PAF 586 – Data for the Public Good

Lab 01 Solutions

These papers represent two data-driven approaches to understanding and predicting neighborhood change. Machine learning algorithms are used to “discover” coherent communities within the city by grouping census tracks into clusters that all have similar characteristics.

Market Value Analysis: A Data-Based Approach to Understanding Urban Housing Markets. pp 49-59 [PDF]

Delmelle, E. C. (2017). Differentiating pathways of neighborhood change in 50 US metropolitan areas. Environment and planning A, 49(10), 2402-2424. [PDF]

How did each author identify coherent “neighborhoods” (or groups) in each model?

Both models use a clustering technique, which are bottom-up inductive approaches to identifying groups. In other words, the neighborhoods in the study were not defined by administrative boundaries such as planning department neighborhoods boundaries, zip codes, or voting districts. They were identified by the algorithms in the study so they represent as much as possible coherent units that share similar characteristics.

Did the two models use the same data to create the groups?

Both use census data as inputs, but they use different variables. The MVA paper focuses almost exclusively on housing market characteristics, and the differentiating pathways paper includes economic measures and population characteristics.

Census Variable	Market Value Analysis	Differentiating Pathways
	ECONOMIC MEASURES	
% with a college degree		X
% manufacturing		X
% service industry		X
% below poverty		X
% unemployed		X
	HOUSING	
median home value	**X**	**X**
variance in sale price	X	
% homes over 30 years old		X
% homes vacant	**X**	**X**
foreclosure rates	X	
% homes owner occupied	**X**	**X**
% housing units are multiunit		X

% of subsidized rental units	X	
number of commercial units	X	
% households arrive < 10 years ago		X
POPULATION CHARACTERISTICS		
population density	X	
% > 60 years old		X
% < 18 years old		X
% black		X
% white		X
% Hispanic		X
% Asian		X
% foreign born		X

How do the labels and descriptions of the groups differ in each model and why?

The neighborhood types are a function of the input data. The Market Value Analysis paper is trying to model home values, so they focus almost entirely on variables describing housing markets. The Differentiating Pathways paper includes characteristics of people in the neighborhoods, so they groups they identify describe population characteristics more than housing markets.

Market Value Analysis

- Competitive home markets
- Strong home markets
- Transitional home markets
- Distressed neighborhoods

Differentiating Pathways

- Wealth, white, educated
- New single-family homes, white
- Older homes, white, some Hispanic, blue collar
- Hispanic and black, higher poverty, aging homes
- Black, high poverty, vacant homes
- Hispanic & black, higher poverty, aging homes
- Black, high poverty, vacant homes
- Hispanic, high poverty, single-family homes, foreign born
- Mixed race, average poverty, renters
- Asians, foreign born, muli-unit, high poverty, recent in-movers
- White & Asian, multiunit housing, educated, recent in-movers, few kids

It should be noted that the machine learning algorithms are good at identifying groups within the data. However, the groups will change completely with the input variable. The first paper focuses on housing characteristics and ignores demographics. But demographic changes drive home prices. And the second ignores housing market characteristics, but changes in home values drive demographic shifts in neighborhoods.

Many data-driven approaches to modeling are good at predicting change using rich input data, but they are not always good at uncovering the underlying theory or mechanisms driving the changes.

Would these "neighborhoods" line up with neighborhoods that are defined on a city's zoning maps?

No, neighborhoods on a city planning map are often defined for historical or arbitrary purposes. They usually do not consist of coherent and homogeneous groups and will have some thriving parts and some blighted parts. This can make it difficult to use traditional neighborhood boundaries difficult to use for planning purposes.

The authors of the MVA paper discuss the advantages of using the groups identified by the clustering algorithms instead of traditional neighborhoods, since the policy prescriptions for would be similar for all blocks within a given category. This can help the city better direct resources (pp 55-56).

- *Stronger markets, displayed in the blue and purple ranges, can often operate as nodes of strength upon which interventions in markets that manifest early signs of blight can be based.*
- *Large areas of distress dictate larger-scale interventions because the degree of blighting influences is so great that minor interventions (e.g., a small number of scattered site housing rehabilitations) are not likely to promote market change.*
- *Yellow markets (labeled in some cities as "transitional markets") draw attention, especially when adjacent to more stable (blue) markets because, in Baltimore, they may be being undermined by high levels of financial distress as evidenced by elevated foreclosure levels. That which is destabilizing the yellow markets may threaten the blue areas.*

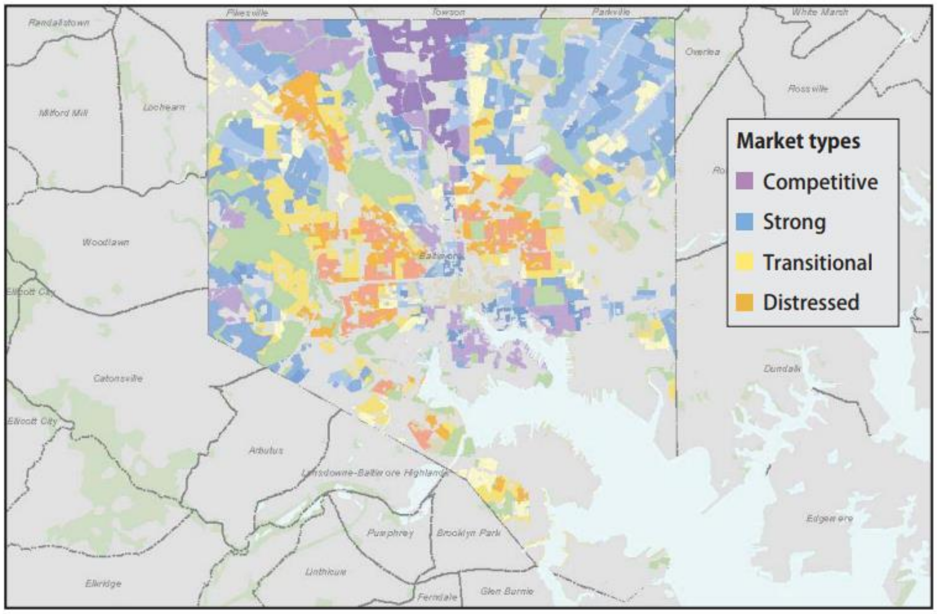
It is true that cities often use groups of census blocks or census tracks to construct their neighborhood boundaries, but each neighborhood defined by the city would likely contain several of each of the groups identified in the study.

Looking Ahead

Clustering algorithms start with a set of observations, like census tracts, and a set of characteristics. The algorithm will assign all of the census blocks to the categories that maximize the within-group similarities and across-group differences (i.e. create the categories that are most distinctive from each other). Afterwards the researcher will look at the characteristic of each group and try to assign meaningful labels. The characteristics of groups from the two studies are reported in the tables below.

In the next lab you will work with census data to create a new scale by combing five separate variables. You will need to name this scale according to the variables you select. This exercise requires some tea leaf reading, and some creativity. Take a look at the variables below to see if you could come up with better names for the groups than what the researchers used.

Market Value Analysis



Baltimore MVA category characteristics, 2008

Market	Distribution summary: Block groups	Distribution summary: Housing units	Distribution summary: % of city housing	Analysis variables: Foreclosures 2006–07 as % of all owner-occupied units	Owner-occupied housing units, %	Vacant housing notices as % of all housing units	Vacant lots as % of all parcels	Occupied single-family units as % of all	Housing units per acre	Commercial land use area as % total block group area	Section 8 units 2008 as % of all rental units	Median 2006–07 sales price
A	12	7,071	2.40	0.87	67.42	0.01	0.00	98.65	6.22	2.05	4.10	$615,915
B	63	33,584	11.41	1.98	51.21	0.92	0.04	93.90	19.43	0.36	1.56	$293,598
C	30	16,261	5.52	3.37	45.19	2.45	0.48	91.93	17.19	18.77	2.22	$244,309
D	117	51,865	17.61	4.50	58.60	1.69	0.46	93.91	11.31	0.00	8.51	$161,447
E	72	37,488	12.73	4.33	54.30	1.68	0.25	93.57	10.86	4.71	7.55	$153,311
F	59	24,700	8.39	5.45	50.09	5.72	0.75	87.54	14.63	18.79	16.37	$97,409
G	118	45,714	15.53	5.75	51.93	6.09	1.01	87.63	15.52	0.07	17.37	$80,315
H	89	29,374	9.98	6.53	34.84	24.01	5.24	66.85	15.91	8.50	11.07	$40,409
I	85	25,786	8.76	6.72	33.40	27.28	4.91	63.18	18.56	0.00	10.30	$36,119
Block group average				5.03	48.02	9.69	2.10	83.87	14.89	4.35	10.26	$130,712

Differentiating Pathways

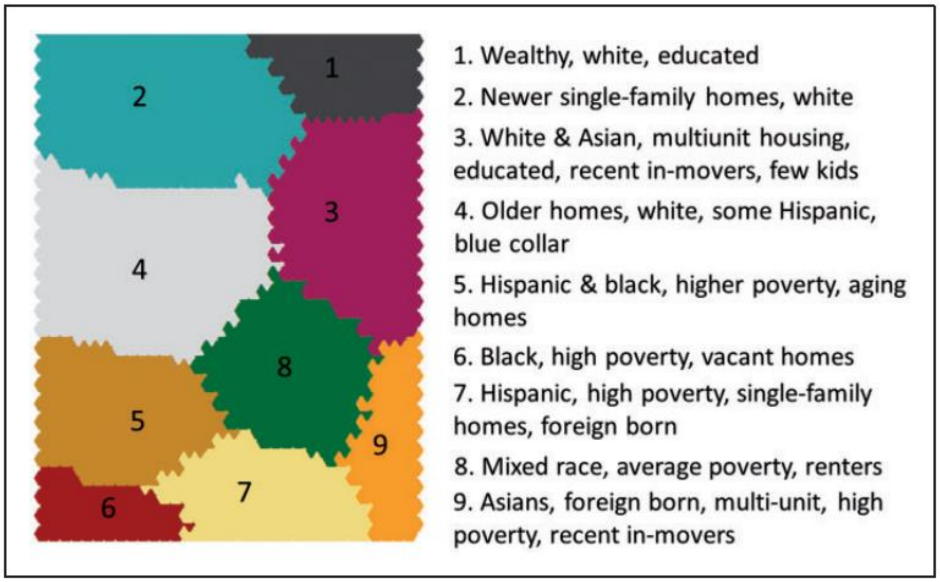


Table 1. Mean z-score of each cluster.

Cluster Number	Median Home Value	% with 4-year college degree	% unemployed	% manufacturing	% below poverty	% homes built >30 years ago	% households moved <10 years ago	% > 60 years old	% service industry
1	1.39	1.45	–0.63	–0.55	–0.67	–0.21	–0.13	0.57	1.19
2	0.23	0.22	–0.45	0.02	–0.64	–0.62	–0.1	–0.12	0.26
3	0.36	0.74	–0.41	–0.46	–0.4	–0.26	0.53	0.23	0.03
4	0.4	–0.5	–0.13	0.31	–0.28	0.01	–0.22	0.12	–0.29
5	–0.71	–0.8	0.41	0.34	0.38	0.57	–0.32	–0.17	–0.56
6	–0.96	–0.01	1.35	–0.04	1.75	0.79	–0.12	–0.3	–0.78
7	–0.65	–0.82	0.68	0.28	1.17	0.54	0.66	–0.47	–0.58
8	–0.35	–0.3	0.03	–0.06	0.08	0.25	0.38	0	–0.4
9	–0.12	0.23	0.35	–0.54	0.73	0.26	1.07	–0.18	–0.42

Cluster Number	% vacant housing	% owner occupied housing	% multiunit structures	% < 18 years old	% black	% white	% Hispanic	% Asian	% Foreign Born
1	–0.04	0.6	–0.35	–0.38	–0.5	0.7	–0.52	0.27	–0.08
2	–0.35	0.85	–0.78	0.25	–0.47	0.66	–0.45	–0.14	–0.48
3	0.04	–0.43	0.65	–0.87	–0.36	0.44	–0.3	0.42	0.15
4	–0.2	0.3	–0.34	0.03	–0.3	0.32	–0.12	–0.19	–0.27
5	–0.07	–0.06	–0.19	0.5	0.54	–0.88	0.67	–0.3	0.24
6	0.84	–0.86	0.47	0.98	2.03	–1.85	0.34	–0.55	–0.21
7	0.39	–1.2	1.02	0.48	0.45	–1.14	1.48	–0.09	1.05
8	–0.01	–0.74	0.75	–0.31	–0.07	–0.12	0.26	0.3	0.48
9	0.41	–0.61	1.7	–1.05	0.11	–0.49	0.55	1.3	1.36

Note. Values shaded by intensity. Darker red signifies larger positive number, darker blue is a larger negative number.