From Individual Tracts to Community Segments: an Unsupervised Learning Approach for the Chicago Community Areas

Zahra Ferdowsi, Raffaella Settimi, Daniela Raicu, and Winifred Curran
School of Computing, CDM,
DePaul University
Chicago, IL
{zferdowsi, rsettimi, dstan}@cdm.depaul.edu, wcurran@depaul.edu

Abstract

This paper presents the results of a cross-sectional study that analyzes housing characteristics and the demographic makeup of Chicago community areas. The study is part of a larger project that uses data mining techniques to evaluate and model changes in neighborhoods that may lead to gentrification or abandonment, and ultimately to the loss of affordable housing. Our initial analysis has defined a typology for the seventy-seven Chicago community areas by classifying them in five segments according to median income level, education, race, and crime rates. Such a classification is important to understand similarities among the communities and to predict where changes are more likely to occur. For instance it can be hypothesized that gentrification is in progress for those communities that do not fit in just one type, but show subareas that belong to a higher-income segment.

1. Introduction

This paper reports the initial findings of a study that applies data mining techniques to monitor and analyze changes in Chicago neighborhoods, using a variety of data sources on Chicago housing characteristics and population demographics.

The main goal of the project is to predict those changes that may lead to loss of affordable rental housing. Data mining techniques will be used to identify a set of variables or risk factors for the loss of affordable rental housing through gentrification or abandonment. The definition of affordable rental housing varies, typically a rental unit is considered affordable when its rent price is less than 30% of the tenant's income. Gentrification often leads to loss of affordable rental housing since it changes the neighborhood class structure through housing investments which attract higher income people to the area [1].

The goal of this initial analysis is to identify similarities among Chicago neighborhoods and to create a typology for the seventy-seven Chicago community areas.

A previous study used clustering techniques to analyze a set of 1996 Census variables on occupation, tenure, household structure and mobility for the city of London [2]. This study showed that community areas in London could be grouped into twelve different clusters, characterized by working class residents, young, mobile, middle class renters, working class private renters, and middle class owner occupied suburbs [2].

A study by Helms [3] analyzed Chicago building permit data from 1995 to 2000 to predict building renovations. Permit data describe alteration and repair work done on a building including the type and estimated cost of work. The study demonstrated that building characteristics such as age, number of dwelling units, and number of stories were significant predictors of renovation. The data analysis also showed that 3 neighborhoods in Chicago would be more likely to be gentrified [3].

There are several other studies that used multivariate analysis for predicting gentrification. Ley and Dobson produced a regression model for gentrified communities in Vancouver from 1971-2001, and found that the areas with smaller distance to the beach and higher distance to commercial areas are more likely to be gentrified [4]. Wyly and Hammel applied stepwise discriminant analysis using census data from 1960 to 1990 for the cities of Chicago, Milwaukee, Minneapolis-St. Paul, and Washington, DC, and discovered that the most powerful indicators for gentrification in Chicago were income and education [1].

The following section describes the datasets used in the initial analysis described in this paper. The cluster analysis technique and results are also described in section 2. In section 3, conclusions and future work are discussed.

2. Methodology and results

The available datasets and their variables are listed in detail in table 1. Observations are aggregated at the census tract level and are limited to city of Chicago. Census tracts are small, relatively permanent statistical subdivisions of a county, designed to be relatively homogeneous units with respect to population characteristics, economic status, and living conditions for the purpose of presenting data. There are 878 census tracts in Chicago based on 2000 Census data.

Table 1. Available data sets and their variables

Data set Name	Variables
Census 2000	Race (Proportion of White, Hispanics, Black, Asian and others), Rent (Proportion of renter Occupied housing, Rent price), Education (Proportion of +25 that Completed 0-8 years school, Bachelor or Graduate), Population (Income, Language spoken at home, Migration), Age of building
Permit (1998-2004)	Proportion of Alteration and Repair, New construction and Demolition for Residential and Business
Building (2000-2004)	Proportion of Building violation, Building illegal conversion, and No heat
Crime (1998-2007)	Proportion of Total crime, Murder, Robbery, Rape, Assault, Theft/Larceny, Property crime, and Personal crime
Foreclosure (1998-2008)	Proportion of Total foreclosure, Apartment foreclosure, Condo foreclosure, Single family homes among foreclosure, Vacant building among foreclosure

As discussed before, the studies by Daly [2] and Helm [3] showed that cluster analysis is very useful to analyze geographical data and to create a topological map of community areas. We applied case-based K-means clustering to the datasets listed in Table 2 below, using the statistical software SPSS [5, 6]. K-means clustering is a powerful and efficient technique for clustering from large data sets. The K-means algorithm classifies the observations in a certain number (k) of clusters based on the set of selected variables. Cases are grouped according to their similarity with respect to these variables. Thus the variability within clusters is small and the one

between clusters is large. The application of the k-means clustering depends on the number of clusters and the choice of the similarity metric [5, 6]. Unfortunately there is no general solution to find the optimal number k of clusters. Typically the number of clusters is selected by comparing the results of several runs of the clustering algorithm with different k classes. In our study the optimal number of classes was found to be five, after applying the clustering technique with a number of classes k varying from 4 to 7. The center of the cluster was chosen as the average of the elements of that cluster and the selected distance metric was the Euclidean distance; the clustering algorithm converged at the 17^{th} iteration (there was no change in the cluster assignment after this iteration).

The variables that were used for the clustering analysis are shown in Table 2. All the variables are binned in 4 equal sized categories as 0 [0-25% of max value], 1 [25% - 50%], 2 [50% - 75%] and 3 [75% - 100%].

Table 2. Variables and datasets that are used for clustering

Data set Name	Variables are used for cluster analysis
Census 2000	Race (Proportion of White, Hispanics, Black, Asian and others), Rent (Proportion of renter Occupied housing), Education (Proportion of +25 that Completed 0-8 years school, Bachelor or Graduate), Median Income
Permit 2000	Proportion of Alteration and Repair, New construction and Demolition for Residential and Business
Building 2000	Proportion of Building violation, Building illegal conversion, and No heat
Crime 2000	Proportion of Total crime

The results of the clustering analysis are displayed in Table 3 below. Each row lists the number of observations and the average values of the discriminating features for each cluster. A color was assigned to each cluster for better visualization of the clustering results in terms of their mapping to the Chicago area (Figure 1).

In summary, Red areas are primarily Hispanic, with low education, low income and low rent price. Green areas have a larger number of White population with medium education, medium income, and low crime rate. Orange areas have predominantly White residents with high education and high income. Yellow areas have a mixed racial makeup. Blue areas have predominantly African American residents with low education and low income, low rent housing, high building violations and high crime rates.

Table 3. Results of cluster analysis

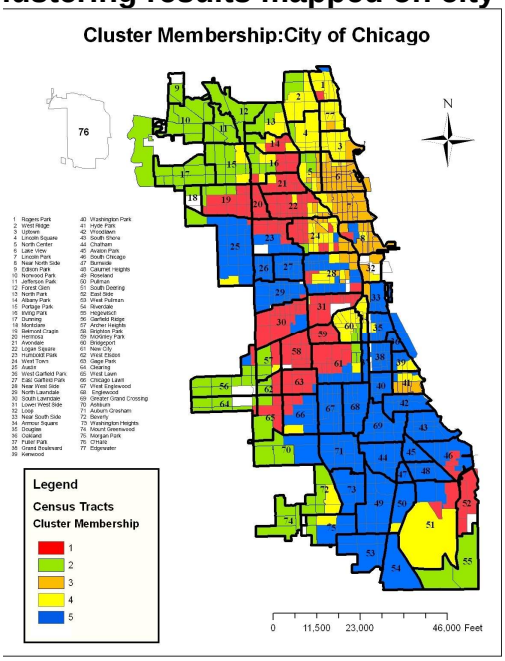
cluster#	number of case	income ($1000)	0-8 years school	bachelor or graduate	Race	rent occupied	rent price	building violation	total crime
1 (Red)	175	36	30%	11%	Hispanic (73%)	60%	$571	2.7%	8%
2 (Green)	105	60	9%	22%	White (75%)	27%	$678	1%	5%
3 (Orange)	96	102	3%	71%	White (78%)	59%	$936	2.3%	12%
4 (Yellow)	127	45	11%	67%	Mix	67%	$635	3.7%	8%
5 (Blue)	341	31	9%	11%	Black (94%)	62%	$521	3.7%	12%

Although both Orange and Blue areas are characterized by high total crime rate, the type of crimes is different. Blue areas are affected by high violent crime rates (murder, rape, robbery,

assault, and personal crime), while the majority of crimes reported in the Orange areas are thefts and property crimes.

Figure 1 maps the clustering results in Table 3 for the city Chicago. There are 34 missing tracts that are colored in white.

Figure 1. Clustering results mapped on city of Chicago



The results of the clustering analysis at the community level illustrate that about 50% of the seventy-seven Chicago communities completely belong to a cluster; it means that 100% of tracts in that community are in a same cluster (these communities are fully colored in one color in the Chicago map in Figure 1). About 20% of communities have more that 80% tracts in a same cluster. About 15% of communities have more that 60% and less than 80% tracts in a same cluster. The communities having different colored areas (meaning that they have tracts in different clusters) represent a mixture of typologies. It can be hypothesizes that this is a sign of a change in the neighborhood. For instance West Town (community n. 24 in Figure 1) shows signs of gentrification as there are higher income tracts (yellow and orange clusters) that are more similar to the neighboring community areas.

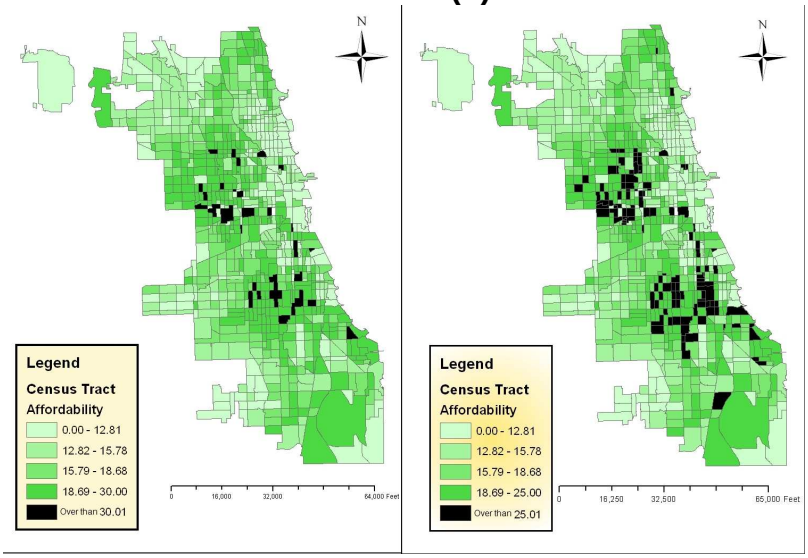
Median income and monthly gross rent are analyzed to evaluate if rent is affordable in a specific tract. A widely used statistics for affordable rent is the ratio of monthly rent price divided by family income. If the ratio is above 30%, the rent is not affordable. Figure 2(a) shows in black the tracts where the median gross rent is more than 30% of the median income, and Figure 2b shows the tracts where the ratio is above 25%. Most tracts where people are burdened by unaffordable rents are in the Blue and in the Red areas that have low income population. This may indicate that affordable housing is lacking in poorer areas, where families have low income.

3. Conclusion and future work

The analysis described in the paper has allowed us to understand the typology of Chicago communities. Further analyses will explore changes over time using census data from 1980 to

2000, and other additional datasets that are being recently acquired. Multivariate time series analysis and sequence data mining will be used to investigate how the typology of community areas has changed over the years and to find significant factors that describe most intensive changes in neighborhoods in the last 20 years. Such factors can be used to understand changes in the complex demographic makeup of the city of Chicago and to predict future gentrification or abandonment of neighborhoods.

Figure 2. Affordability analysis – ratio of median gross rent over median income
(a) Black tracts have ratio > 30% (b) Black tracts have ratio > 25%



Acknowledgments

This study is funded by the Institute for Housing Studies at DePaul University, through a MacArthur Foundation grant.

References

[1] Wyly, Elvin K., and Hammel, Daniel J., "Modeling the Context and Contingency of Gentrification" *Journal of Urban Affairs*, vol. 20, No. 3, 1998, pp. 303-326

[2] Daly, Martin, "Characteristics of 12 clusters of wards in Greater London" *Department of Planning and Transportation*, Research report No.13, 1971

[3] Helms, Andrew C., "Understanding gentrification: an empirical analysis of the determinants of urban housing renovation" *Journal of Urban Economics*, Vol. 54, 2003, pp. 474-498

[4] Ley, David and Dobson, Cory, "Are There Limits to Gentrification? The Contexts of Impeded Gentrification in Vancouver" *Journal of Urban Studies*, vol.45, No.12, 2008 Nov., pp. 2471-2498

[5] Kanungo, Tapas, Netanyahu, Nathan S., and Wu, Angela Y., "An Efficient k-Means Clustering Algorithm: Analysis and Implementation" *IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, July 2002*, pp. 881-892

[6] Han, J and Kamber, M, *Data Mining Concepts and Techniques*, Second Edition, 2006, Morgan Kaufman, San Francisco, CA