show the MSE values in the table below.

Table 2: Predictive performance of Bayesian regression models

| Model | MSE |
|---|---|
| B | 0.202 |
| C | 0.238 |
| D | 0.246 |
| A | 0.306 |

I proceed with model B, which uses square footage, number of bedrooms, number of bathrooms, and distance from nearest metro station, grocery store, post office, school, and pharmacy as independent variables. I show the posterior distributions for its model coefficients below. The posterior median is given as a point estimate in pink.

Since I've standardized my explanatory variables, the units can be understood as in the following example: "a one standard deviation increase in distance from the nearest metro station corresponds with an average decrease of 0.04 standard deviations in the log-transformed price of a residential property for sale in Washington DC, with 95% credible interval (-0.09, 0.00)."

I'll note however that the posterior distributions for the coefficients of the distance variables are quite wide, indicating that the model is not very certain about their values. Additionally, I have not accounted for spatial correlation in the data at this point, which could mean that the estimates are biased.

## Dirichlet Process Mixture Model

The results of my DPMM analysis are shown below. As stated above, the number of clusters for DPMMs is not specified in advance, but rather is inferred from the data. I find some difficulty in the sampling process for the DPMM and the results are only moderately stable. However, the clusters are quite reasonable given my prior knowledge of the distribution of properties in D.C. and the plot of prices across the city shown earlier.

I name the clusters W, X, Y, and Z by descending order of median price. I show median values by cluster for all variables in the table below.

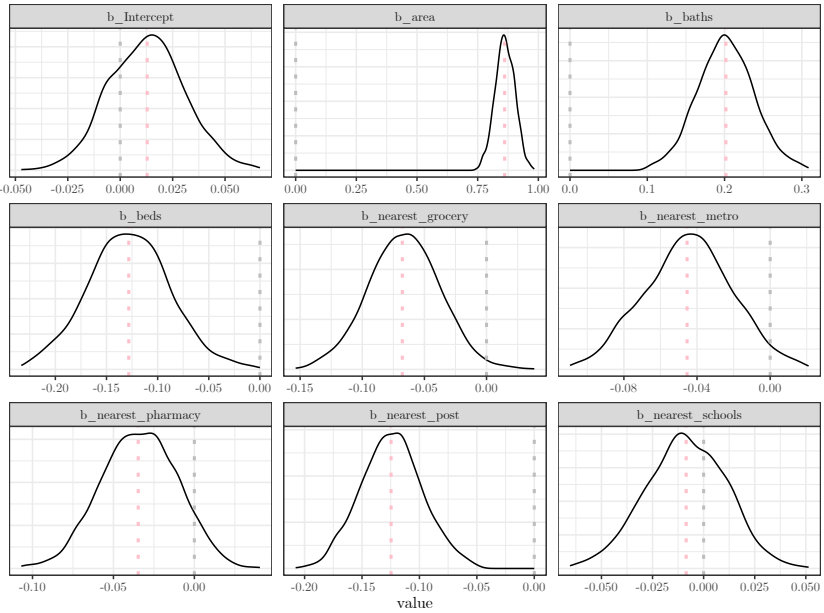


Figure 2: Posterior distributions for model coefficients

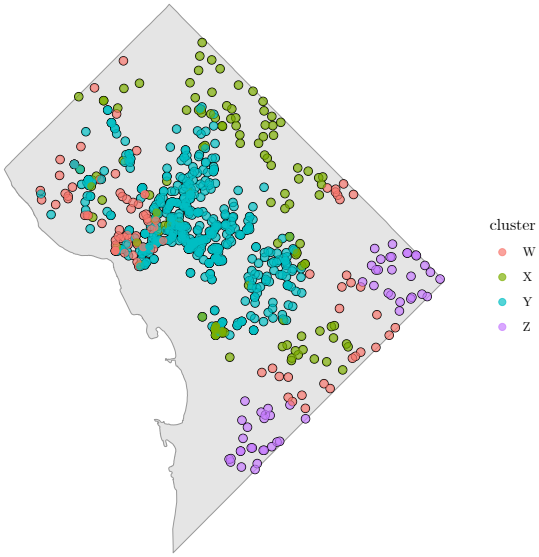


Figure 3: Dirichlet Process Mixture Model clusters

Table 3: Median values for selected variables by cluster

| name | W | X | Y | Z |
|---|---|---|---|---|
| price | 3822450 | 547000 | 575000 | 439500 |
| area | 3826 | 1170 | 949 | 1305 |
| beds | 4 | 3 | 2 | 3 |
| baths | 4 | 2 | 2 | 2 |
| nearest_metro | 1274 | 1093 | 559 | 1251 |
| nearest_schools | 393 | 367 | 303 | 282 |
| nearest_grocery | 748 | 501 | 342 | 1215 |
| nearest_pharmacy | 554 | 433 | 298 | 501 |
| nearest_post | 699 | 718 | 529 | 1324 |

## Bayesian Regression within DPMM

Having identified four clusters of properties distinguished by location and price, I fit separate Bayesian regression models for each cluster, using the same formula specified in Model B previously. I show the posterior distributions for the coefficients of the models in the plot below. The variation in sample size between clusters is evident in the width of the posterior distributions.

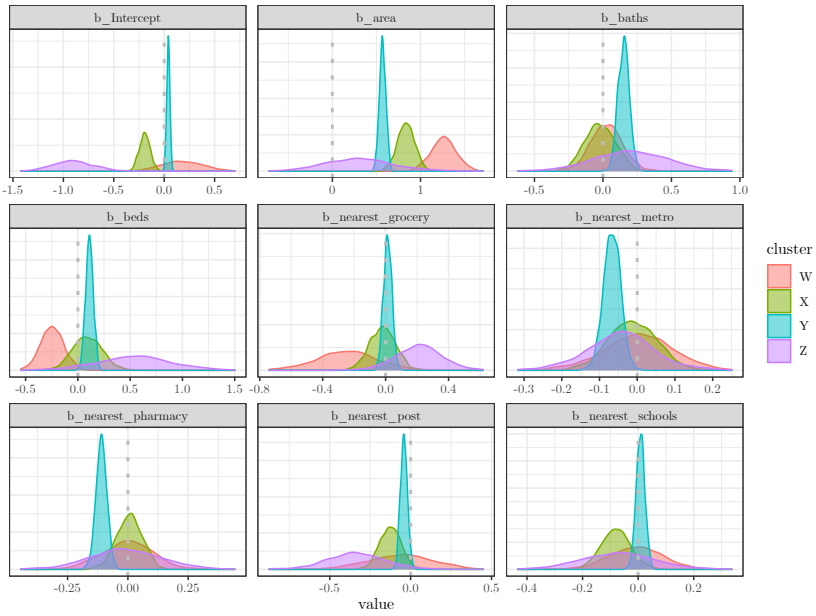


Figure 4: Bayesian regression coefficients from models fit by cluster

However, I do find differences in effects from proximity to amenities by cluster. For example, distance from the nearest metro station appears to have a large effect on the price of properties only in cluster Y, while distance from the nearest school appears to have a large effect on the price of properties only in cluster X.

## Gaussian Process Regression

I fit a GPR model using the same formula specified in Model B previously, but now fully accounting for spatial correlation in the data by my use of the Matérn kernel. I find a MSE of 0.09 on the test set, which is meaningfully lower than the MSE of 0.20 obtained from the Bayesian

regression model. I also fit additional GPR models, which are described along with their test set MSE in the table below.

Table 4: Predictive performance of GPR models

| Variables | Kernel | Test MSE |
|---|---|---|
| lat, lon, beds, baths, area, distance_* | Matérn + White Noise + RBF | 0.09 |
| beds, baths, area, distance_* | White noise + RBF | 0.18 |
| beds, baths, area | White noise + RBF | 0.30 |
| distance_* | White noise + RBF | 0.65 |

By printing standardized length scale values for each feature in a GPR model, "relative importance of different inputs [can] be inferred from the data" (Bishop [2006]). For the model incorporating the Matérn kernel, the relative importance of the distance variables is said to be trivially small. For all other applicable models, distance to the nearest metro station is assigned the highest relative importance of all distance variables.

# 5 Discussion

## 5.1 Summary

The results show that, of the three different engineered measures of proximity to amenities, distance in meters is the most predictive of property price. This likely reflects the fact that the distance in meters contains more information than the other measures. Additionally, it's possible that most of the value in living near an amenity is captured by proximity to at least one instance of that amenity (i.e. that the marginal value of living near a second metro station is less than the marginal value of living near the first metro station).

I also show that the effect of distance to amenities on property price can be partially represented by a linear model, as I find that the models including some measure of proximity to amenities outperform the baseline model. However, the large majority of variability in property price is simply explained by the property's square footage.

Using DPMM, I produce reasonable subgroups within the property listings in the city that represent actually distinct groups of properties. I then show that the effect of distance to amenities on property price varies by subgroup. This is consistent with the idea that the value of living near different amenities depends on the location of the property. For example, the distance to the nearest metro station appears to have a meaningful effect on properties in the city center, but not necessarily in the richer exterior.

Finally, I show that a GPR model that fully accounts for spatial correlation in the data far outperforms the baseline Bayesian regression model and the regression models fit within clusters. This shows that the location of a property contains more information than just the distance to the nearest amenity. Additionally, the GPR model that used only the same variables as those contained in the Bayesian regression model outperformed the baseline model, suggesting there is non-linearity in the relationship between property price and the distance to amenities.

## 5.2 Implications

I demonstrate the importance of considering both spatial and non-linear relationships when analyzing the impact of walkability-related factors on property values. While distance to amenities does matter, the relationship is not always linear or easily interpretable, and the importance of proximity to certain amenities can vary across different neighborhoods.

While relationships between proximity to amenities varies in nature and in space, this paper shows evidence in several places that proximity to transportation (distance from nearest metro station) appears to have a practically significant relationship with property price. This suggests that transportation infrastructure may be a particularly important factor to consider when designing walkable cities.

### 5.3 Limitations

This paper has several limitations that should be considered when interpreting the results. First, the analysis is based on data from property listings in Washington, D.C., which may not be representative of the broader population of residential properties or the preferences of residents in other cities. Additionally, the dataset only includes a limited number of amenity types, which may not capture the full range of factors that influence walkability and property values. Furthermore, the use of linear distance rather than network distance (i.e. distance traveled via roads or paths) may not accurately represent the actual accessibility of amenities for residents. Additionally, the amenity data used is only available within the city boundary, which may bias the results against properties near the border.

Furthermore, the study relies on listed property prices rather than actual sale prices, which may not fully reflect the true value of properties. The properties for sale in the dataset may also not be a random sample of all homes, which could potentially bias the results. Finally, the study does not account for other important factors that may influence property values, such as the age of homes or the quality of local schools.

## 6 Conclusion

Given the limitations of the current study, there are several avenues for future research in this area. First, researchers could explore the relationship between walkability-related factors and property values using data from other cities or regions, in order to assess the generalizability of the findings. Additionally, future studies could incorporate a wider range of amenity types and other relevant factors, such as network distance or the quality of local schools, to provide a more comprehensive assessment of the impact of walkability on property values.

In any case, relevant stakeholders should use empirical methods that account for non-linear relationships and spatial correlation when assessing the impact of walkability-related factors on property values. This will help ensure that the results are accurate and representative of the true relationship between these factors and property values.

# 7 References


Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, New York, NY, 1 edition, August 2006.

Data Ranch. Kaggle Dataset: 100,000+ Zillow Properties. Website, August 2022. URL https://www.kaggle.com/datasets/dataranch/100000-zillow-properties-us-canada.

Carlos Moreno. La ville du quart d'heure : Pour un nouveau chrono-urbanisme, Oct 2016. URL https://www.latribune.fr/regions/smart-cities/la-tribune-de-carlos-moreno/la-ville-du-quart-d-heure-pour-un-nouveau-chrono-urbanisme-604358.html.

Office of the Chief Technology Officer. Open Data DC Data Collections. Website, April 2023. URL https://opendata.dc.gov/search?collection=Dataset.

Georgia Pozoukidou and Zoi Chatziyiannaki. 15-minute city: Decomposing the new urban planning eutopia. *Sustainability*, 13(2), 2021. ISSN 2071-1050. doi: 10.3390/su13020928. URL https://www.mdpi.com/2071-1050/13/2/928.