Deriving and Verifying a Community Vitality Index through a Machine Learning Model Based on House Price Data

By Mark Li

Author Biography

Mark Li, a 16-year-old Junior at Avon High School, CT, excels in mathematics, computer science, and physics. Mark's passion extends to data science and social science research, where he combines machine learning techniques with social sciences to explore complex issues. Beyond academics, he enjoys absorbing new knowledge and finds fulfillment in guiding his peers. In the future, Mark aspires to tackle societal well-being or dilemmas by incorporating AI design as a generalized calculator for medicine, making a meaningful impact on the world's challenges.

Abstract

Recognizing the complexity and limitations of the approach to produce a traditional Multi-dimensional Community Index (MD-CVI) as well as high cost and time-intensive, the study explores another approach that establishes a machine-learning model to derive a community index by leveraging contemporary house prices. This method has several significant advantages over traditional methods, providing a more efficient and reliable way to assess community well-being. The model produces a dynamic Community Vitality Index (CVI) that quickly captures changes within a community and enhances the MD-CVI's sophistication by selecting and combining any geographic regions generated by clustering algorithms, employing machine learning model and house sold price reduces human errors from questionnaires, and potentially improves accuracy. The study also exemplified and verified the model via house data from Minnesota, Texas, and Connecticut, spanning the last five years. Subsequently, comparing the house-price-based CVI with traditional MD-CVIs reveals a consistent alignment in their community index rankings. The conclusion is that the house-price-based CVI, as second research, is a viable proxy and a benchmark for a traditional MD-CVI, indicating an innovative approach for the stakeholders, including the government, education institutions, health and welfare organizations, etc. in assessing community well-being and making informed policy decisions regarding the flow of public funding of community-based projects. This research has the potential to significantly impact policy decisions, ensuring that public funding is directed to the communities that need it most.

Keywords: Real-Time Community Vitality Index, Multi-Dimensional Community Index, House-Dependent Factors, Location-Dependent Factors, House Price, Census, and Machine Learning

Introduction

A Community Vitality Index (CVI) provides a comprehensive overview of community well-being. (Curtis et al., 2012). Governments, non-profit organizations, researchers, and even private companies have used various community indexes to assess and measure a community's vitality and allocate funding accordingly. A traditional multi-dimensional CVI (MD-CVI) typically involves collecting data from multiple aspects of a community, such as demographics, employment, and Census. Traditional CVIs rely on collecting surveys and Census information updated once every decade, and CVIs are adjusted annually based on extrapolation instead of actual community changes. The process is complex, resulting in significant delays in CVI updating, thereby limiting the MD-CVI's ability to capture urgent matters in community dynamics. To address the discreteness of a traditional MD-CVI, this study introduces a machine-learning model to derive a community index from house prices, which reflects the combined influence of various factors affecting the real estate market and the community's overall well-being simultaneously.

Two categories of factors influence house prices: those directly linked to the physical house structure, referred to as "house-dependent factors," and those tied to the overall health of a community, referred to as "location-dependent factors." The study formulates an index for evaluating the impacts of location-dependent variables, representing the community vitality index. Numerous academic studies have demonstrated the influence of location-dependent factors on house prices. Communities offering a high quality of life, excellent schools, healthcare, and low crime rates tend to draw more residents. The high demand for housing in such communities can impact property prices. A study by Florida State University shows that a 10% increase in violent crimes within a neighborhood reduces house values by 6% in Miami-Dade County, Florida. (Goncalves, 2009) Attractive community amenities can positively impact the value of real estate. "Homes adjacent to natural resources like parks and open spaces hold an 8%-20% higher value than comparable properties." (Wolf, 2018). Rising housing prices suggest a community has a healthy economy, including a low unemployment rate, steady incomes, and a flourishing business environment, a critical component of overall community well-being. Therefore, the house-price-based index can serve as a proxy for the MD-CVI and will remedy the deficiencies of the traditional MD-CVIs, such as formidable data collection, inconsistent variables, coarse granularity, and static out-of-date data.

Materials and Methods

A house price is the cumulative impact of house-dependent and location-dependant factors. The relationship can be represented as the formula:
P = f (house-dependent factors) + g (Location-dependent factors) + e

Where P is the house price, f(x) is the cumulative impacts of house-dependent factors, g(x) is the cumulative impacts of location-dependent factors, representing the unprocessed CVI, and e is the error.

This study employs a proportioning algorithm to determine the optimal distribution of price value between the proto-CVI and the regression of the house-dependent factors concerning the least regression R2 value. The resulting CVI can be derived through the price split, as illustrated below:

Cumulative impacts of Location-dependent factors (CVI) = Price – Cumulative impacts of House Dependent Factors - Error

$$\sum \frac{|O_i - E_i|^2}{|E_i|}$$

The significance of the correlation is calculated through the Chi-Square test. Oi represents the vector of observed values, and Ei represents the vector of expected values.

2.1 Data Source and Two Overarching Parts of the Regression Model

Data Source: This study uses nearly 100,000 housing data fetched from the Redfin crossing Southern, Eastern, and Northern regions of the US.

House-dependent Factors: Square Footage, Square Footage of Heated Area, Lot Size, Number of

Rooms, Number of Bedrooms, Number of Bathrooms, The Number of Garages, Whether the house has a pool, Flooring material, External Siding material, Heating fuel type.

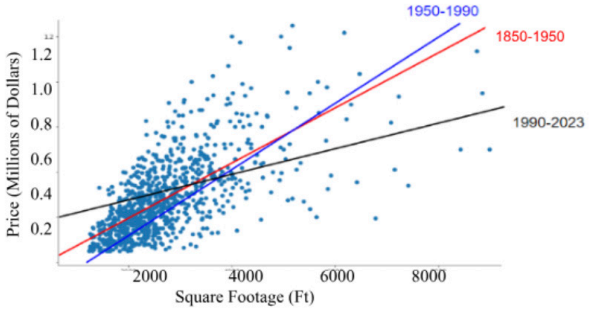
Location-dependent Factors: Crime Rate, Police Count, Number of Nearby Highways, Number of Nearby Local Roads, and Number of Nearby Service Roads.

The dataset was split into binary and non-binary variables, and categorical variables were converted into a series of binary variables.

2.2 Model Development

This study uses a hierarchical regression model (Price Determination Model ~PD-Model) to calculate housing prices and the CVI. It processed the continuous numeric, categorical, binary, and variables separately. Two primary factors, square footage and the segmented build year, were initially added as independent variables. The correlation is 0.6792.

Figure 1. *Year Partitioned Comparison of Sq. Footage and House Price*



The graph depicts the year-segmented relationship. It shows a direct relationship between square footage and housing price, indicating the use of regression.

Afterward, adding more factors into a multiple linear regression model forms a circular cluster, resulting in a reduced R2 value of 0.5381, indicating increased coherence among the model's variables. To address the data circulation issue, HDBSCAN and K-Means++ were applied to cleanse the datasets. K-means was paired with K-means++ to find the optimal number of clusters by location. They also identified outliers via the longitude and latitude of houses and clusters to control location-dependent factors. For Hartford County, location clustering graphed by MatPlotLib is shown below.

Figure 2. Housing Clusters in Hartford County, CT

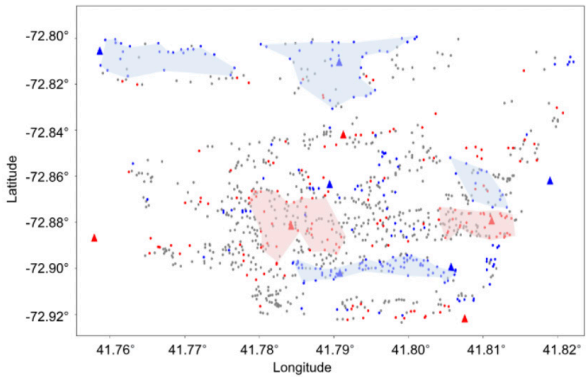


Figure 2 visualizes the regions where the community index is calculated, reflecting low and high house prices. K-means++ identified six high-priced, five low-priced, and ten medium-priced clusters. Red denotes house prices below $350,000, and blue indicates above $750,000. The shaded regions represent actual clusters of houses. Triangles represent clusters of houses.

In outlier detection clustering, HDBSCAN used a distance threshold between two neighborhood data points. If the shape of clusters is elliptical, the accuracy and the R2 value decrease significantly. To obtain the accuracy and reliability model, a clean function was created to remove outlier data in the three steps below to improve the accuracy to 0.7302.

1. HDBSCAN clustering produces a membership score for each data point. All data points with a membership score of 0.7 or higher were excluded, creating a controllable outlier removal system;

2. Linear regression analysis estimates the coefficient for each variable in the housing dataset;

3. The slope values defined two parallel hyperplanes, and the y-intercept was adjusted to encompass 65% of the data within the range of hyperplanes; other data was dropped.

Next, the result is adjusted for inflation as housing prices have increased by 30% since 2019. The inflation rate is obtained from the average price change. Finally, by weighting the proportions, 71.42% of the price was allocated to the house-dependent factors and 28.58% to the location-dependent factors (implying the CVI). The calculated CVI was normalized into the final index, subtracting N to make variance across different CVIs more apparent: Normalized CVI = Unprocessed Community Index - N

Results

The table below lists all the independent variables, coefficients, and confidence for house-dependent factors used in the regression analysis of location-dependent variables. The coefficients describe the strength of correlation and, thus, the importance of each house-dependent factor contributing to housing price and the ranking of the importance of location-dependent variables. The margin for statistical significance is 0.1 or less. The Confidence value used is a P-value from Chi-Square analysis. Categorical and Binary variables are handled separately in the regression model.

Table 1: *House Dependent Variables and Coefficient (* is Statistically Significance)*

House Dependent ~**Numerical Variables**		Coefficient	Confidence
Square Footage		2618*	0.088
Square Footage of Heated Area		149*	0.075
Lot Size		125*	0.018
No. of Rooms		1724*	0.047
No. of Bedrooms		1392	0.715
No. of Bathrooms		1102	0.692
House Dependent ~**Binary/Categorical Variable**		Coefficient	Confidence
Has a Pool?		19983*	0.045
Flooring	Has Hardwood Floor?	19276*	0.098
Exterior Siding	Wood or Other	-912	0.158
	Brick or Stone	15407	0.466
Heating Utility	Natural Gas or Oil	93	0.539
	Electric power	-463	0.572

Table 2: Location Dependent Variables and Coefficient

Location dependent Variable			Coefficient	Confidence	
				Individual	Combined
Location dependent Factors	Accessibility	No. of Nearby Highways	-5161	0.654	0.238
		No. of Nearby Local Roads	4353	0.476	
		No. of Nearby Service Roads	1043	0.142	
	Crime	Crime Rate Normalized(x1000)	-4.76	0.276	0.194
		Police Count	-102	0.445	
	Public Participation		92.4	0.120	

In a multivariate regression model, the factors can be displayed as such, where n is the number of variables, w is the list of all factor coefficients, y is the dependent variable, and n an independent variable:

$$C + \sum_{i=1}^{n} x_i \cdot w_i = y$$

The independent variables with the most significant magnitude of weightings impact the dependent variable the most. By extension, if a list of various terms (xi) is plugged into a multivariate regression, the xi terms with the most significant coefficient magnitude will have the strongest correlation overall. The above formulation can be used to determine the importance of the variables and, by extension, suggest a ranking system to measure how important each variable is to community vitality given the housing-based CVI.

The values in the Coefficient column in Table 2 ranked from most to least important were Accessibility, Crime Rate, and Public Participation sequentially. The correlation comparison produces a ranking of the least and most important factors towards community vitality. The government or a social welfare organization can establish a cutoff threshold to filter out a portion of funding applications and grasp the priority of applications, thereby mitigating the scenario faced by Flint, Michigan. Allocating government spending is fraught with oversight. Flint was overlooked in competitive federal grants since it could not maintain non-federal funding during a downturn in public health, civic participation, and employment, among other factors. (Jeff Arkin) By plugging the factor CVIs into the ranking system in this paper, a set of coefficients describing the importance of each factor is produced, allowing Flint to identify the most critical issues and enable

the community to apply for federal aid grants that effectively address their most pressing needs.

3.1 Verifying the accuracy of House-price-based CVI compared to traditional MD-CVI

The following figures compare the house-price-based CVI with traditional MD-CVIs for the towns of Greater Hartford County. This ensures that the CVI by PD-MODEL is reliable and the results are consistent with the values of conventional MD-CVIs.

Figure 3. *Scaled House-Price-Based CVI and Economic Innovation Group DCI*

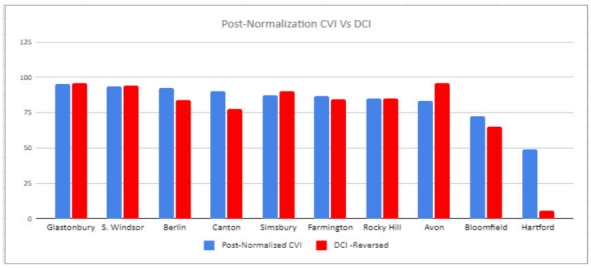


Figure 3 compared the house-price-based CVI with the Distressed Communities Index (DCI) developed by The Economic Innovation Group, a bipartisan public policy think tank. (Economic Innovation Group, 2017) The scale on the left describes the post-normalization scaled house-price-based CVI and the scale on the right represents the DCI.

Figure 4. *House-price-based CVI and DataHaven Community Well-being Index*

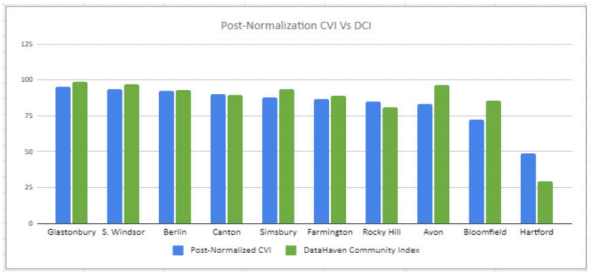


Figure 4 compared the house-price-based CVI with the Community Well-being Index developed by DataHevan. The two figures show that the town ranking of the house-price-based CVI consistently aligns with the town ranking in the two traditional MD-CVIs.

Figure 5. *Comparison of School Ranking, Scaled house-price-based CVI and DCI.*

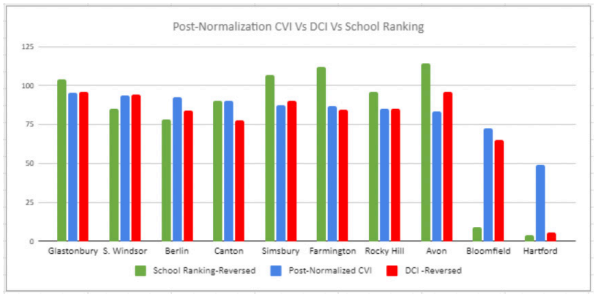
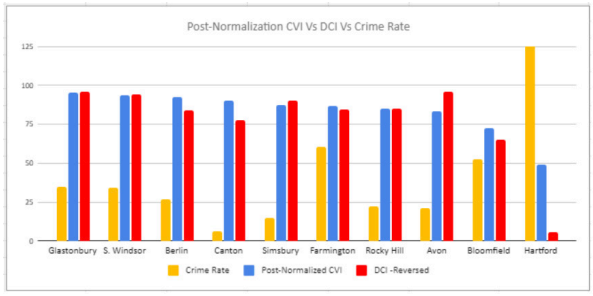


Figure 5 shows that school quality correlates well with the CVI, with both indexes coinciding with the school rankings. The school ranking, DCI, and normalized CVI are scaled to fit in one graph.

Figure 6. *Comparison of Crime Rate, Scaled house-price-based CVI, DCI*



There is a slight trend in terms of the crime rate. The crime rates are similar in communities with a higher vitality index or lower DCI. However, crime rates in areas with high DCI or low vitality index are much higher. Hartford was not included due to its exceptionally high crime rate and DCI preventing scaling of the other crime rates. Note: The data for Hartford/NE Hartford overshoots the graph. The crime rate was obtained from Census Information.

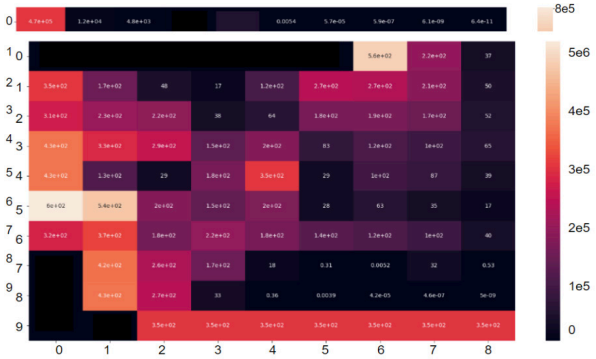
3.2. Exemplification

The figures below showcase PD-MODEL being applied to other metro areas. They depict the average house price of each grid square by region

of Zipcode combinations. Light colors indicate high prices, while dark colors represent low prices. The similarity and high accuracy enable the utilization of PD-MODEL for any community. It will produce a community region and a valid CVI for it.

Figure 7. *Heatmap of average housing prices (Greater Hartford County, CT)*

The high-valued house cluster is at (2,1) because the high-valued region at (4,0) pulls the center to (2,1).
Figure 8. Heatmap of average Texas housing prices (Plano and Houston, TX)



High-valued house cluster is at (1,8).

Figure 9. *Heatmap of average Minnesota housing prices (Plymouth and Minnetonka, MN)*

High-valued house clusters concentrated around (0, 5) and (6,0)

Discussion

The study validated the model based on single-family house prices in suburban areas, where the abundant data increases the model's accuracy. In rural areas where the house price data may not reach a specific density, sparse data can make it difficult for the models to generalize effectively and capture the underlying patterns in the data. In urban areas, different features, such as renting homes, proximity to public transportation, amenities, and population density, may influence the community index differently compared to suburban areas. Therefore, the model must be fine-tuned to accommodate the data pattern in rural or urban areas.

The initially stated goal of the house-price-based CVI model is to provide a synthesized view of various factors affecting the real estate market and the community's overall well-being through a new methodology, showing its advantage in simplicity, accessibility, and real-time monitoring. It may not capture different communities' nuances or unique characteristics and dynamics. In future studies, we endeavor to enhance the conceptual model to address these issues. The work is underway.

PD-MODEL computes the CVI results using data from each house within the zip code area. As data for individual houses was unavailable within census tracts, the CVI wasn't standardized based on them. Nonetheless, this model remains applicable to census tracts, provided they can link and furnish addresses for individual houses.

As each CVI determination region becomes geographically smaller, the effects of sparsely distributed outliers become more significant. For example, even though crime rates usually tend to concentrate on specific "Crime Hotspots," isolated instances of crime still occur in areas with very low concentrations. In a particular case, a region with an outlier crime rate can disproportionately negatively impact the vitality index. Consequently, important factors with sparse distribution may introduce inaccuracies when using smaller areas and data subsets.

Conclusion

This paper adopts an agglomerative hierarchical regression model to derive a community vitality index from house sales information. It examines its close correlation with other existing multi-dimensional community well-being indicators. The house-price-based CVI reflects the collective impacts of the indicators showing a community's well-being.

The house-price-based CVI can be updated based on the latest house sales data, and the community can immediately show its ever-changing dynamics before the census data. It can provide real-time insights into the spatial distribution of economic well-being during economic and social volatility. The level of details in the house sales information determines the CVI's granularity, such that the CVI's flexibility for both small and large areas allows users to create their search criteria on either a city, a county, or a geographic region controlled by zip code combinations. The house-price-based CVI is not intended to offer a complete and definitive view of a community but a new perspective different from traditional multidimensional CVIs. It should be used as a proxy for conventional multidimensional CVIs. Policymakers, organizations, and researchers can use it as a benchmark to verify the effectiveness of traditional multidimensional CVIs and different community-based funding allocations. The PD-CVI model functions as a validation tool, specifically evaluating human intervention within the MD-CVI to refine its value further. I hope this work can attract more curiosity and encourage participation in the ongoing effort to improve our understanding of the community's vitality to identify areas of need.

Acknowledge

I want to thank Professor G.M, Graduate Student Christ. C of Public Policy, Mimi. W of Economics, Staffings of Civil Planning departments, County Finance Department in the Great Hartford, and RA. Shu. C of Neighborhood Reinvestment. Their expertise and knowledge were invaluable during this research.

References

About the Community Index. (2020). Shinyapps. io. https://fourtheconomy.shinyapps.io/Community_Index/#section-about

Arkin, J. (2023). Observations on Challenges with Access, Use, and Oversight. https://www.gao.gov/assets/gao-23-106797.pdf

Azimlu, F., Rahnamayan, S., & Makrehchi, M. (2021, July 7). House price prediction using clustering and genetic programming, along with conducting a comparative study. Proceedings of the Genetic and Evolutionary Computation Conference Companion. Presented at the GECCO '21: Genetic and Evolutionary Computation Conference, Lille France. doi:10.1145/3449726.3463141

Bulatao, R. A. (2000). Read "Beyond six billion: Forecasting the world's population" at nap.edu. The Accuracy of Past Projections | Beyond Six Billion: Forecasting the World's Population | The National Academies Press. https://nap.nationalacademies.org/read/9828/chapter/4

Ceccato, V., & Wilhelmsson, M. (2020). Do crime hot spots affect housing prices? Nordic Journal of Criminology, 21(1), 84–102. doi:10.1080/2578983x.2019.1662595

Community Vitality Index. (2015). PDF. Marron, J. (n.d.). Retrieved 28 December 2023, from https://archives.iupui.edu/bitstream/handle/2450/10109/ThrivingComm_Vitality_Web.pdf

Curtis, J., & Cunningham, M. (2012). Muskie school capstones and dissertations - University of Maine System. USM Digital Common. https://digitalcommons.usm.maine.edu/muskie_capstones/

DataHaven. (2023, March 3). Greater Hartford Community Wellbeing Index. https://www.ctdatahaven.org/reports/greater-hartford-community-wellbeing-index

Economic Innovation Group. (2017). Distressed Community Index 2017. Economic Innovation Group. https://eig.org/wp-content/uploads/2017/09/2017-Distressed-Communities-Index.pdf Compared DCI with CVI

Freemark, Y., et al. (2023, October 25). Is Federal Infrastructure Investment Advancing Equity Goals? Urban Institute. https://www.urban.org/research/publication/is-federal-infrastructure-investment-advancing-equity-goals

Goncalves, J. (2009). EEB--UNDERGRADUATE ECONOMICS JOURNAL. Empirical Economic Bulletin. https://digitalcommons.bryant.edu/cgi/viewcontent.cgi?article=1020&context=eeb

Hallisey, E., Flanagan, B., Kolling, J., & Lewis, B. (2014, March 7). CDC's Social Vulnerability Index. A Social Vulnerability Index (SVI) from the CDC. https://svi.cdc.gov/Documents/Publications/CDC_ATSDR_SVI_Materials/SVI_Poster_07032014_FINAL.pdf

National Association of REALTORS® Research Group Home Buyers and Sellers Generational Trends Report. (n.d.).

Ihlanfeldt, K., & Mayock, T. (2009). Crime and Housing Prices. Department of Economics and DeVoe Moore Center. https://coss.fsu.edu/dmc/wp-content/uploads/sites/8/2020/09/02.2009-Crime-and-Housing-Prices.pdf

Labor Market Trends and Local Job Strategies. (1997). In Labor Market Trends and Local Job Strategies.

Lowe, Kate, et al. "Capacity and Equity: Federal Funding Competition between and within Metropolitan Regions." Journal of Urban Affairs, vol. 38, no. 1, Feb. 2016, pp. 25–41, https://doi.org/10.1111/juaf.12203. Accessed 3 Jan. 2022.

Niche (Ed.). (2023, December 30). 2024 Best School Districts in Connecticut [Fact sheet]. Niche. Retrieved December 30, 2023, from https://www.niche.com/k12/search/best-school-districts/s/connecticut/

Obtained School Data for comparison with CVI Open Street Map. (2023, July 17). Tags - OpenStreetMap Wiki. Wiki.openstreetmap.org; Open Street Map. https://wiki.openstreetmap.org/wiki/Key:Used for Location-Dependent verification of CVI

The American Public Transportation Association and The National Association of Realtors®, "The Real Estate Mantra –Locate Near Public Transportation," October 2019

Pavel K., April 10, 2018. Unemployment and the US Housing Market during the Great Recession. https://web.stanford.edu/~pavelkr/jmp.pdf

The Vermont Community Index: Technical Documentation. (2023, April 13). The Vermont Community Index: Technical Documentation. Retrieved December 27, 2023, from https://finance.vermont.gov/sites/finance/files/documents/VCI%20Technical%20Documentation%20-%20MTAP.pdf

Wolf, K.L. 2010. Community Economics - A Literature Review. In: Green Cities: Good Health (www.greenhealth.washington.edu). College of the Environment, University of Washington.