A Comparison of Submarket Separation Techniques: A Case in the City of Atlanta, GA

Mid-Continent Regional Science Association 53rd Annual Conference

Thursday, June 8, 2023

Hierarchical linear model

$$price = \beta_0 + \beta_1(sqft) + \beta_2(sqft * testscore) + \beta_3 age + \beta_4 age^2 + e$$

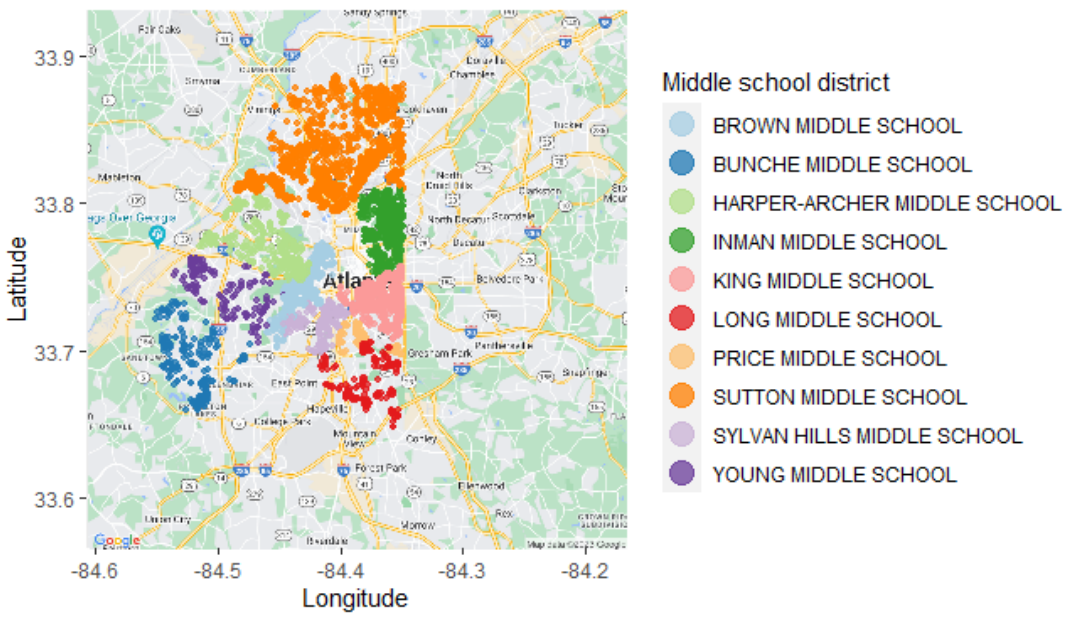
- Estimate the parameters of the equation for any two adjacent school zones.
- If the parameter of the interaction term between square footage and test score is statistically different from zero, then the school zones are assigned to different submarkets
  - It not, they belong to the same submarket and another adjacent school zone is added, and the process is repeated for all school zones.

Hierarchical linear model

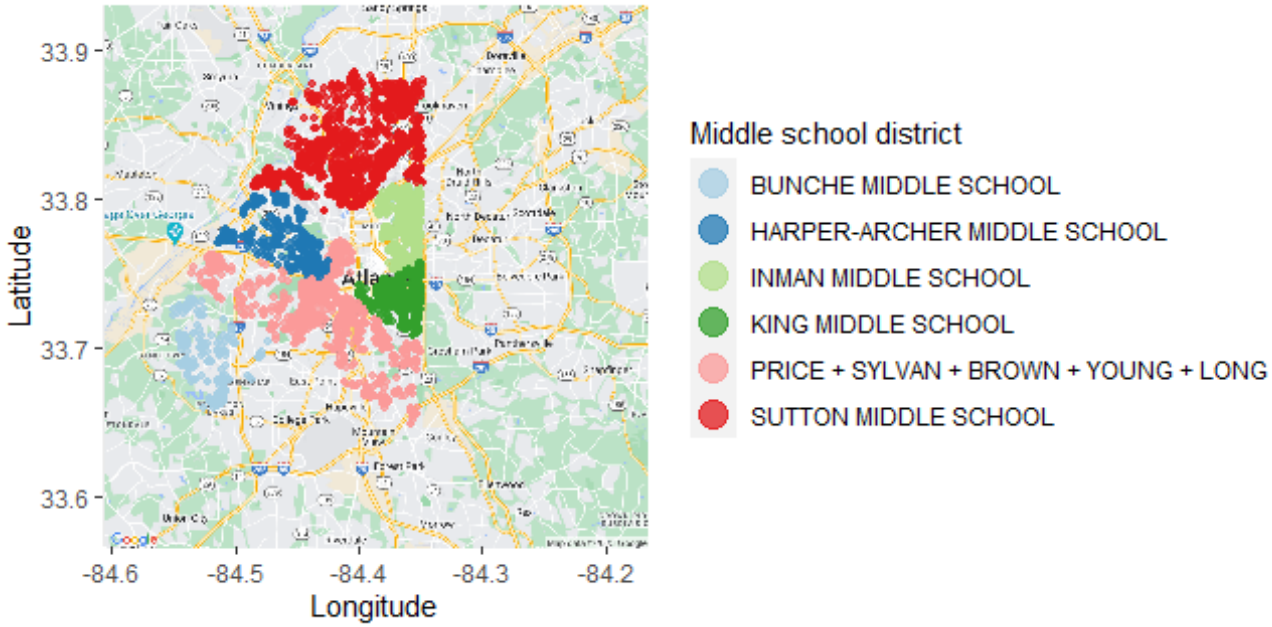
	Obs.	RMSE	Adj. R-squared	AIC
All school districts	5,069	179,687	0.7908	20,306.24
BUNCHE MIDDLE SCHOOL	234	30,474	0.9426	123.21
HARPER-ARCHER MIDDLE SCHOOL	386	31,213	0.8052	224.52
INMAN MIDDLE SCHOOL	628	133,073	0.7170	2,168.02
KING MIDDLE SCHOOL	898	50,656	0.8618	1,352.90
PRICE + SYLVAN + BROWN + YOUNG + LONG	1,173	34,699	0.7628	871.63
SUTTON MIDDLE SCHOOL	1,750	243,349	0.7056	8,105.91

Hierarchical linear model

Atlanta school districts



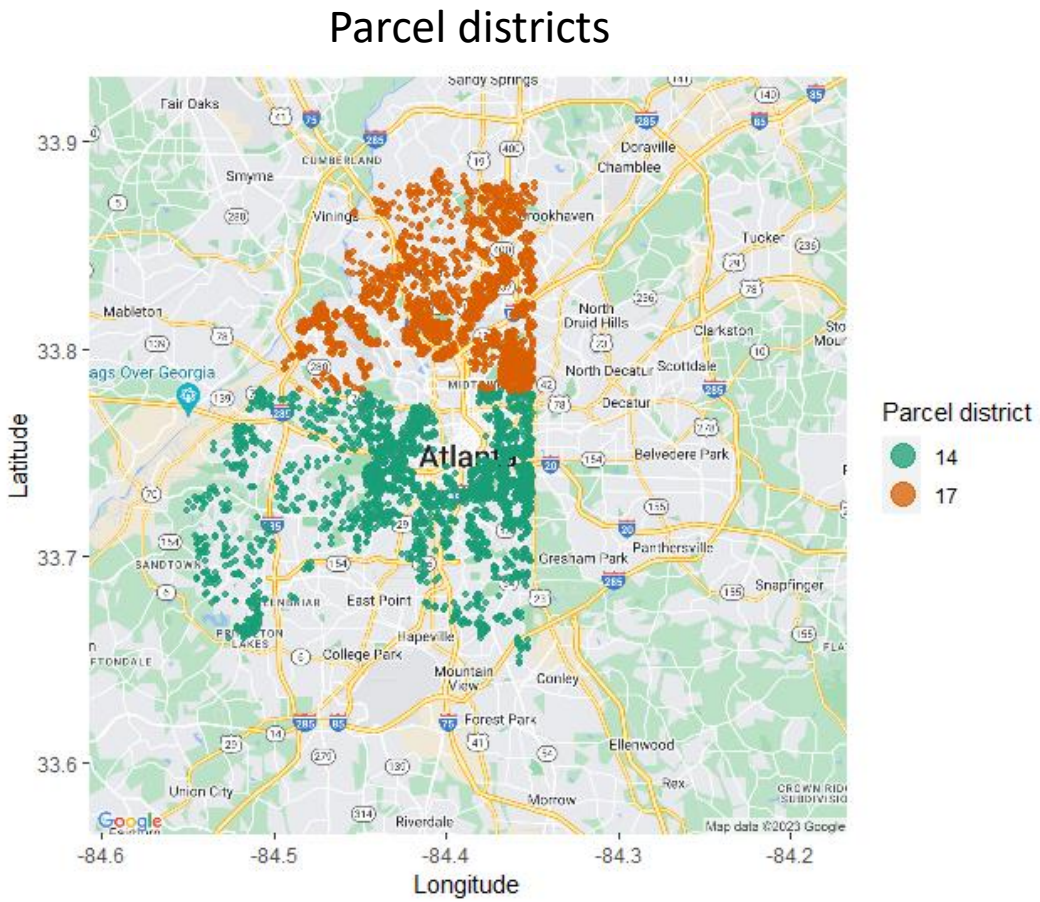
After GT segmentation



Quasi-hierarchical (market segmentation by parcel district)

	Obs.	RMSE	Adj. R-squared	AIC
Parcel 14	2,834	63,627.62	0.8986	5,472.613
Parcel 17	2,235	229,884.7	0.7146	10,084.48

Quasi-hierarchical (market segmentation by parcel district)



Fully endogenized finite mixture model: EM algorithm (1/5)

- This model employs a finite mixture model to sort households into endogenously determined latent submarkets. The finite mixture model to predict home prices is:

$$h(P_i|x_i,\beta_j,p_j)=\sum_{j=1}^{m}\pi(z_i)f(P_i|x_i,\beta_j)$$

- The mixing model $\pi(z_i)$, is used to assign each observation a percentage chance of belonging to each latent submarket and $f(.)$ is a submarket specific conditional hedonic regression. The home price is therefore a weighted average of predicted values across submarkets weighted by the probability of being located in the submarket.

Fully endogenized finite mixture model: EM algorithm (2/5)

- We also define $(d_i = d_{i1}, d_{i2}, \ldots, d_{im})$ to be binary variables that indicate the inclusion of household i into each latent group. These are incorporated into the likelihood function based on a logistic function which are conditional on factors that do not directly influence the price of the house.
- Since the submarket identification (d) is not directly observable, an expectation maximization (EM) algorithm is used to estimate the likelihood of class identification:

$$d_{ij} = \frac{\pi_j f_j(P_i|x_i, \beta_j)}{\sum_{j=1}^{J} \pi_j f_j(P_i|x_i, \beta_j)}$$

Fully endogenized finite mixture model: EM algorithm (3/5)

- The Expectation step – the E step – involves imputation of the expected value of d_i given the mixing covariates, interim estimates of γ, β, π. The Maximization step – the M step – involves using estimates of d_i from the E step to update the component fractions of π_j and β. The EM algorithm can be summarized as:

1. Generate starting values for γ, β, π
2. Initiate iteration counter for the E-step, t(initial t at 0)
3. Use β^t and π^t from Step 2 to calculate provisional d^t from $d_{ij} = \frac{e^{\gamma_j z_i}}{1+\sum_{C=1}^{C} e^{\gamma_j z_i}}$

Fully endogenized finite mixture model: EM algorithm (4/5)

4. Initiate second iteration counter, v, for the M-step

5. Interim estimators of d^{t+1} are then used to impute new estimates of β^{v+1} and π^{v+1} with $d_{ij} = \frac{\pi_j f_j(P_i|x_i,\beta_j)}{\sum_{j=1}^{J} \pi_j f_j(P_i|x_i,\beta_j)}$

6. For each prescribed latent class, estimators of β^{v+1} are imputed, via M-step, as well as π^{v+1}

7. Increase v counter by 1, and repeat M-step until: $f(\beta^{v+1}y, x, \pi, d) - f(\beta^{v}y, x, \pi, d) < a$ prescribed constant; if so, then $\beta^{t+1} = \beta^{v+1}$

8. Increase t counter and continue from step 3 until: $f(\beta^{t+1}, \pi^{t+1}, d|y) - f(\beta^{t}, \pi^{t}, d|y) < a$ prescribed constant

Fully endogenized finite mixture model: EM algorithm (5/5)

- d_{ij} is estimated simultaneously with the estimation of the hedonic regression parameters, which are conditional on class identification.
- This process is repeated until there is no change in the likelihood function: $LogL = \sum_{i=1} \sum_{j=1} d_{ij} \log\left[f_j\left(P_i \middle| x_i, \beta_j\right)\right] + d_{ij} \log\left[\pi_j\right]$
- The steps above, particularly from Step 3-8 do not necessarily occur sequentially as outlined above but occur simultaneously as the continual updating of estimators. Each v iteration conditionally maximizes the likelihood function using interim estimates of observation latent class membership probabilities in one of the latent classes; while each t iteration updates latent class memberships.
- The modified hedonic regression is: $y_{ij} = d_{ij}\left(\beta_j X_i\right) + \epsilon_{ij}$

Fully endogenized FMM

	AIC	Weighted RMSE	R-squared
2 submarkets	23,366.83	163,129.2	0.8295
3 submarkets	25,565.62	143,928.4	0.8673
4 submarkets	27,603.37	108,951.3	0.9240
5 submarkets	29,597.14	106,008.2	0.8669

Summary statistics of FMM submarkets assigned to highest probability (2 submarkets)

Not much distinction in submarkets by race: Just separates ***white north*** from ***black south***

	N	Mean	SD
Sales price	769	890,354	605,620
Sqft	769	3,426	1,849
Math score	769	521	23.3
Lot size	769	0.451	0.45
Median income	769	122,239	66,410
Age of house	769	56.6	31.0
Pct white	769	77.3	26.0
Pct black	769	16.3	25.6
Pct over-65	769	8.40	6.50
Pct college degree	769	77.4	18.1
Tract cover	769	23.6	7.72
Crime	769	0.106	0.08
Pct renter	769	32.7	26.8

Summary statistics of FMM submarkets assigned to highest probability (2 submarkets)

Not much distinction in submarkets by race: Just separates ***white north*** from ***black south***

	N	Mean	SD
Sales price	4,159	333,172	262,444
Sqft	4,159	1,706	972
Math score	4,159	498	27.8
Lot size	4,159	0.271	0.222
Median income	4,159	70,327	47,849
Age of house	4,159	58.3	30.1
Pct white	4,159	46.5	36.8
Pct black	4,159	48.2	38.9
Pct over-65	4,159	8.20	6.80
Pct college degree	4,159	55.6	27.3
Tract cover	4,159	21.5	7.32
Crime	4,159	0.161	0.103
Pct renter	4,159	46.7	23.9

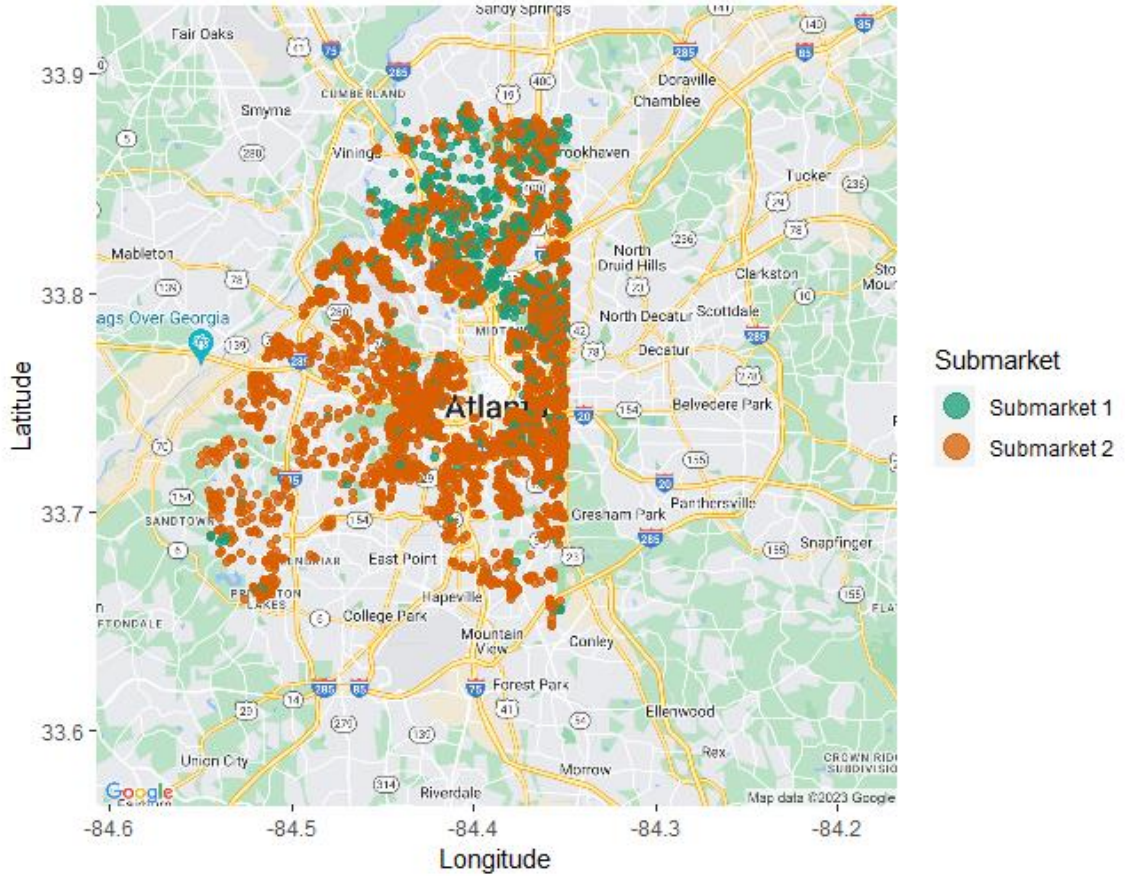
Regression coefficients for probability of being assigned to submarkets (2 submarkets)

	Prob Sub1	Prob Sub2
Intercept	-109.227***	209.227***
Pct black	-0.019*	0.019*
Pct renter occupied	-0.016	0.016
Pct college degree	-0.015	0.015
Math score	0.143***	-0.143***
Log median income	6.874***	-6.874***

Not much distinction in submarkets by race: Just separates ***white north*** from ***black south***

Fully endogenized FMM (2 submarkets)

2 submarkets



Summary statistics of FMM submarkets assigned to highest probability (3 submarkets)

Not much distinction; ***value/sqft & lot size*** differ more than typical

	N	Mean	SD
Sales price	2,843	379,062	331,087
Sqft	2,843	1,554	1,077
Math score	2,843	498	28.2
Lot size	2,843	0.269	0.228
Median income	2,843	72,030	52,318
Age of house	2,843	61.0	29.3
Pct white	2,843	46.8	37.2
Pct black	2,843	48.0	39.1
Pct over-65	2,843	8.30	6.80
Pct college degree	2,843	55.7	27.7
Tract cover	2,843	21.7	7.46
Crime	2,843	0.164	0.107
Pct renter	2,843	46.9	24.7

Summary statistics of FMM submarkets assigned to highest probability (3 submarkets)

Not much distinction; ***value/sqft & lot size*** differ more than typical

	N	Mean	SD
Sales price	1,672	354,627	255,170
Sqft	1,672	2,192	1,160
Math score	1,672	503	27.4
Lot size	1,672	0.308	0.269
Median income	1,672	77,923	49,135
Age of house	1,672	52.7	31.1
Pct white	1,672	52.2	36.2
Pct black	1,672	42.0	38.2
Pct over-65	1,672	8.00	6.80
Pct college degree	1,672	60.0	26.3
Tract cover	1,672	21.7	7.25
Crime	1,672	0.143	0.091
Pct renter	1,672	43.6	23.9

Summary statistics of FMM submarkets assigned to highest probability (3 submarkets)

Not much distinction; ***value/sqft & lot size*** differ more than typical

	N	Mean	SD
Sales price	413	967,884	719,584
Sqft	413	3,403	2,051
Math score	413	523	23.8
Lot size	413	0.468	0.488
Median income	413	124,513	66,983
Age of house	413	59.6	30.6
Pct white	413	78.1	26.5
Pct black	413	15.6	26.1
Pct over-65	413	8.60	6.50
Pct college degree	413	78.3	18.0
Tract cover	413	23.5	7.65
Crime	413	0.108	0.083
Pct renter	413	32.2	26.5

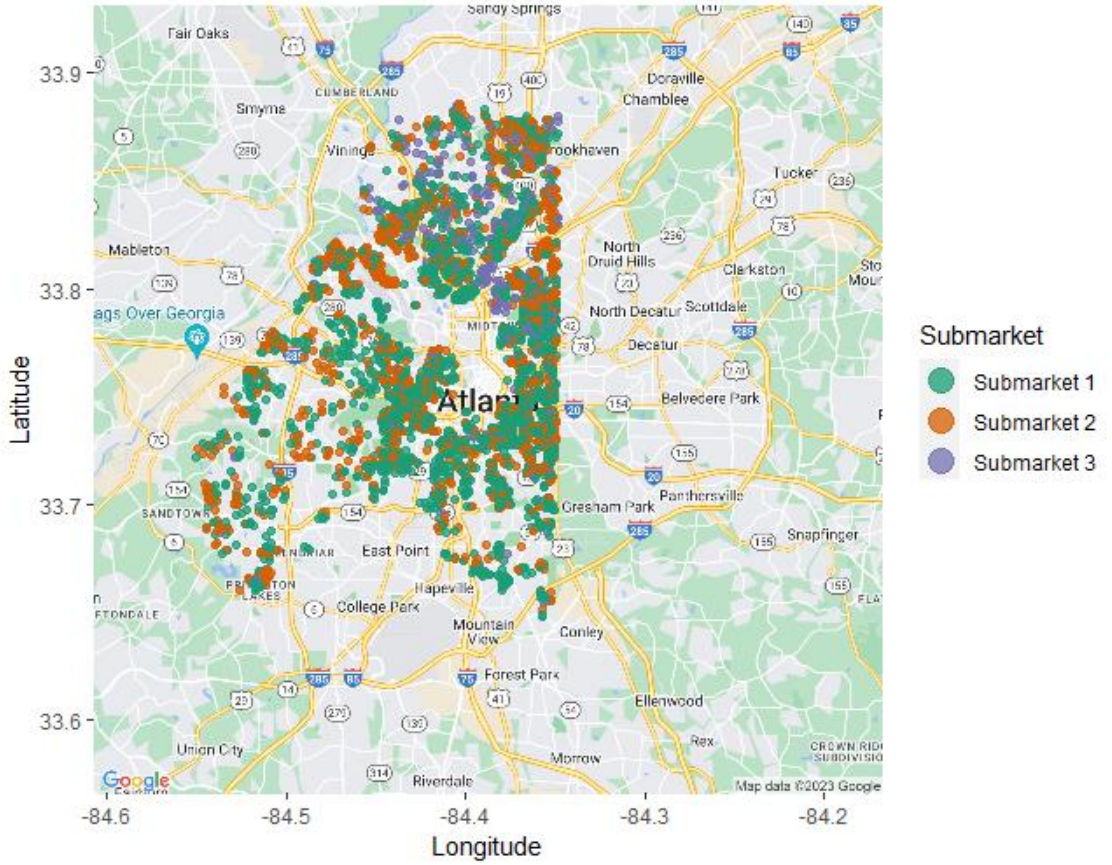
Regression coefficients for probability of being assigned to submarkets (3 submarkets)

	Prob Sub1	Prob Sub2	Prob Sub3
Intercept	89.927***	111.426***	-101.400***
Pct black	-0.032	0.005*	0.027
Pct renter occupied	0.047*	-0.026	-0.022
Pct college degree	-0.006	0.011	-0.005
Math score	-0.084***	-0.053***	0.136***
Log median income	-0.878	-4.083***	4.961***

Not much distinction; ***value/sqft & lot size*** differ more than typical

Fully endogenized FMM (3 submarkets)

3 submarkets



Summary statistics of FMM submarkets assigned to highest probability (4 submarkets)

Submarket 3 most distinct

	N	Mean	SD
Sales price	2,629	381,036	342,189
Sqft	2,629	1,906	1,103
Math score	2,629	498	28.3
Lot size	2,629	0.269	0.24
Median income	2,629	71,608	51,913
Age of house	2,629	61.5	29.0
Pct white	2,629	46.6	37.3
Pct black	2,629	48.1	39.2
Pct over-65	2,629	8.30	6.80
Pct college degree	2,629	55.4	27.8
Tract cover	2,629	21.7	7.44
Crime	2,629	0.164	0.107
Pct renter	2,629	47.0	24.6

Summary statistics of FMM submarkets assigned to highest probability (4 submarkets)

Submarket 3 most distinct

	N	Mean	SD
Sales price	1,100	411,642	290,447
Sqft	1,100	2,455	1,249
Math score	1,100	505	26.7
Lot size	1,100	0.322	0.287
Median income	1,100	83,501	54,004
Age of house	1,100	50.9	32.3
Pct white	1,100	55.6	35.3
Pct black	1,100	38.5	37.2
Pct over-65	1,100	7.80	6.70
Pct college degree	1,100	62.0	26.3
Tract cover	1,100	21.9	7.42
Crime	1,100	0.142	0.091
Pct renter	1,100	42.7	24.6

Summary statistics of FMM submarkets assigned to highest probability (4 submarkets)

Submarket 3 most distinct

	N	Mean	SD
Sales price	257	1,222,571	745,874
Sqft	257	3,486	1,709
Math score	257	528	19.3
Lot size	257	0.471	0.485
Median income	257	135,354	67,393
Age of house	257	62.5	29.2
Pct white	257	83.9	18.8
Pct black	257	9.96	17.1
Pct over-65	257	9.30	6.70
Pct college degree	257	82.0	13.0
Tract cover	257	23.0	7.85
Crime	257	0.104	0.08
Pct renter	257	29.2	26.1

Summary statistics of FMM submarkets assigned to highest probability (4 submarkets)

Submarket 3 most distinct

	N	Mean	SD
Sales price	942	320,164	225,162
Sqft	942	2,528	1,461
Math score	942	503	28.5
Lot size	942	0.311	0.271
Median income	942	76,005	48,534
Age of house	942	55.6	29.9
Pct white	942	50.3	37.3
Pct black	942	44.2	39.2
Pct over-65	942	8.10	6.80
Pct college degree	942	59.5	26.2
Tract cover	942	21.90	7.20
Crime	942	0.144	0.094
Pct renter	942	44.0	24.0

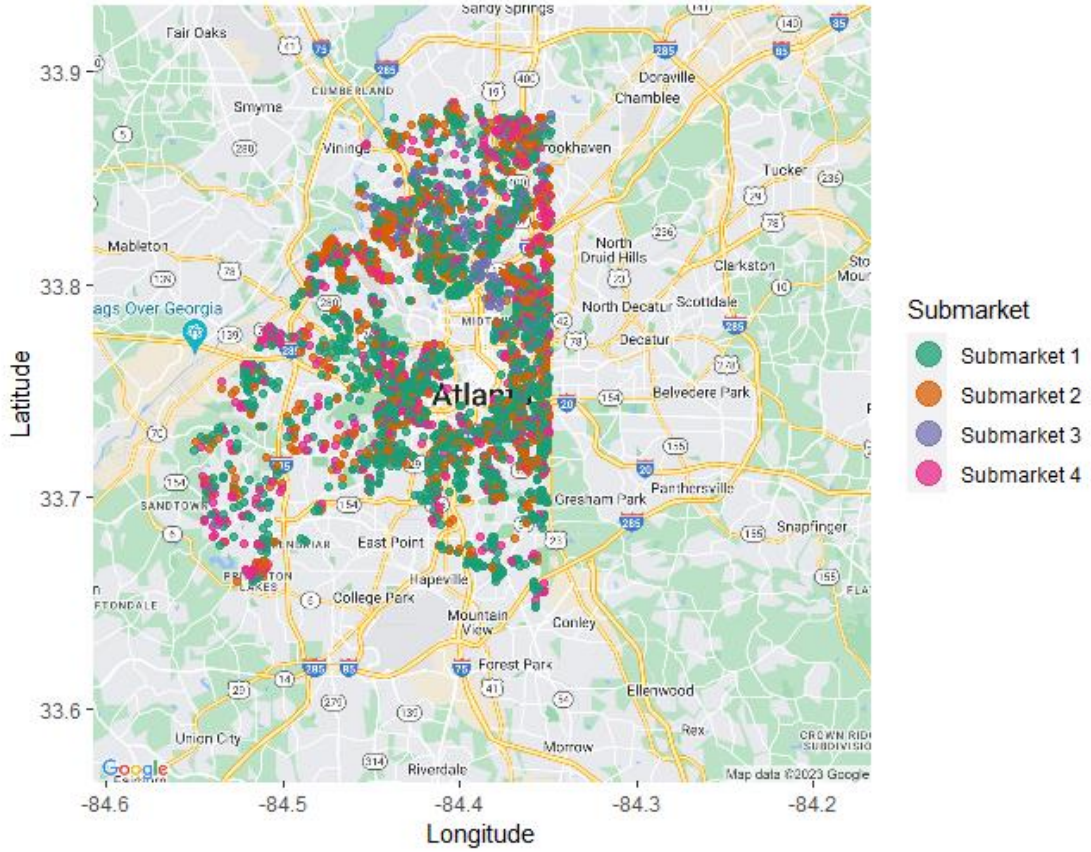
Regression coefficients for probability of being assigned to submarkets (4 submarkets)

	Prob Sub1	Prob Sub2	Prob Sub3	Prob Sub4
Intercept	35.329**	68.597***	-68.166***	64.240***
Pct black	-0.019	-0.021	-0.011*	0.029*
Pct renter occupied	0.033*	0.004	-0.014	-0.023
Pct college degree	-0.025	-0.014	-0.017	0.056*
Math score	-0.043**	-0.054***	0.111***	-0.014
Log median income	1.411	-1.033	2.626**	-3.003***

Submarket 3 most distinct

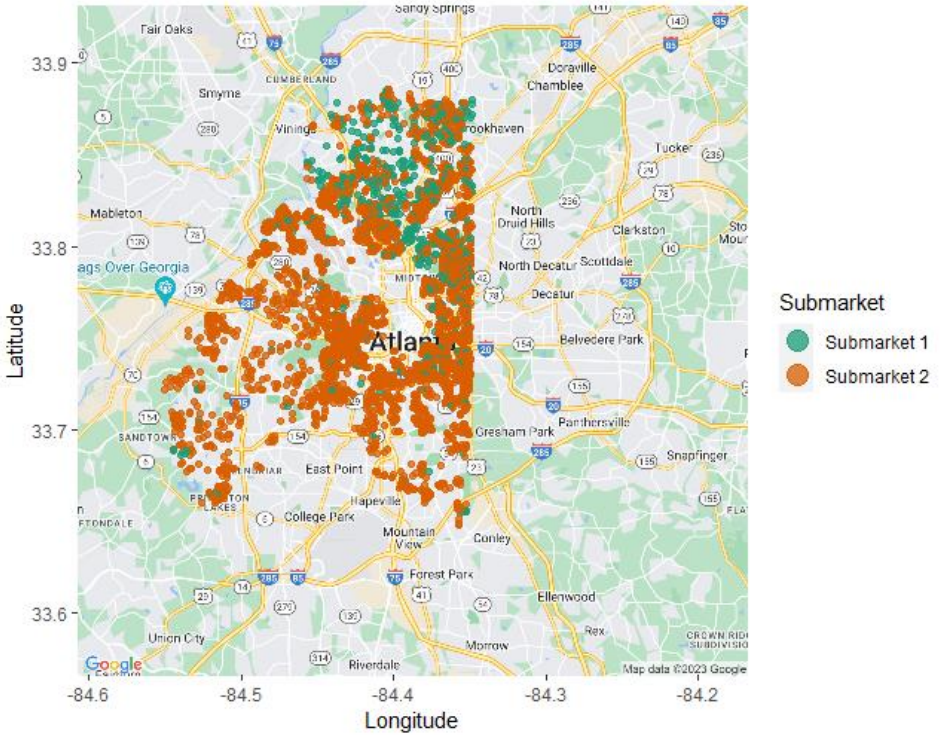
Fully endogenized FMM (4 submarkets)

4 submarkets

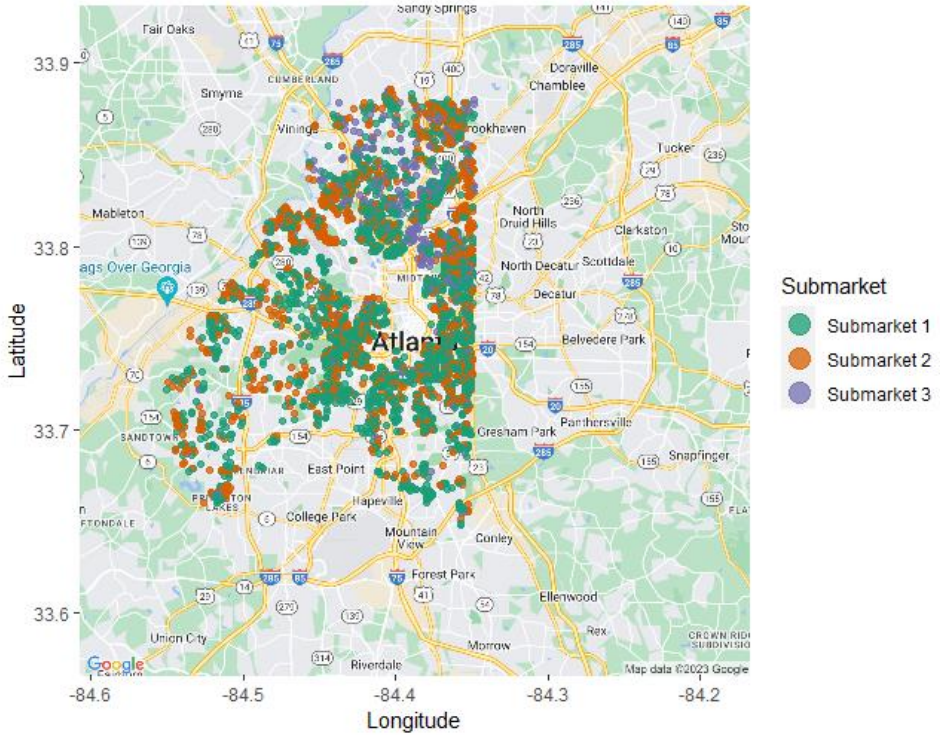


Fully endogenized FMM (2 and 3 submarkets)

2 submarkets

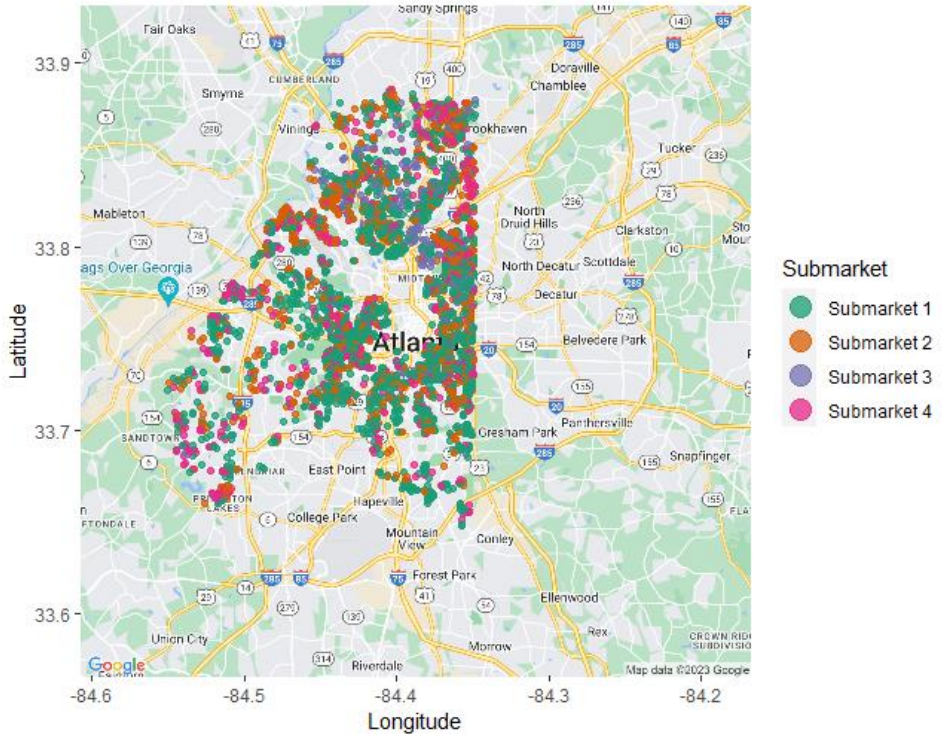


3 submarkets



Fully endogenized FMM (3 and 4 submarkets)

4 submarkets



5 submarkets