**Using machine learning in GIS to predict housing affordability changes**

**Author:** Lydia Akalu  
**Course:** DATA-205  
**Date:** December 11, 2025

**1. Introduction and Project Overview**

**1.1 Context and Motivation**

Housing affordability has become an increasingly urgent issue in Montgomery County, Maryland, particularly along major transit investments such as the Purple Line corridor. While transit-oriented development is often promoted as a tool for improving accessibility and economic opportunity, it can also contribute to rising housing costs and displacement pressures in surrounding neighborhoods. Understanding how affordability changes spatially and what factors are most strongly associated with those changes is critical for equitable planning and policy decisions.

This project investigates spatial drivers of housing affordability change in Montgomery County, with a specific focus on block groups along and near the Purple Line corridor. The analysis combines traditional spatial regression, multiscale local models, and machine-learning-based prediction to examine both explanatory relationships and future affordability patterns and to demonstrate how geospatial analysis tools in ArcGIS can be used to analyze and model spatial patterns in housing affordability

**1.2 Data**

The primary dependent variable is the Composite Housing Affordability Index (CHAI), a distribution-based measure summarizing how much household income is spent on housing. CHAI combines renters and owners with mortgages using the formula:

CHAI = R\_share × R\_mean + O\_share × O\_mean

where renter and owner cost burdens are derived from ACS distribution tables rather than simple medians.

**Data sources include:**

* **American Community Survey (ACS) 5-year estimates (2019–2023)**
  + Income distribution (B19001)
  + Educational attainment (B15003)
  + Rent burden (B25070)
  + Owner cost burden (B25091)
* **Montgomery County Planning Department**
  + Zoning snapshots (single-family and multifamily zoning)
  + Development pipeline data
* **Transitland GTFS + Esri tools**
  + Transit service frequency and accessibility (LOSA)

**Inclusions:**

* Block-group-level CHAI (t1)
* Socioeconomic, zoning, and transit variables
* Purple Line corridor context

**Exclusions:**

* Earlier ACS timeframe (2009–2013)
* Change (delta) variables between t0 and t1- due to weaker relations
* Independent municipalities (Rockville and Gaithersburg), due to boundary and data inconsistencies

**1.3 Project Goals**

The goals of this project are to:

1. Identify key socioeconomic, zoning, and transit factors associated with changes in housing affordability.
2. Examine how these relationships vary spatially across Montgomery County.
3. Explore the feasibility of predicting future affordability change under assumed Purple Line service and development scenarios.
4. Demonstrate how geospatial analysis tools in ArcGIS can be used to analyze and model spatial patterns in housing affordability

**1.4 Tools and Methods**

* **Python**: Census API data extraction, preprocessing
* **ArcGIS Pro**:
  + Exploratory Regression
  + Ordinary Least Squares (OLS)
  + Multiscale Geographically Weighted Regression (MGWR)
  + Forest-Based Regression & Prediction (FBCR)
* **ChatGPT/Copilot:** for code generation and explanation of GIS features.
* **GitHub:** for documentation

**2. Summary of Data Cleaning and Pre-Processing**

Data preparation focused on ensuring spatial and temporal consistency across multiple sources:

* Downloaded ACS tables using the Census API and standardized all variables to block-group geography.
* Calculated CHAI using distribution-based methods, excluding “not computed” categories.
* Created t1-only variables (2019–2023) to avoid compounding uncertainty from earlier ACS vintages.
* Performed spatial joins to integrate zoning data, transit accessibility, and development pipeline information.
* Aggregated transit frequency measures into percentage-based accessibility metrics at the block-group level.
* Removed block groups with missing data, inconsistent boundaries, or incomplete zoning classifications.

**3. Basic Descriptive Statistics**

**Table 1. Key Variables and Summary Statistics**

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | ACS Table(s) | Period | Definition / Calculation |
| LOWINC\_SHR\_t0 / LOWINC\_SHR\_t1 / dLOWINC\_SHR | B19001 | t0 / t1 / diff | Households with income <$30,000. Sum bins: <10k, 10–14,999, 15–19,999, 20–24,999, 25–29,999; divide by total households. |
| HIGHEDU\_SHR\_t0 / HIGHEDU\_SHR\_t1 / dHIGHEDU\_SHR | B15003 | t0 / t1 / diff | Percent of population age 25+ with Bachelor’s or higher (BA, MA, Professional, Doctorate ÷ pop 25+). |
| NONWHITE\_SHR\_t0 / NONWHITE\_SHR\_t1 / dNONWHITE\_SHR | B02001 | t0 / t1 / diff | Percent non-White = 1 − (White alone ÷ total population). |

Gis based data

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Source | Period | Definition / Calculation |
| SF\_Zone\_t0 / SF\_Zone\_t1 / dSF\_Zone | Montgomery County Planning (zoning snapshots) | 2007 / 2017 / diff | % of buildable area (non-park) zoned single-family-only. Parks removed; crosswalked categories applied; intersected with 2020 BGs. |
| MF\_Zone\_t0 / MF\_Zone\_t1 / dMF\_Zone | Montgomery County Planning (zoning snapshots) | 2007 / 2017 / diff | % of buildable area where multifamily is allowed (incl. PD districts permitting MF). |
| SF\_Units\_t0 / SF\_Units\_t1 / dSF\_Units | Department of Permitting Services (occupancy permits) | t0 / t1 / diff | Single-family occupancy permits per 1,000 households (BG-normalized). Lag applied to align permits → occupancy → ACS. |
| MF\_Units\_t0 / MF\_Units\_t1 / dMF\_Units | Department of Permitting Services (occupancy permits) | t0 / t1 / diff | Multifamily occupancy permits per 1,000 households (BG-normalized). Lag applied as above. |
| LOSATransit\_t0 / LOSATransit\_t1 / dLOSTransit | Transitland GTFS + Esri Public Transit 'Calculate Transit Service Frequency' (Areas) | Nov 2013 / Nov 2023 / diff | % of BG area with ≥6 trips/hour during AM peak (Areas workflow). |

**Distribution and Stastical description of CHAI variable**

**A map of housing affordability

AI-generated content may be incorrect. A graph of a distribution of a number of bars

AI-generated content may be incorrect.**

This map illustrates housing affordability across block groups in Montgomery County. Darker blue areas indicate higher housing affordability pressure, while lighter blue areas represent locations where residents experience lower affordability pressure (i.e., better affordability.

**4. Final Data Products and Results**

**4.1 Exploratory Regression and OLS**

Exploratory Regression was used to identify stable and meaningful predictors of affordability change. Results show:

* Low-income share: 100% significant, consistently positive
* Higher-education share: 100% significant, consistently negative
* Renter share: 100% significant, strongly positive
* Multifamily zoning: Significant in ~61% of block groups
* Transit accessibility (LOSA): Significant in ~51% of block groups, with mixed effects

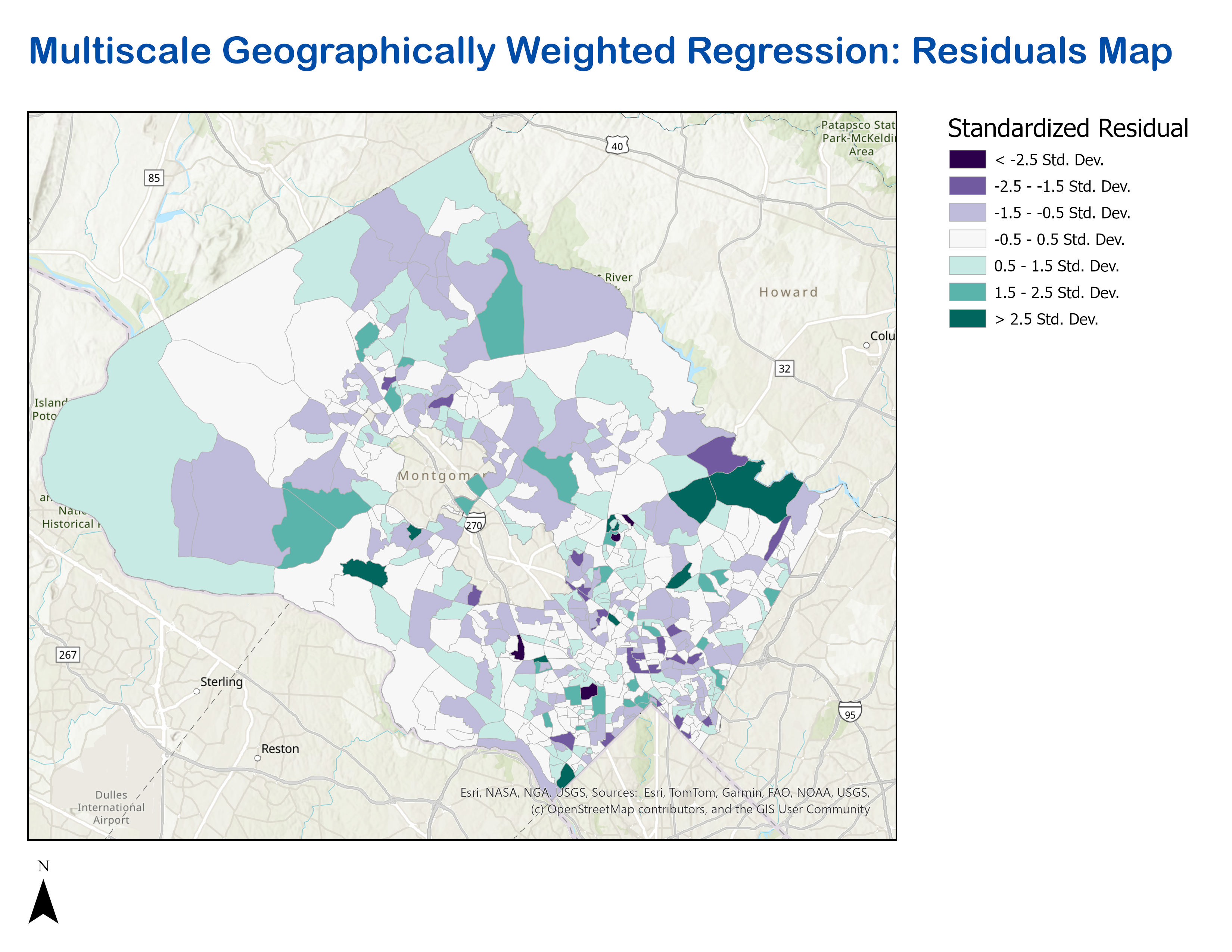
An OLS model using these variables achieved an **R² ≈ 0.54**, indicating moderate explanatory power at the global scale.

**4.2 Multiscale Geographically Weighted Regression (MGWR)**

MGWR revealed that relationships between affordability and explanatory variables operate at different spatial scales:

* Socioeconomic variables (income, renter share) showed strong, spatially consistent effects.
* Zoning and transit variables exhibited localized and context-dependent impacts.
* Some variables were non-significant in large portions of the county, emphasizing that global models alone can mask local dynamics.

MGWR provided critical insight into where certain drivers matter most, even when global significance is weak. As shown In the map below is the residual map for the MGWR model.



**4.3 Forest-Based Regression and Prediction (FBCR)**

A forest-based machine learning model was used to explore **future affordability change** under assumed Purple Line conditions:

**Prediction inputs included:**

* Future transit accessibility assuming 7.5-minute Purple Line headways
* Estimated future renter share from the development pipeline
* Purple Line station proximity

**Key findings:**

* Transit accessibility alone had a weak predictive effect.
* Areas with higher predicted affordability change were more strongly associated with development intensity than with transit proximity.
* Prediction errors showed spatial clustering, suggesting missing neighborhood-level factors and structural complexity beyond the model’s inputs.

**4.4 Value of the Results**

This project demonstrates that:

* **Add additional income variables**  
  Include multiple income categories (not only low-income) to better capture affordability changes across different household income levels.
* **Expand and refine transit data**  
  Separate bus services from rail/metro lines to reflect their different service patterns, coverage, and impacts on housing affordability.
* **Use alternative transit accessibility measures**  
  Move beyond service frequency and test measures such as job accessibility by transit, travel-time accessibility, or access to essential services.
* **Run scenario-based analyses**  
  Evaluate different future assumptions, such as varying Purple Line headways or alternative development pipeline scenarios, to assess potential affordability outcomes.
* **Incorporate domain expertise**  
  Collaborate with housing economists and urban planners to improve variable selection, interpretation, and alignment with real-world housing market dynamics.

**5. References and Acknowledgements**

**References**

* [Census API](https://www.census.gov/data/developers/data-sets.html)
* [Census Geographic Crosswalks](https://www.nhgis.org/geographic-crosswalks)
* [GTFS](https://gtfs.org/documentation/schedule/reference/)
* [TransitLand](https://www.transit.land/)
* [Esri Public Transit Model](https://pro.arcgis.com/en/pro-app/3.4/help/analysis/networks/transit-data-model.htm)
* [Transit Frequency Tool](https://pro.arcgis.com/en/pro-app/latest/tool-reference/public-transit/calculate-transit-service-frequency.htm)

**Acknowledgements**

I would like to express my sincere gratitude to:

* Russell Provost, for mentorship and guidance throughout this project.
* Professor Lori Perine, for continuous feedback and academic support.
* The Montgomery County Government data team, for access to data and applied learning opportunities.
* Montgomery College and Professor Saidi, for providing the academic foundation and opportunity to engage in a real-world, community-impact project.