INSTITUTE FOR
HOUSING STUDIES AT
DEPAUL UNIVERSITY

IHS Regional Housing Market Segmentation Analysis

June 2016

Using clustering techniques and data on current housing market and socioeconomic conditions, IHS developed a segmentation analysis to help regional policy makers and housing stakeholders understand variation in local housing submarkets. These data and analysis were developed to allow for more strategic and targeted outreach and implementation of the Homes for a Changing Region program and more broadly, to inform the development of new regional housing initiatives that are responsive to local needs.

Table of Contents

ABBREVIATED METHODOLOGY

About the technique

When developing housing and community development strategies, policy makers, and urban planners rely on a deep understanding of the characteristics of the communities in which they work. While it is simple to evaluate a community's median income or unemployment rate, it becomes challenging to characterize communities when comparing many factors, such as educational attainment, population changes, levels of mortgage credit, or the underlying housing stock.

A housing market segmentation study can explore differences by building an integrated model that identifies geographic units with similar characteristics. This data modeling approach can be invaluable for urban planning, real estate development, geo-demographic research, legislative policy work, economic investment, and more. An extensive history of housing segmentation and its applications can be found in *Housing Market Segmentation: a Review* by Islam and Asami, and *The definition and identification of housing submarkets* written by C.A. Watkins.[1,2]

Clustering algorithms incorporate multiple variables to group items by their overall similarity. When applied to community-level data, clustering provides a way to compare different geographic units that share similar traits, regardless of their physical proximity. The goal of this study was to use clustering techniques to perform a housing market segmentation analysis of the seven county region surrounding Chicago, IL namely Cook, DuPage, Kane, Kendall, Lake, McHenry, and Will counties. The model incorporated data on housing stock and affordability, housing market activity, resident demographics, and socioeconomic indicators to identify communities with similar characteristics.

The segmentation algorithm and associated techniques are discussed in greater detail in the technical appendix.

About the data

Data were collected from five sources: the 2000 US Decennial Census; 2013 5-year sample estimates from the US Census Bureau's American Community Survey (ACS); the US 2010 Project at Brown University; Housing and Urban Development's 2013 Location Affordability Index; and housing market conditions data from the Institute for Housing Studies at DePaul University. The communities studied comprise a seven-county region in northeastern Illinois, and include Cook, DuPage, Kane, Kendall, Lake, McHenry, and Will

[1] K. S. Islam and Y. Asami, "Housing Market Segmentation : a Review," Rev. Urban Reg. Dev. Stud., vol. 21, no. 2, 2009.

[2] C. A. Watkins, "The definition and identification of housing submarkets," Environ. Plan. A, vol. 33, no. 12, pp. 2235–2253, 2001.

counties. The region was studied at the census-tract level, with census tracts defined by the year 2010 Decennial Census; a total of 1985 census tracts were analyzed.[3]

Data for this study were collected to analyze a number of topics related to housing demand and supply, including variables associated with current and changing demographic and socioeconomic conditions, housing affordability, housing stock, and investment and market patterns. These data included information on population change, income level, household size, age, housing tenure, rents and home values, cost burden, vacancy, educational attainment, density, housing stock age and type, mortgage activity, foreclosure distress, and characteristics of property sales. By request, IHS excluded data on poverty status and race and ethnicity so that these features could be analyzed separately by the Chicago Metropolitan Agency for Planning (CMAP) and project partners. A full list of variables and sources is included in the digital appendix.

[3] See the technical note for more on the treatment of the data for analysis

CLUSTERING RESULTS

Overview of cluster patterns and maps

The clustering application identified eight distinct clusters in the Chicago seven county region. It clustered wealthy, economically distressed, and high-growth census tracts consistently and distinctly. Though geographic location was not included as a variable, the results also highlighted a strong geographic pattern related to the historical evolution of urban development outward from downtown Chicago based on age of the housing stock. It also identified a cluster unique in its high levels of growth, housing stock age, and income but geographically distinct, with tracts concentrated in downtown Chicago and dispersed in pockets across the region.

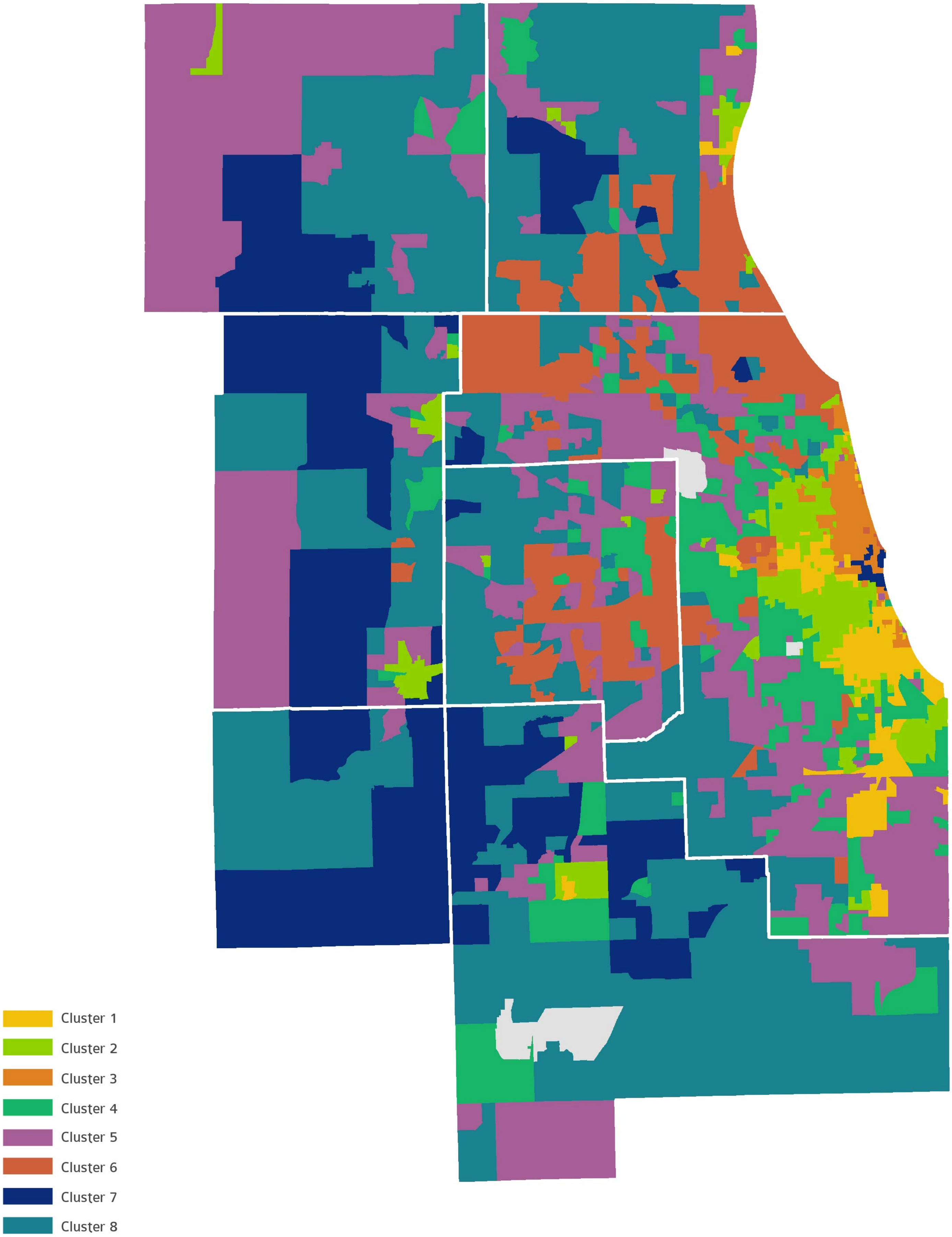
Clusters were most strongly differentiated by age of housing, household income, population growth, and economic hardship indicators such as unemployment and foreclosure. The algorithm identified two low- to moderate-income primarily suburban clusters differentiated by housing stock, an aging population, and certain economic indicators. The algorithm also identified two high-wealth clusters: one typified by middle-aged homeowners in communities with moderate sales activities, and a second cluster of young, highly educated, urban professionals. Finally, two economically-distressed clusters were identified, differentiated by the varying degrees of unemployment, subsidized housing, and investment indicators.

A summary of characteristics are in Table 1 below. A full summary table with value ranges for each cluster and data variable analyzed in the segmentation model are included in the digital appendix.

Table 1. IHS Regional Housing Market Segmentation Analysis, Summary of Cluster Characteristics

Cluster	Housing Affordability	Housing Stock	Investment and market conditions	Demographic and socioeconomic characteristics
1	•High and increasing levels of cost burden •Housing stock is low cost, but incomes are low and have seen large declines •Low transportation costs	•Relatively high density urban cluster with older housing stock built prior to 1940 •High levels of renters and a largest share of subsidized housing	High distress area with high levels of foreclosure activity and vacancy •Low levels of mortgage investment, but high levels of cash sales	•Low levels of educational attainment •Lower-income with the largest income declines •Unemployment is high and increasing •Increase in senior population •High, but decreasing levels of children
2	•High and increasing levels of cost burden, largely due to declining incomes •Low transportation costs	•Relatively high density urban and suburban cluster with older housing stock built prior to 1940 •Moderate to high levels of renters with lower levels of subsidized housing	•High distress area with higher levels of foreclosure activity but moderate levels of vacancy •Low levels of mortgage investment, but high levels of cash sales	•Low levels of educational attainment •Households have lower- and moderate-income and have seen declining incomes •This area is characterized by large households and high levels of children
3	•House prices and rents are high and increasing, but lower levels of cost burden due to high and growing incomes •Low transportation costs	•High density urban cluster with older housing stock •High level of renters, but only area with declining levels of renters •Generally lower levels of levels of subsidized housing	•Very active housing market with high levels of mortgage activity and turnover and low vacancy •Low foreclosure distress	•High/middle income, younger, and educated households •Only cluster to see incomes increase •High levels of small 1-person households with low levels of children
4	•Moderate cost burdened with substantial increases in burdened households •Housing costs are generally low, but incomes are declining •High transportation costs	•Largely suburban cluster with some lower density urban areas and largely post-war housing stock built 1940–59 •Largely owner occupied cluster with low levels of subsidized housing	•Moderate levels of foreclosure activity and high/moderate level of distressed sales •Moderate levels of mortgage lending and housing market activity	•Primarily middle and moderate income households with generally lower levels of educational attainment •Family area with moderate share of children but somewhat older population
5	•Moderate cost burdened with substantial increases in burdened households •Housing costs are generally low, but incomes declining •High transportation costs	•Lower-density suburban cluster with housing stock largely built 1960-79 •Largely owner occupied, but higher level of renters and subsidized housing	•Moderate levels of foreclosure activity and high/moderate level of distressed and cash sales •Only cluster with house price decline	•Primarily moderate income households with generally lower levels of educational attainment •Family area with moderate share of children
6	•Expensive cluster in terms of housing costs (both house prices and rents) •Low cost burden due to high incomes •High transportation costs	•Largely suburban, low density with a mix of both older and newer housing •Largely owner occupied with low levels of subsidized housing	•Not heavily impacted by foreclosures •Low vacancy with higher levels of mortgage investment, and higher levels recent housing turnover	•Largely higher income and highly educated •High share seniors and low share of younger adults •Family cluster with high share of 2-4 person households but moderate levels of children.
7	•This area has moderate levels of cost burden with slight increases •Home prices are moderate, but rents are high (core), and incomes are high •Transportation costs are very high (fringe areas)	•Bimodal cluster includes areas with high levels of recent, post-2000 development in the urban core and fringe •Outside of core, primarily low density •Lowest level of renters	•Moderate levels of foreclosure activity and distressed sales •Moderate levels of lending, somewhat stagnant recent market activity	•Cluster with most significant population growth •Higher/middle-income cluster, but only moderate levels of educational attainment •Large family households and high share of children, likely at fringe •Primarily middle-aged households
8	•Moderate levels of cost burden with slight increases •Home prices and rents are moderate due to higher incomes •Very high transportation costs	•Suburban, low density cluster with a housing largely built after 1980 •Largely owner occupied cluster with low levels of subsidized housing	•Not heavily impacted by foreclosures but moderate levels of distressed sales •Low vacancy with moderate levels of mortgage investment	•Moderate levels of population growth •Higher/middle-income cluster, but only moderate levels of educational attainment •High share of 2-4 person households but large decline in children •Primarily middle-aged households

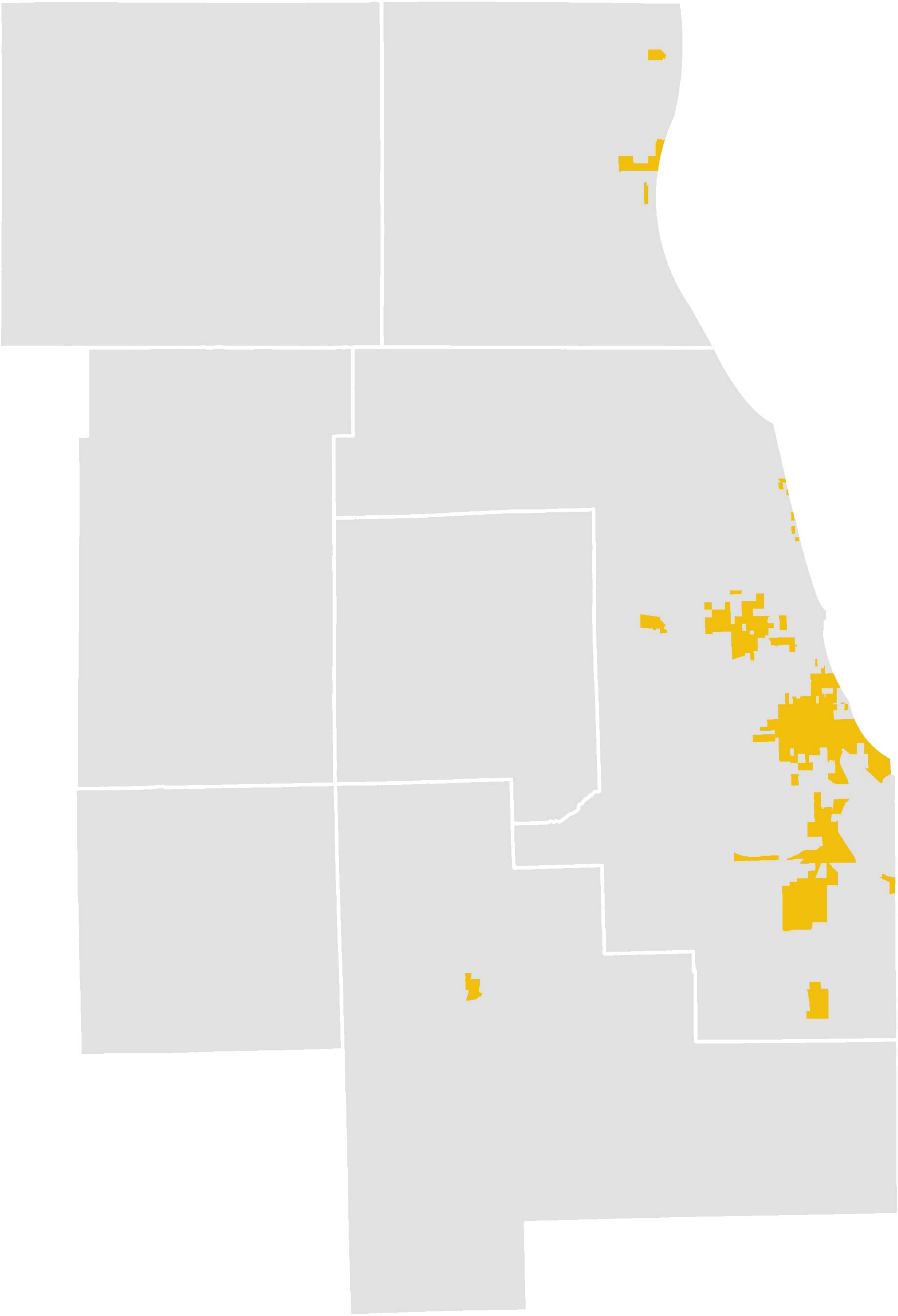
IHS Regional Housing Market Segmentation Analysis
Chicago Region



SOURCE: IHS CALCULATIONS

INSTITUTE FOR HOUSING STUDIES
AT DEPAUL UNIVERSITY

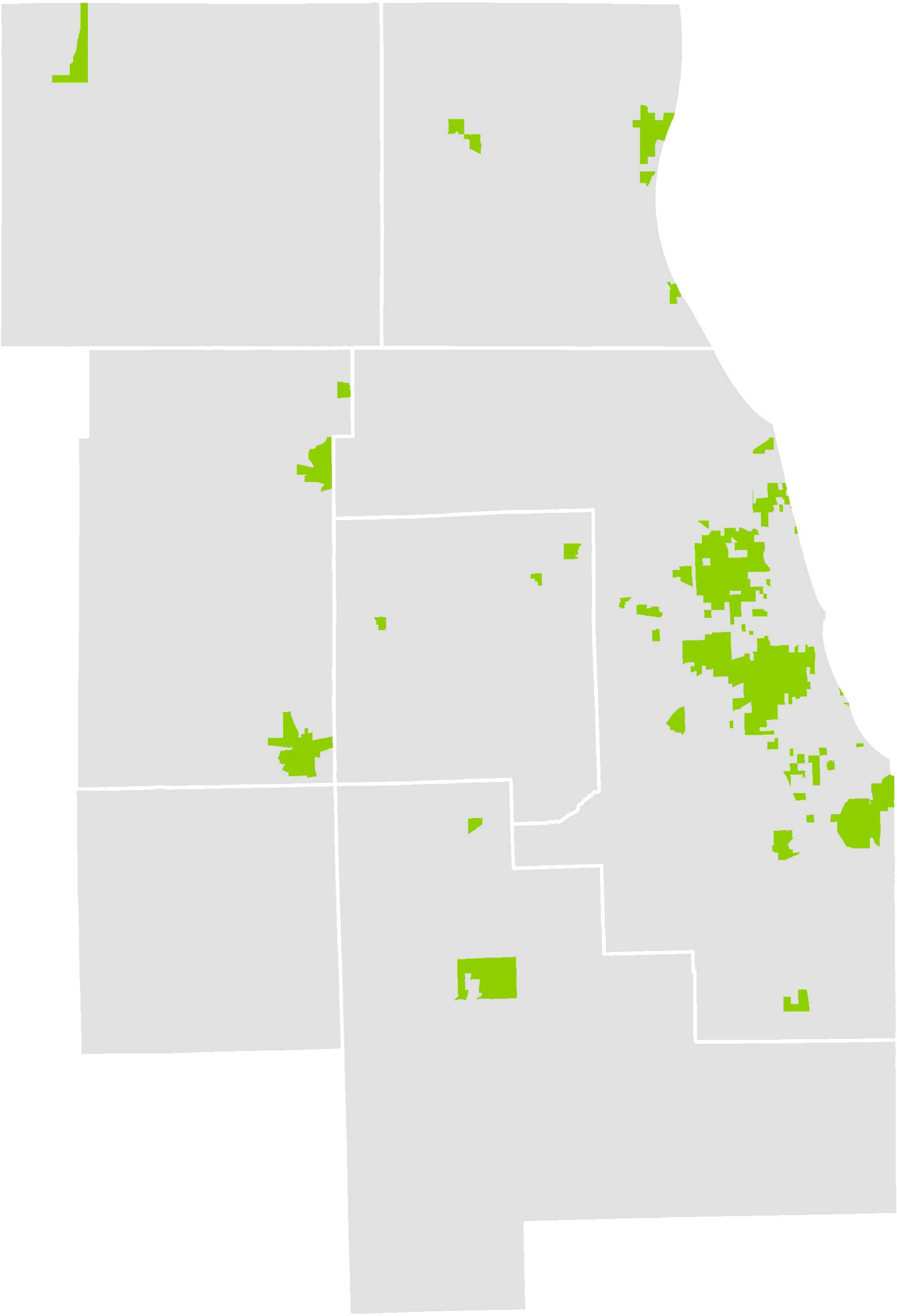
IHS Regional Housing Market Segmentation Analysis
Chicago Region, Cluster 1



SOURCE: IHS CALCULATIONS

INSTITUTE FOR **HOUSING STUDIES**
AT **DEPAUL UNIVERSITY**

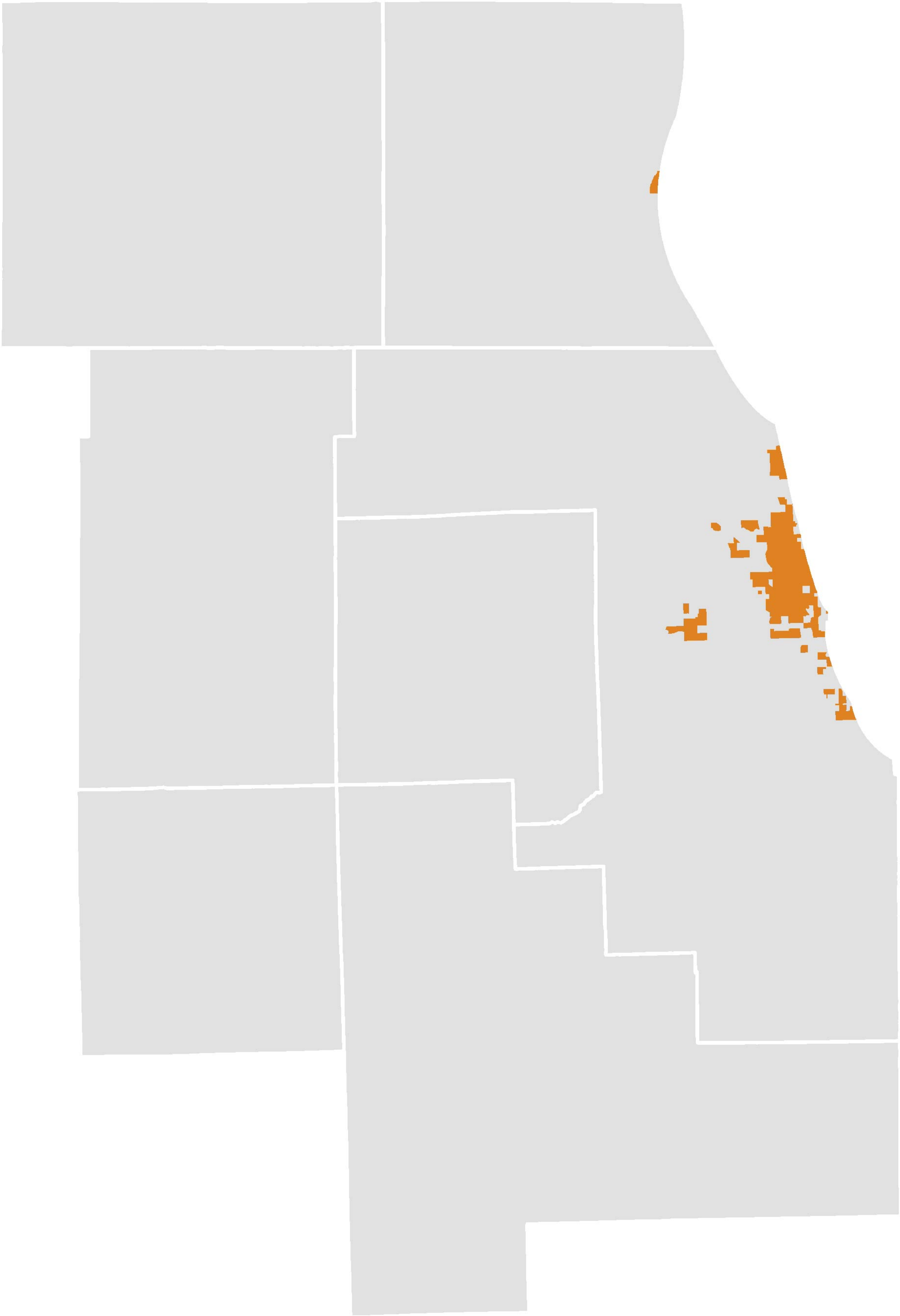
IHS Regional Housing Market Segmentation Analysis
Chicago Region, Cluster 2



SOURCE: IHS CALCULATIONS

INSTITUTE FOR **HOUSING STUDIES**
AT **DEPAUL UNIVERSITY**

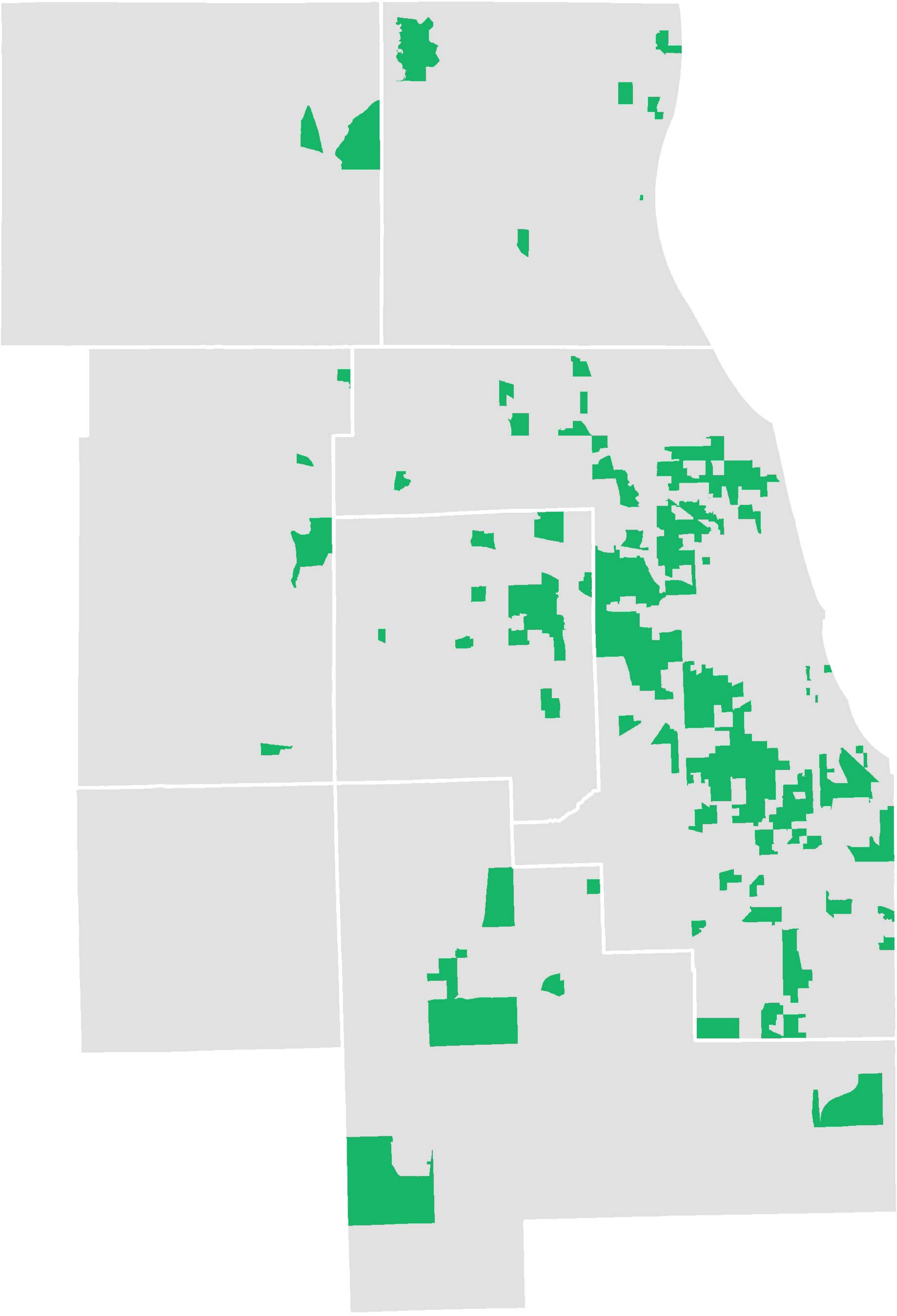
IHS Regional Housing Market Segmentation Analysis
Chicago Region, Cluster 3



SOURCE: IHS CALCULATIONS

INSTITUTE FOR HOUSING STUDIES
AT DEPAUL UNIVERSITY

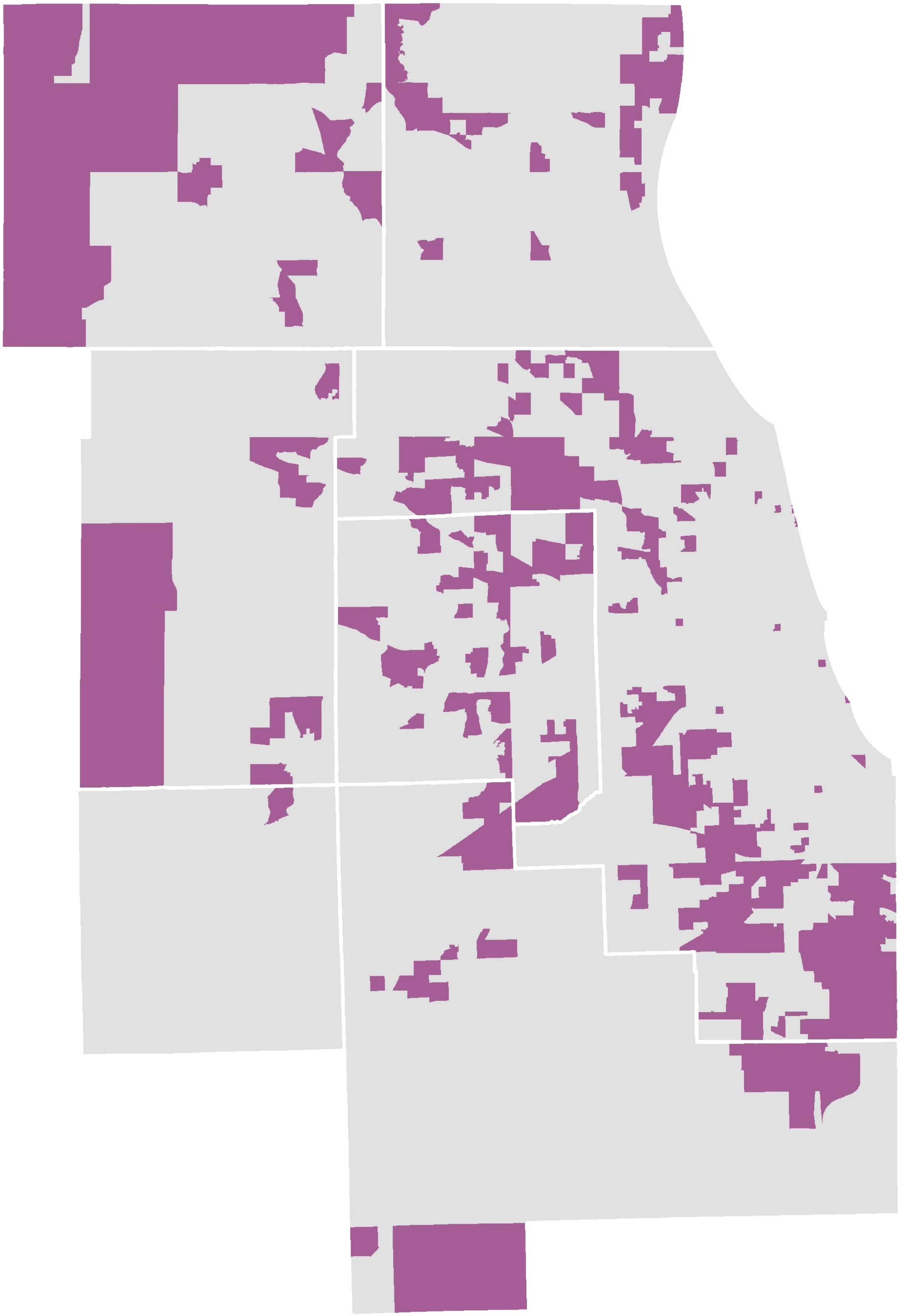
IHS Regional Housing Market Segmentation Analysis
Chicago Region, Cluster 4



SOURCE: IHS CALCULATIONS

INSTITUTE FOR HOUSING STUDIES
AT DEPAUL UNIVERSITY

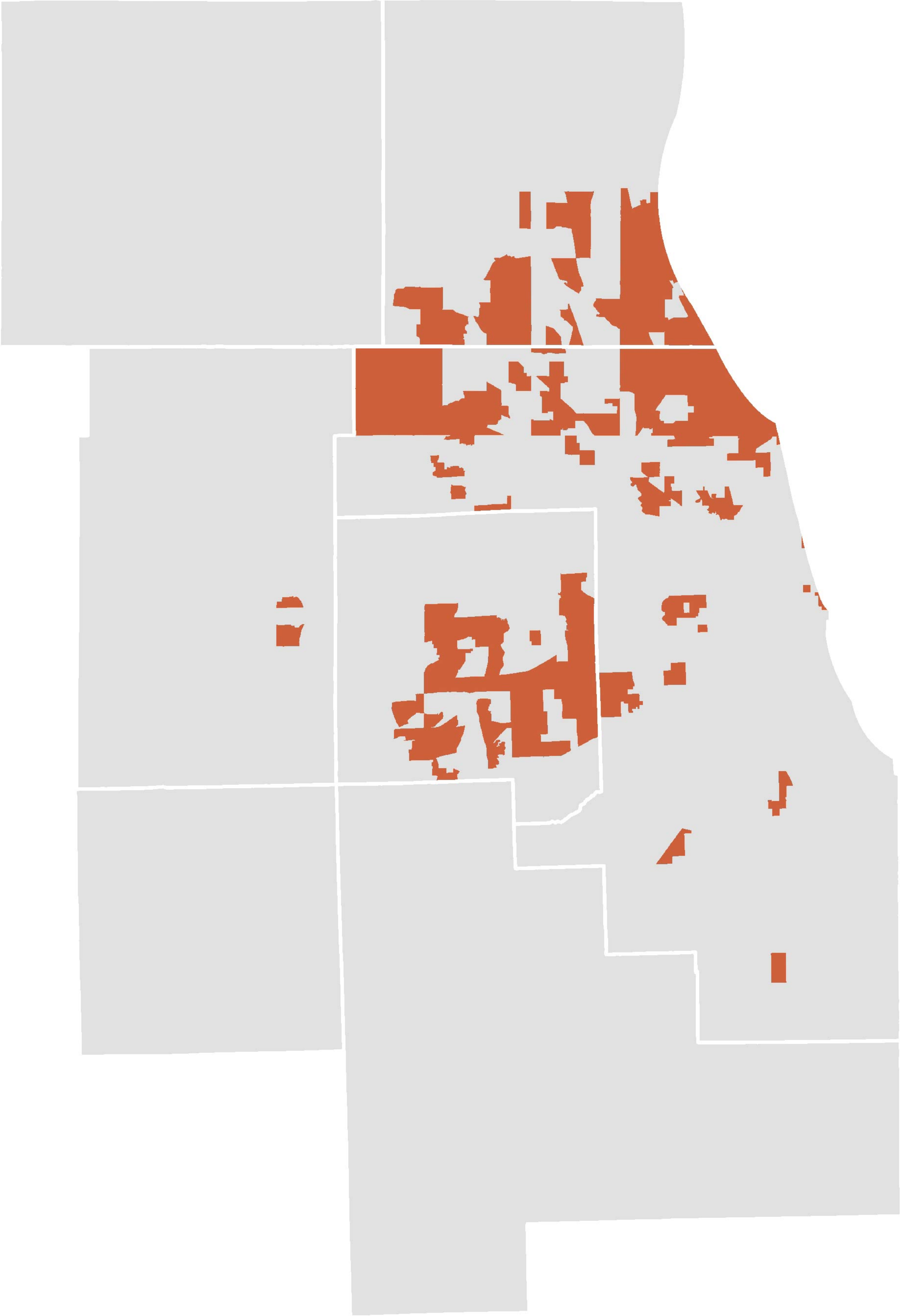
IHS Regional Housing Market Segmentation Analysis
Chicago Region, Cluster 5



SOURCE: IHS CALCULATIONS

INSTITUTE FOR HOUSING STUDIES
AT DEPAUL UNIVERSITY

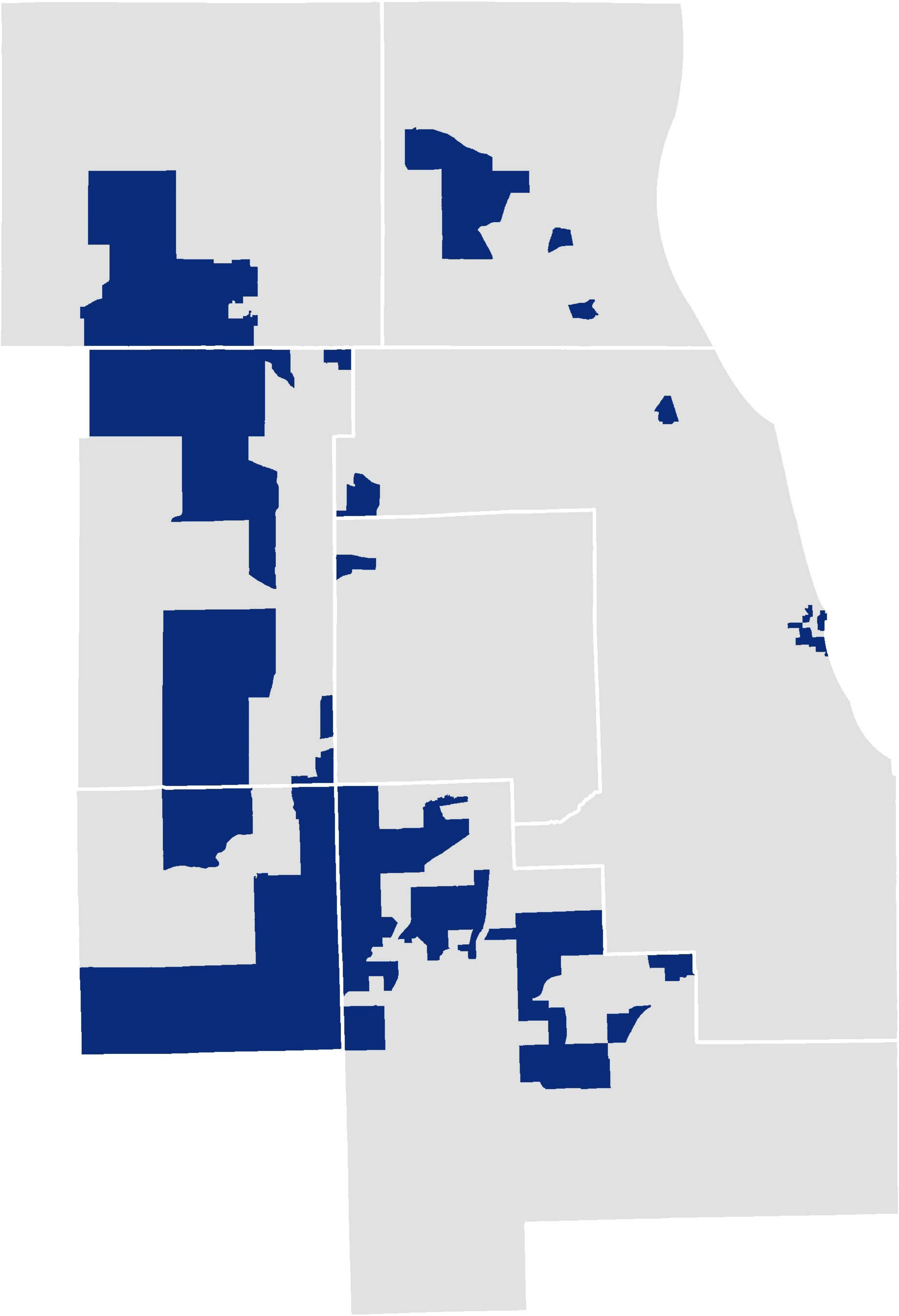
IHS Regional Housing Market Segmentation Analysis
Chicago Region, Cluster 6



SOURCE: IHS CALCULATIONS

INSTITUTE FOR HOUSING STUDIES
AT DEPAUL UNIVERSITY

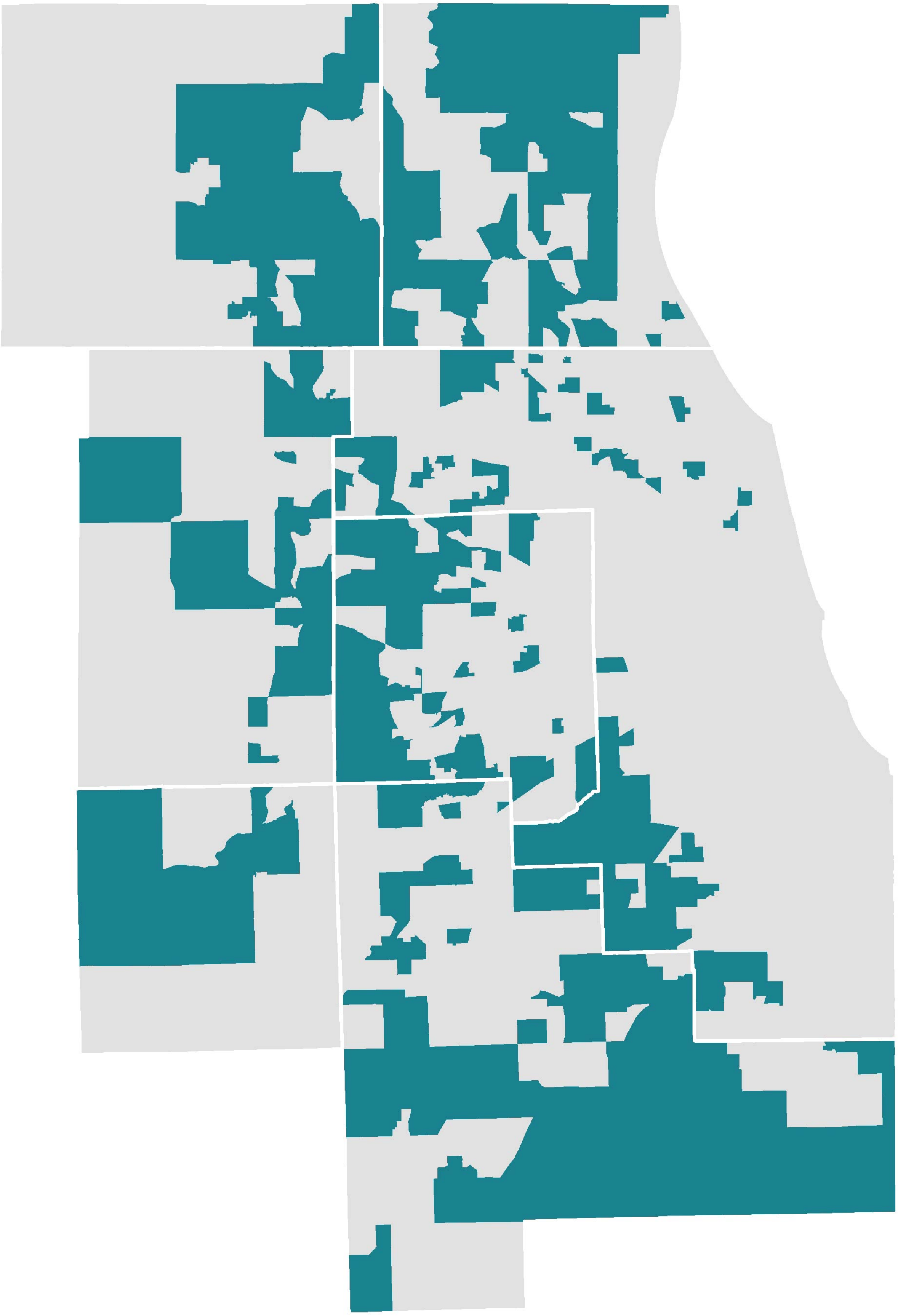
IHS Regional Housing Market Segmentation Analysis
Chicago Region, Cluster 7



SOURCE: IHS CALCULATIONS

INSTITUTE FOR HOUSING STUDIES
AT DEPAUL UNIVERSITY

IHS Regional Housing Market Segmentation Analysis
Chicago Region, Cluster 8



SOURCE: IHS CALCULATIONS

INSTITUTE FOR **HOUSING STUDIES**
AT **DEPAUL UNIVERSITY**

TECHNICAL APPENDIX

Data preprocessing

Normalization of census tracts for longitudinal analysis

Since 2000 Census tracts differ from 2010 tracts, IHS was required to normalize year 2000 data to 2010 census boundaries. For certain data, IHS utilized transformations of Census data from Brown University's US 2010 Project. Where data were not available, IHS utilized an application developed by Brown University to normalize the results.[4] After removing five census tracts that did not represent residential areas, 1,980 census tracts were included in the analysis.[5]

Normalization of original data values

In clustering applications, a typical preprocessing step is to standardize variables so that all data are transformed to a comparable range of value. This is because variables measured at different scales will likely skew an analysis, where a variable with a larger range might outweigh variables with smaller ranges. To correct this, the following transformations were applied:

- Each count variable was converted to percentages ranging in the [0,1] interval.
- Each continuous variable with a dollar amount such as median household income, home value, or contract rent was converted to a new variable in the [0,1] range using a Min-Max Scaling. Year 2000 variables were also adjusted for inflation.[6]
- Variables describing changes between year 2013 and year 2000 were computed by subtracting the percentage values in 2013 from the percentage values in 2000 and/or subtracting the inflation adjusted year 2000 amounts from year 2013 amounts.

Margin of error analysis

The Census ACS estimates are based on a sample and as a result, may be affected by high levels of sampling variability. The reliability of each ACS estimate can be analyzed using the published margin of error that is based on a 90-percent confidence level.[7] The margin of error (MOEs) measures the variation in the random samples due to chance.

[4] See (http://www.s4.brown.edu/us2010/Researcher/ltdb1.htm)

[5] Two tracts, 17031980100 and 17031980000, represent Midway and O'Hare airport, respectively. Tract 17197980000 is an Army Munitions. Tract 17031381700 and 17097863006 both have zero population.

[6] The dollar values of Year 2000 variables were converted to a 2013 Inflation Adjusted amount by multiplying them by a factor of 1.37811525, obtained from the Bureau of Labor Statistics.

[7] U.S. Census Bureau (2008). A Compass for Understanding and Using American Community Survey Data: What General Data Users Need to Know. Washington DC: U.S. Government Printing Office https://www.census.gov/content/dam/Census/library/publications/2008/acs/ACSGeneralHandbook.pdf

Method for standard errors

A commonly used technique to decide whether a certain ACS variable estimate is reliable employs the coefficient of variation (CV) of the sample estimate. The coefficient of variation is defined as the ratio between standard error and estimated value, and measures the relative amount of variability associated with the sample estimate. Low CV values indicate more reliable estimates. In line with these criterion, only ACS estimates with CV values below 30 percent were used in this analysis. In order to include certain ACS variables with CV values exceeding 30 percent, IHS followed Census Bureau protocols to create a new derived variable with a reduced and acceptable margin of error. The CV of the aggregated estimate was computed to assess its reliability and the new aggregated variable was used in the analysis if the CV was below 30 percent.

K-Medoids clustering technique

About the method

This analysis uses a K-Medoid technique as the method for defining clusters of census tracts with similar characteristics. K-Medoid is a distance-based partitioning method that divides the set of data points into non-overlapping subsets (or clusters) such that each data point is exactly in one subset. Objects within a subset are more similar to one another and different from the objects in other clusters.

The K-Medoid technique groups data points by calculating their pairwise distance from a central point in each cluster. The central-most point (medoid) of the cluster can be regarded as the representation of that cluster. Each data point is then assigned to the closest medoid, and the collection of points assigned to a medoid forms the associated cluster. Extensive discussion of this technique can be found in the textbook by L. Kaufman and P.J. Rousseeuw.[8] K-medoid clustering was chosen for this analysis as it can better handle the variation and outliers present in housing data utilized for the study.[9] The analysis was computed using the PAM implementation in the R "cluster" package. Eight clusters were created using the Euclidian distance measure.

Choosing the number of clusters "k"

One major challenge among clustering methodologies is the need to pre-select an appropriate number (k) of clusters. The intended use of the final clustering results can cause additional complexity. If there are too few clusters the segmentation is coarse, resulting in broad, non-specific clusters. With too many clusters, there are very small differences among variables, and it becomes difficult to characterize the clusters. One common quantitative approach to choosing the appropriate number of clusters is to cluster the data multiple times and choose a different number of clusters each time. An internal validity metric measuring

[8] L. Kaufman and P. J. Rousseeuw, Finding groups in data : an introduction to cluster analysis. Wiley, 2005.
[9] H. S. Park and C. H. Jun, "A simple and fast algorithm for K-medoids clustering," Expert Syst. Appl., vol. 36, no. 2 PART 2, pp. 3336–3341, 2009.

the quality of the clustering results is recorded for each trial and the optimal k is selected according to some criterion specific to the chosen metric.

Methods for evaluating cluster quality

Silhouette width is a common internal validity measure for clustering and has been shown to be robust when applied to many clustering algorithms.[10] To choose the appropriate number of clusters for this study, silhouette width was recorded for values of k from two to eleven. Eight clusters were selected as they were associated with both a narrow silhouette width and an acceptable level of granularity for the intended use.

Qualitative testing

Clustering seeks to create useful, understandable, and insightful groupings. Considering these goals, qualitative evaluations of cluster quality are also relevant. For this study, mapping and evaluation of geographic patterns and trends verified that the algorithms produced clusters with merit by assessing whether clusters made sense intuitively and accurately reflected the observed characteristics of the areas.

Clustering results were analyzed using a three-step process. First, silhouette distances were computed as a quantitative assessment of clusters quality. Census tracts and associated clusters were then mapped to determine whether the results were consistent with the observed characteristics in the region. Finally, the values for each variable included in the segmentation were compared among clusters to identify if significant differences among clusters and to descriptively characterize each cluster. The results were further refined through meetings with project partners, resulting in the final housing market segmentation results presented in this document.

[10] L. Vendramin, R. Campello, and E. Hruschka, "Relative Clustering Validity Criteria: A Comparative Overview," Stat. Anal. Data Min., vol. 3, no. 5, 2010.