Predicting Gentrification in Houston's Low- and Moderate-Income Neighborhoods
Francisca Winston and Chris Walker
November, 2012

City and community leaders in Houston have a strong interest in knowing which of the city's low- and moderate-income neighborhoods are at risk of gentrification, a process of economic change that results from the in-migration of households with higher education and income levels than the people who already reside there. If accompanied by rising property values, this process can lead to displacement of low-income residents who can no longer afford to pay increased rents or rising taxes on the homes they own.

To support local decisions on the allocation of scarce housing subsidies, LISC research staff developed a predictive model to show where gentrification in Houston is most likely to occur.

The attached map shows the low- and moderate-income census tracts that our statistical analysis predicts are most likely to display signs of gentrification over the next five-to-ten years. The map distinguishes among tracts that are more than 75 percent likely to gentrify (dark orange), those that have a 51-75 percent chance (light orange), and those with a less than 50 percent chance of gentrification (no color).

In this analysis, *gentrification* refers to increases in a neighborhood's median incomes, median housing values, and percentage of residents with a college degree that are at least 10 percent higher than for all Houston neighborhoods. This does not mean that neighborhoods will be fully "gentrified" in the near future; only that its trends are consistent with those normally associated with neighborhoods that do so.

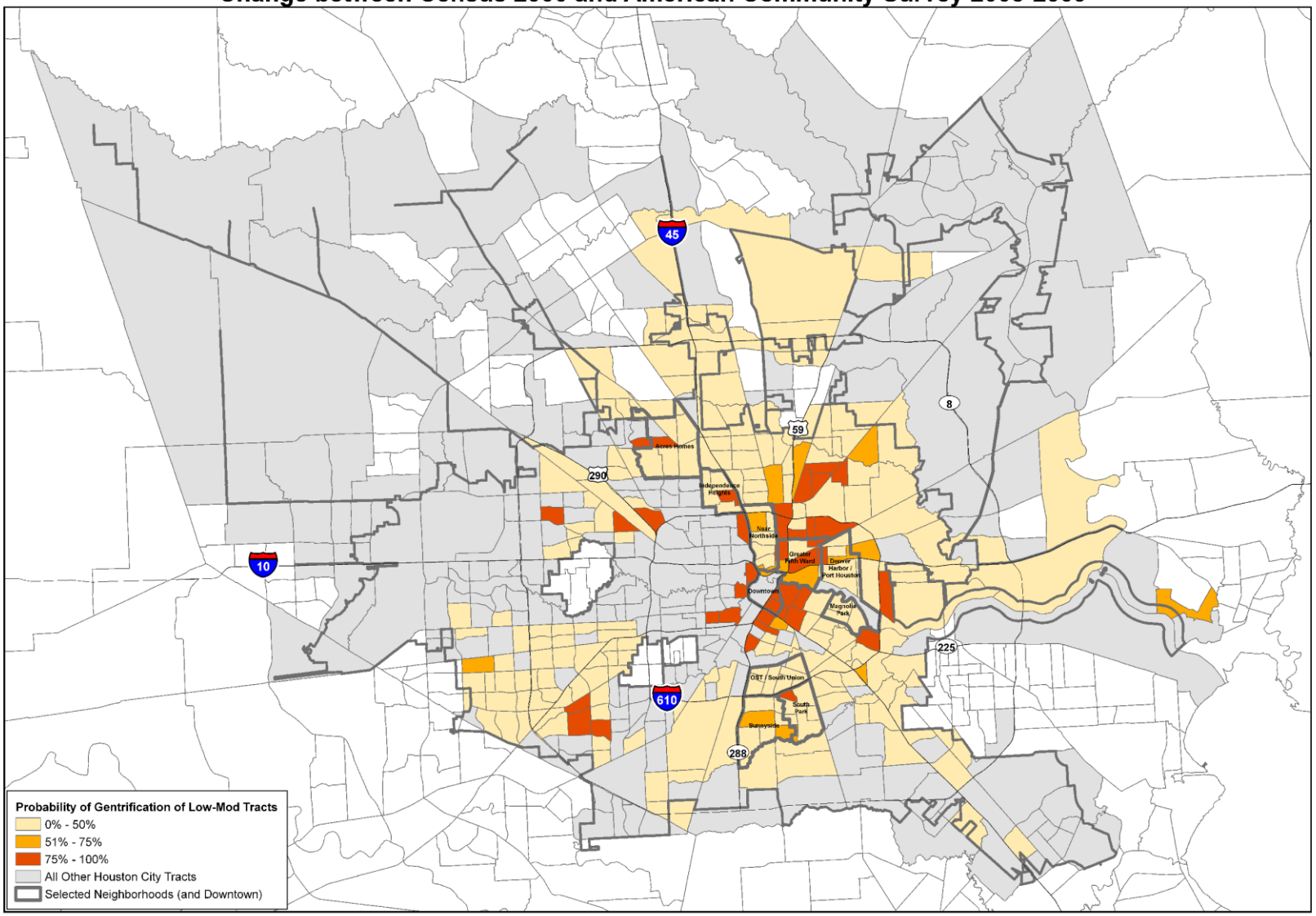
These projections are the result of a statistical procedure that takes factors known to be tied to past gentrification in Houston, and uses these same factors measured over a more recent period (roughly 2000-2007) to project where neighborhoods might gentrify in the future. The accuracy of the projection depends on whether these past factors will continue to influence change going forward.

The remainder of this short paper goes into more detail on the purposes, methods, data, and results of our analysis.

Analysis Approach

Our analysis follows an approach to predicting gentrification in the San Francisco Bay area developed by Karen Chapple in *Mapping Susceptibility to Gentrification: the Early Warning Toolkit.* Following past research, Chapple defined gentrification as "as a process of neighborhood change that encompasses economic change in the form of increases in both real estate investment and household income, as well as demographic change in the form of increases in educational attainment." To find tracts in the process of gentrification, Chapple identified low- and moderate-income census tracts that displayed an increase in income, education and house value greater than the San Francisco metropolitan area average between 1990 and 2000.

Probability of Gentrification in Low and Moderate Income Census Tracts in Houston, TX
Change between Census 2000 and American Community Survey 2005-2009



LISC Data Sources: 2000 US Census, US Census 2005-2009 American Community Survey 5-Year Estimates
Map created by LISC Research & Assessment

We adopted Chapple's core definition of gentrification, but made three modest changes. First, we stipulated that values for each variable used to designate gentrified tracts must increase by 10 percent or more compared to the city average. Second, we used the city as a point of comparison, not the overall metropolitan area. Third, we extended the time period to run from 1990 to roughly 2007, which is the mid-point of the 2005-2009 average reported in the US Census American Communities Survey, or ACS.

We also adopted Chapple's basic analysis approach, which used a statistical model known as binary logistical regression (logit) to "predict" gentrification between 1990 and 2000 using factors measured in 1990. She then used these same factors measured in 2000 to predict gentrification in 2010. (At the time of her analysis, she did not have 2010 census data available to her, as we do, nor had the 2005-2009 ACS data been released.) In our analysis, we use Census 1990 and ACS 2005-2009 data to measure gentrification between 1990 and 2007. We then predict gentrification between 1990-2007 using factors measured at one point in time (1990) as well as over time (1990-2000).

Variables and Data Sources

Drawing upon the literature on gentrification, Chapple tested a wide range of independent variables to determine which types of neighborhoods were more likely to gentrify. We began with her list of variables, selected those that were available to our analysis, and added others that were consistent with the gentrification literature. For example, Chapple included the percent of nonfamily households in 1990; we included both the 1990 percentage and the percent change between 1990 and 2000. We also added some additional baseline (1990) variables, including poverty rate and vacancy rate. Additionally, we examined possible amenity and disamenity factors, such as industrial land uses, as indicated by Harris County parcel level data for the city of Houston.

Table 1 compares the factors Chapple found to be statistically associated with gentrification to the factors tested in our model. The first three columns in the table provide (1) the variables used in Chapple's model, (2) the direction of each variable's effect on gentrification – positive or negative, and (3) its rank of importance in that model. (The greater the effect, the higher the ranking, with '1" being the most influential and "19" being the least influential.) The last three columns show (4) the variables used in our model, (5) the direction of each variable's influence, and (6) whether it was significant or not.

The rows in the table are grouped according to whether the variables were statistically significant in both the Chapple and LISC models, otherwise used in both models, were significant only in the Chapple model, or used only in our model, so far as we know. (We did not have the list of variables that were used in the Chapple model that turned out not to be statistically significant predictors.)

A variable is *significant* if its relationship with the dependent outcome is very unlikely to be the result of chance occurrence. In this case, a significant variable means it is correlated with the probability of gentrification and that relationship is not likely to be coincidental.

Table 1: Factors tested in Chapple's model and our LISC model						
	Chapple Variables	**Chapple Direction**	**Chapple Rank**	**LISC Variables**	**LISC Direction**	**LISC Significance**
Shared Significant Variables	Percent of non-family households	Positive	8	Percent of non-family households	Positive	Yes
	Percent of dwelling units in buildings with 5+ units	Positive	7	Percent of dwelling units in buildings with 5+ units	Negative	Yes
	Percent of dwelling units with three of more cars available	Negative	2	Percent of dwelling units with three of more cars available	Negative	Yes
	Youth Facilities Per 1000	Positive	3	Number of Youth Facilities	Positive	Yes
Other Shared Variables	Median gross rent	Negative	18	Median gross rent		No
	Percent non-Hispanic white	Negative	12	Percent non-Hispanic white		No
	Percent of dwelling units in buildings with 3-4 units	Positive	10	Percent of dwelling units in buildings with 3-4 units		No
	Percent of owners paying more than 35% of income	Negative	15	Percent of owners paying more than 35% of income		No
	Percent of married couples with children	Negative	9	Percent of married couples with children		No
	Percent of renters paying more than 35% of income	Positive	11	Percent of renters paying more than 35% of income		No
	Percent of workers taking transit	Positive	4	Percent of workers taking transit		No
	Percent renter-occupied	Positive	13	Percent renter-occupied		No
	Public housing units	Positive	19	Public housing units		No
	Recreational facilities per 1000	Negative	1	Recreational facilities per 1000		No
	Small parks per 1000	Positive	17	Small parks per 1000		No
KC's Exclusive Variables	Public Space per 1000	Positive	5			
	Income diversity	Positive	6			
	Distance to San Jose	Positive	14			
	Distance to San Francisco	Negative	16			
LISC's Exclusive Variables				Change In Percent Of Married Couples with Children 1990 – 2000	Negative	Yes
				Change In Percent Of Non-Family Households 1990 – 2000	Positive	Yes
				Change In Percent Of Renter-Occupied Units 1990 – 2000	Negative	Yes
				Change In Median Gross Rent 1990 – 2000		No
				Change In Percent Of Non-Hispanic Black 1990 – 2000		No
				Change In Percent Of Non-Hispanic White 1990 – 2000		No
				Change In Percent Of Owners Paying More Than 35% Of Income 1990 – 2000		No
				Change In Percent Of Renters Paying More Than 35% Of Income 1990 – 2000		No
				Change In Percent Of Units In Buildings With 3 To 4 Units 1990 – 2000		No
				Change In Percent Of Units In Buildings With 5 Or More Units 1990 – 2000		No
				Change In Percent Of Units With 3 Or More Vehicles 1990 – 2000		No
				Change In Percent Of Workers Taking Transit 1990 – 2000		No
				Change In Vacancy Rate 1990 – 2000		No
				Percent Acreage General Commercial Vacant		No
				Percent Acreage Residential Vacant Table Value		No
				Percent Vacant 1990		No
				Poverty Rate 1990		No

Columns two and five indicate the direction of each variable's effect, either positive or negative, on the probability of gentrification. For example, "Percent of nonfamily households" exerts a positive effect in both Chapple's model and ours. This means, holding all other variables constant, that areas with higher percentages of nonfamily households have greater probability of gentrification. Conversely, a variable with a negative direction has the opposite effect; areas with higher values have lower probability of gentrification.

To construct the logit model, we used forward selection based on maximum likelihood. Forward selection is stepwise selection method where each variable is tested independently against the outcome variable, which in this case, is whether the tract gentrified. The variable with the strongest relationship – the most significant variable – is the first factor added to the model. Each remaining variable is then tested in a model with the first selected factor. The most significant remaining variable is selected and added to the model. This continues until there are no more factors to add using a threshold of .05 significance level; that is, less than a 5 percent likelihood that the relationship is coincidental.

The seven factors we tested that turned out to be statistically significant predictors to gentrification between 1990 and 2007 are given in Table 2.

Table 2: Factors included on our model

Percent of non-family households 1990
Percent of dwelling units in buildings with 5+ units 1990
Percent of dwelling units with three of more cars available 1990
Number of Youth Facilities
Change In Percent Of Married Couples with Children 1990 – 2000
Change In Percent Of Non-Family Households 1990 – 2000
Change In Percent Of Renter-Occupied Units 1990 – 2000

Analysis Results

As noted above, the analysis proceeds in two steps. In Step 1, we develop a model that predicts gentrification using factors thought to be associated with that process. This step produces a list of variables measured in 1990 and in 2000 and associated coefficients that together "predict" whether a tract gentrifies in 2007. (This prediction can be examined for accuracy.) In Step 2, we take the variables as measured in 2000 and 2007 and predict gentrification in the future.

Step 1: Predicting gentrification between 1990 and 2007

To predict gentrification in Houston between 1990 and 2007, we used a logistic regression, which enables researchers to examine the effects of multiple "predictor" variables on an outcome, in this case gentrification, which is measured as a simple yes-or-no. That is, a tract is either "undergoing gentrification" or "not undergoing gentrification."

The basic formula for a logit model is: logit (g) = a + bx, where *g* is the probability of the outcome variable, *a* is a constant and *b* is the coefficient for the explanatory variable *x*. [1]

[1] Logistic regression
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1065119/

Using the coefficients from the output table below, our model with multiple variables and no constant is:

logit (probability of gentrification) =
6.2049*Percent of non-family households 1990 –
6.5605*Percent of dwelling units in buildings with 5+ units 1990 –
21.7092*Percent of dwelling units with three of more cars available 1990 +
3.5889*Number of Youth Facilities –
2.1560*Change In Percent Of Married Couples with Children 1990 – 2000 +
2.8821*Change In Percent Of Non-Family Households 1990 – 2000 –
9.1538*Change In Percent Of Renter-Occupied Units 1990 – 2000

In ordinary least squares (OLS) regression, which is what most people think of as "regression," a statistic called the R-square provides a measure of the model's goodness of fit. It ranges from zero to one; a larger R-square means the model does a better job at explaining the variability in the outcome variable. In logistical regression, the OLS R-Square calculation can't be calculated. To measure the strength of a model, *Pseudo R-squares* are often used. They look like OLS R-squares, but do not mean exactly the same thing. [2]

The Nagelkerke R-square is one of many *pseudo R-squares.* It measures how useful the independent factors in our model are at predicting the outcome variable, the probability of gentrification. The Nagelkerke R-square values can range from zero to one with higher values meaning a more useful model. [3] Our overall model produced a Nagelkerke R-square of .715, indicating a reasonably good

Table 3: Output of the model

Variables in the Equation	Coefficient	Standard Error	Wald	Significance
Percent of non-family households 1990	6.20	1.73	12.84	0.0003
Percent of dwelling units in buildings with 5+ units 1990	-6.56	1.64	15.81	0.0001
Percent of dwelling units with 3or more cars 1990	-21.70	5.08	18.22	0.0000
Number of Youth Facilities	3.58	1.49	5.74	0.0165
Pct. change in married couples with children 1990 – 2000	-2.15	0.87	6.08	0.0136
Pct change in non-family households 1990 – 2000	2.88	1.40	4.21	0.0400
Percent change renter-occupied units 1990 – 2000	-9.15	2.66	11.76	0.0006

The Wald Statistic is used to measure each variable's significance. It is equal to (coefficient/Standard Error)^2 and follows a Chi-Square distribution.

Researchers and policymakers are interested in two types of prediction accuracy. Does a tract that is predicted to gentrify do so or not? And do tracts that are not predicted to gentrify do

[2] What are pseudo R-squareds? UCLA: Statistical Consulting Group.
http://www.ats.ucla.edu/stat/mult_pkg/faq/general/Psuedo_RSquareds.htm
[3] Logistic regression
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1065119/

so? As explained above, the binary logit model assigns each tract a probability that it will gentrify. By comparing predicted gentrification with its actual occurrence, researchers can test the accuracy of the model.

In this step, our model assigned a likelihood of gentrification to all 195 low and moderate Houston income tracts in 1990. In 29 tracts, this likelihood was greater than 50 percent. Of these, 21 tracts (or 72 percent) actually gentrified, by our definition. The model assigned a 75 percent or more likelihood of gentrification to 14 of the 29 tracts; of these, 12 actually gentrified, a "success rate" of 86 percent. In other words, as the model assigns higher likelihoods to tracts, the accuracy of the prediction goes up. Conversely, the model predicted that 164 would not gentrify and 144, or 88%, did not gentrify, which means that 20 tracts gentrified even though they were not predicted to do so. Table 4 provides the tract counts and percentages of correct and incorrect predictions.

Table 4: Accuracy of predictions of gentrification between 1990 and 2000

Predicted Likelihood of Gentrification by 2000	**Did Tract Gentrify Between 1990 and 2000?**		**Total**	**Percent Correct**
	No	*Yes*		
Less than 50%	144	20	164	88%
50% to 75%	6	9	15	60%
Greater than 75%	2	12	14	86%
Undefined	1	1	2	
Total	153	42	195	
Percent Correct	94%	50%		

Step 2: Predicting Gentrification in 2015-2020

To forecast gentrification in the future, we take the coefficients from the model in step one and apply them to 2000-2007 data. To illustrate the relationships in the model between predictive factors and gentrification likelihood, Table 5 shows the average values of the predictor variables for each of three categories of gentrification likelihood.

Table 5: Average values of the seven significant factors by likelihood of gentrification

Predicted Likelihood of Gentrification	**Percent of non-family households 2000**	**Percent of dwelling units in buildings with 5+ units 2000**	**Percent of dwelling units with three of more cars available 2000**	**Youth Facilities**	**Change In % Of Married Couples with Children 2000-2007**	**Change In % Of Non-Family Households 2000-2007**	**Change In % Of Renter-Occupied Units 2000-2007**
Less than 50%	29%	37%	11%	0.0146	-11%	8%	5%
50% to 75%	27%	17%	10%	0.1538	-24%	14%	-6%
Greater than 75%	36%	23%	8%	0.4118	-29%	25%	-10%
All Tracts	30%	34%	10%	0.0748	-14%	11%	2%

For example, the next-to-last column in the table shows the percent change in non-family households between 2000 and 2007 for all tracts and for tracts in each likelihood category. Tracts with a less-than-fifty-percent likelihood of gentrifying had an only 8 percent increase in non-family households compared to a 25 percent increase for those tracts with a change of gentrifying of more than 75 percent.

The model predicts that of the 254 low and moderate income tracts in the city of Houston in 2000, 47 (or 19 percent) have a likelihood of gentrifying of 50 percent or more. If we consider only those with more than a 75 percent chance of gentrifying, which is what we recommend, the model predicts that 34 tracts will gentrify. Strictly speaking, the model predicts the likelihood of gentrification in 2012, but as this is a continuing process, it is reasonable to suggest a longer time horizon, such as sometime between 2015 and 2020. Tract counts and percentages are given in the table below.

Table 6: Number of tracts predicted to gentrify in future

Predicted Likelihood of Gentrification by 2000	Number of Tracts	Percent
Less than 50%	206	81%
50% to 75%	13	5%
Greater than 75%	34	13%
Undefined	1	0%
Total	**254**	100

So if trends continue as they have between 2000 and ACS 2005-2009 (roughly 2007), the map shows tracts where we expect to see above average performance on education, income, property values relative to other neighborhoods sometime in the near future.

Predictions, of course, can be wrong. If we assume that the prediction going forward is as accurate as the predictions made in step 1, then of the 34 tracts with a greater than 75 percent likelihood of gentrifying, 3 of them will fail to do so. (That is, 88 percent of them will gentrify, which is the same ratio as in Table 1.)