Paper 092-2012

Multidimensional Scaling on ZIP Codes

Chao Huang, Oklahoma State University, Stillwater, OK
Xiangxiang Meng, SAS Institute Inc., Cary, NC

ABSTRACT

Postal code or ZIP code of the United States is a series of five digits initially serving the purpose of mail delivery, and is now extensively used for geographical partition in many fields. In SAS®, the ZIPCITYDISTANCE function calculates the physical distances between any pair of ZIP codes. The MDS procedure, based on the multidimensional scaling technology, translates the distance matrix of the ZIP codes into relational numeric values. This paper describes a statistical solution that separates the ZIP codes into limited levels by PROC MDS and the clustering methods in SAS/STAT.

INTRODUCTION

Multidimensional scaling (MDS) is an optimization technique that detects similarities or dissimilarities among various objects by a minimizer toward a cost function such as the equation below [1]. Here X_i and X_j are the vectors form the dissimilarity matrix of distances, while σ_{ij} is the distance between the two objects.

$$\boldsymbol{min} \sum_{i<j} (\left\| \boldsymbol{x_i} - \boldsymbol{x_j} \right\| - \boldsymbol{\delta_{ij}})^2$$

In market research, MDS is widely utilized to reflect the perceptions of customers onto the conceptual maps. This method also becomes popular in analyzing social network. The MDS procedure in SAS is based on such a statistical algorithm. Using PROC MDS, Larry extracted the names and email addresses from SAS-L, a popular listserv for SAS users, and mapped all users to a social circle [2].

ZIP codes provide useful information, and also bring many challenges. First, ZIP codes are unbalanced that can be messy in dividing areas. For example, in Texas, some counties contain more than 100 ZIP codes, while some counties share only one ZIP code (Figure 1).

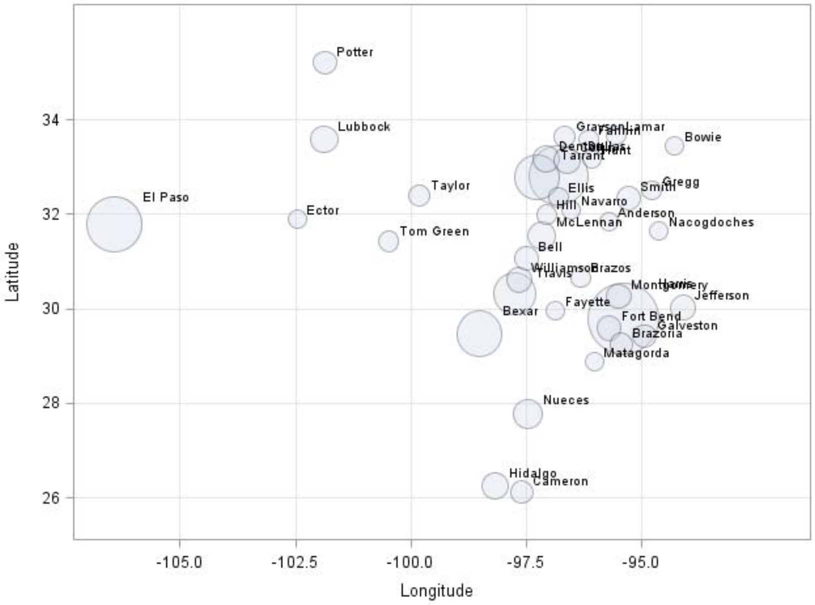


Figure 1. Top 40 counties in Texas that have most of the ZIP codes (Bubble sizes represent the number of ZIP code levels each county has)

Second, ZIP codes tend to change over time. For example, in Texas, the first two numbers of all ZIP codes are 73, 75-79 or 88 (Figure 2A). Thus, 7 regions of Texas can be separated: Dallas-Fort Worth-Arlington Metropolitan Area is divided to two parts. 2 ZIP codes starting with 73 are surrounded by ZIP codes starting with 79 (Figure 2B).

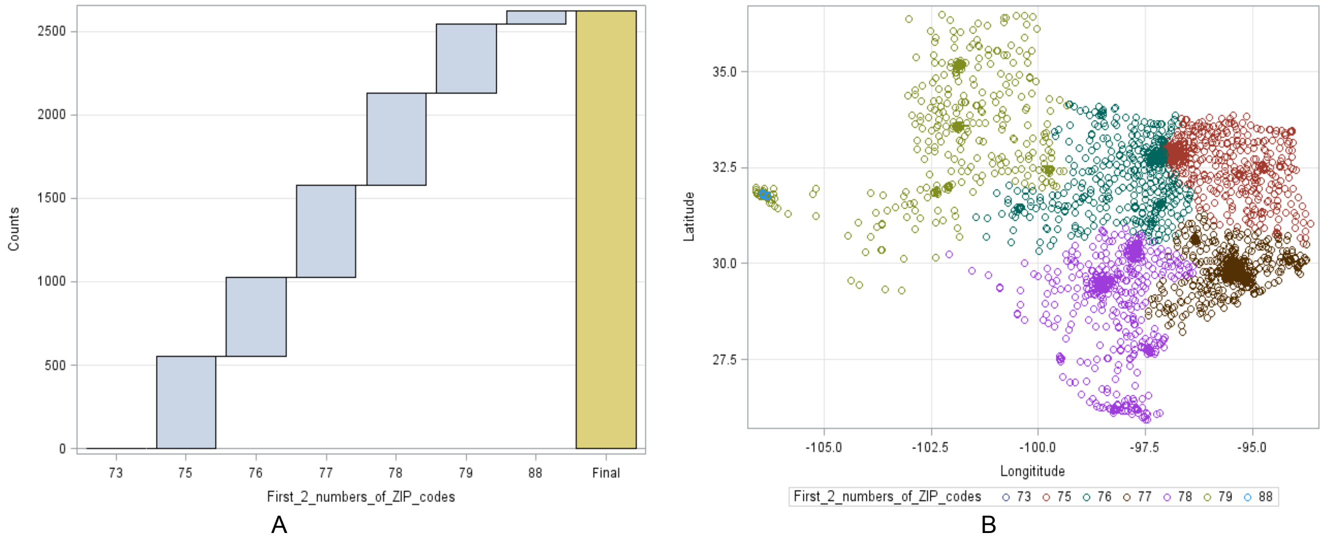


Figure 2. Frequencies of ZIP codes by the first two digits. (A) Counts of each category from 73 to 88. (B) Categorized ZIP codes scattered by their latitude and longitude

To better use ZIP codes, the U.S. Census Bureau created a tool called ZIP Code Tabulation Areas, which is a statistical geographic entity produced by for tabulating summary statistics from the 2010 Census [3]. In this paper, we introduce a simple approach to translate those ZIP codes to meaningful numeric values that are easier to be categorized.

THE ZIPCITYDISTANCE FUNCTION

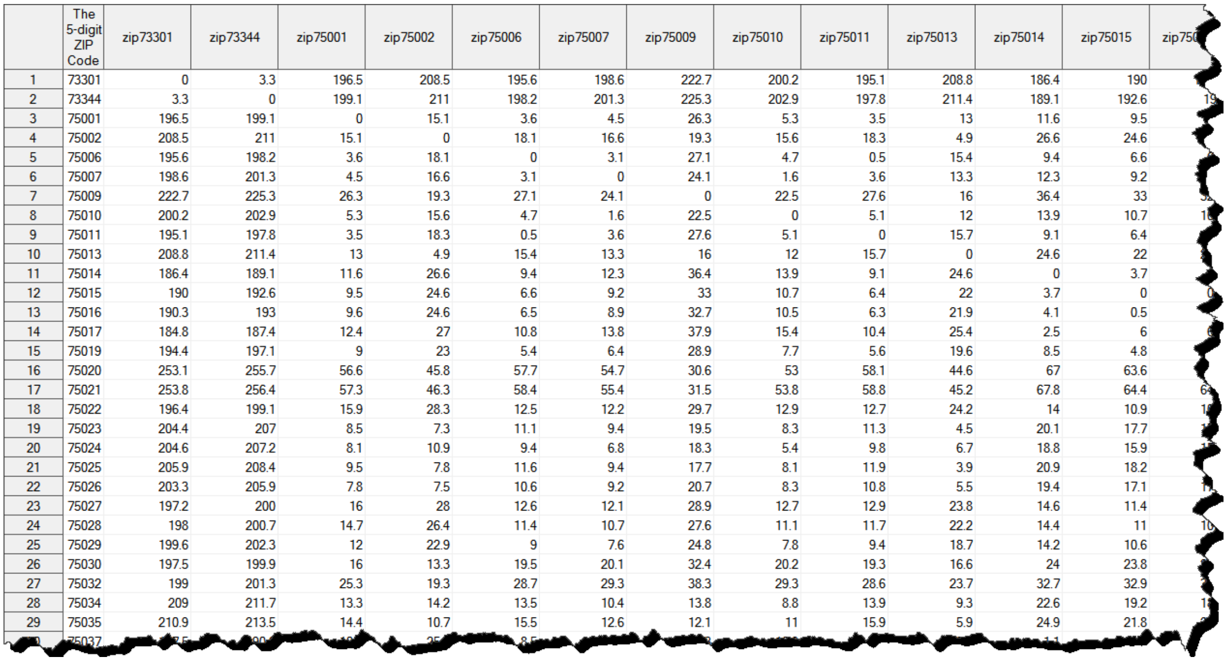
A data set from SAS's HELP library, SASHELP.ZIPCODE, records 41,466 distinctive ZIP codes nationwide. The data set of Texas is extracted from SASHELP.ZIPCODE and contains 2,650 ZIP codes ranging from 73301 to 88595.

```
data txzip;
   set sashelp.zipcode;
   where statecode = "TX";
run;
```

The ZIPCITYDISTANCE function in SAS measures the distances between any two ZIP codes by miles. We apply this function to all of Texas's 2,650 ZIP codes and eventually create a distance matrix with 2,650 columns by 2,650 rows by the codes below (Figure 3).

```
* Compute distances using the ZIPCITYDISTANCE function;
proc sql;
   create table _1 as
   select a.zip as zipa, b.zip as zipb, zipcitydistance(zipa, zipb) as distance
   from txzip as a, txzip as b;
quit;

* Transpose into a distance matrix;
proc transpose data = _1 out = _2 prefix = zip;
   by zipa;
   id zipb;
   var distance;
run;
```



	The 5-digit ZIP Code	zip73301	zip73344	zip75001	zip75002	zip75006	zip75007	zip75009	zip75010	zip75011	zip75013	zip75014	zip75015
1	73301	0	3.3	196.5	208.5	195.6	198.6	222.7	200.2	195.1	208.8	186.4	190
2	73344	3.3	0	199.1	211	198.2	201.3	225.3	202.9	197.8	211.4	189.1	192.6
3	75001	196.5	199.1	0	15.1	3.6	4.5	26.3	5.3	3.5	13	11.6	9.5
4	75002	208.5	211	15.1	0	18.1	16.6	19.3	15.6	18.3	4.9	26.6	24.6
5	75006	195.6	198.2	3.6	18.1	0	3.1	27.1	4.7	0.5	15.4	9.4	6.6
6	75007	198.6	201.3	4.5	16.6	3.1	0	24.1	1.6	3.6	13.3	12.3	9.2
7	75009	222.7	225.3	26.3	19.3	27.1	24.1	0	22.5	27.6	16	36.4	33
8	75010	200.2	202.9	5.3	15.6	4.7	1.6	22.5	0	5.1	12	13.9	10.7
9	75011	195.1	197.8	3.5	18.3	0.5	3.6	27.6	5.1	0	15.7	9.1	6.4
10	75013	208.8	211.4	13	4.9	15.4	13.3	16	12	15.7	0	24.6	22
11	75014	186.4	189.1	11.6	26.6	9.4	12.3	36.4	13.9	9.1	24.6	0	3.7
12	75015	190	192.6	9.5	24.6	6.6	9.2	33	10.7	6.4	22	3.7	0
13	75016	190.3	193	9.6	24.6	6.5	8.9	32.7	10.5	6.3	21.9	4.1	0.5
14	75017	184.8	187.4	12.4	27	10.8	13.8	37.9	15.4	10.4	25.4	2.5	6
15	75019	194.4	197.1	9	23	5.4	6.4	28.9	7.7	5.6	19.6	8.5	4.8
16	75020	253.1	255.7	56.6	45.8	57.7	54.7	30.6	53	58.1	44.6	67	63.6
17	75021	253.8	256.4	57.3	46.3	58.4	55.4	31.5	53.8	58.8	45.2	67.8	64.4
18	75022	196.4	199.1	15.9	28.3	12.5	12.2	29.7	12.9	12.7	24.2	14	10.9
19	75023	204.4	207	8.5	7.3	11.1	9.4	19.5	8.3	11.3	4.5	20.1	17.7
20	75024	204.6	207.2	8.1	10.9	9.4	6.8	18.3	5.4	9.8	6.7	18.8	15.9
21	75025	205.9	208.4	9.5	7.8	11.6	9.4	17.7	8.1	11.9	3.9	20.9	18.2
22	75026	203.3	205.9	7.8	7.5	10.6	9.2	20.7	8.3	10.8	5.5	19.4	17.1
23	75027	197.2	200	16	28	12.6	12.1	28.9	12.7	12.9	23.8	14.6	11.4
24	75028	198	200.7	14.7	26.4	11.4	10.7	27.6	11.1	11.7	22.2	14.4	11
25	75029	199.6	202.3	12	22.9	9	7.6	24.8	7.8	9.4	18.7	14.2	10.6
26	75030	197.5	199.9	16	13.3	19.5	20.1	32.4	20.2	19.3	16.6	24	23.8
27	75032	199	201.3	25.3	19.3	28.7	29.3	38.3	29.3	28.6	23.7	32.7	32.9
28	75034	209	211.7	13.3	14.2	13.5	10.4	13.8	8.8	13.9	9.3	22.6	19.2
29	75035	210.9	213.5	14.4	10.7	15.5	12.6	12.1	11	15.9	5.9	24.9	21.8

Figure 3. Part of the distance matrix constructed using the ZIPCITYDISTANCE function

CLUSTERING ZIPCODES USING MDS COORDINATES

We implement multidimensional scaling directly onto this distance matrix. In the MDS procedure, if the LEVEL at the PROC statement is set to be ABSOLUTE, this procedure produces two dimensions. The FASTCLUS procedure applies the nearest centroid sorting method that minimizes the sum of squared distances from other data points to the centroids. Here we arbitrarily choose 5 cluster numbers to cluster the two dimensions generated by MDS. One thing is worth noting: multidimensional scaling is a computation-intensive algorithm; it may take minutes to finish the PROC MDS operation described below.

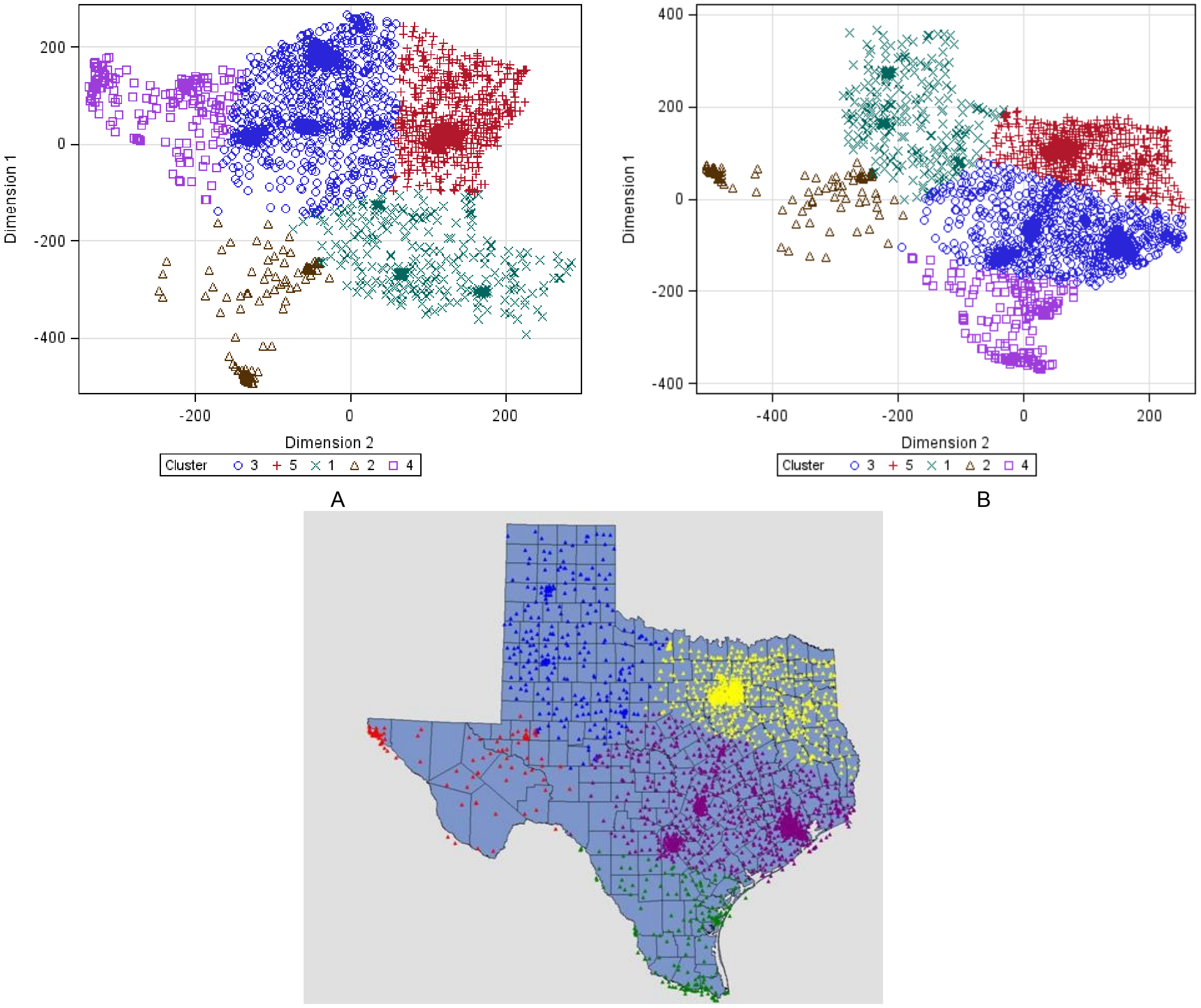
```
* Apply the multidimensional scaling on the distance matrix;
proc mds data = _2 level = absolute out = _3 ;
   id zipa;
run;

* Specify five clusters;
%let cluster = 5;
proc fastclus data = _3 maxc = &cluster out = _4;
   var dim:;
   where _name_ is not missing;
run;

* Draw a scatter plot with cluster labels;
proc sgscatter data = _4;
   plot dim1 * dim2 / group = cluster grid;
run;
```

To visualize the distance-transformed dimensions, we label and plot them with their corresponding cluster numbers. An interesting phenomenon is that the scatter plot successfully recovers the relative distances but with a misleading angle (Figure 4A). We rotate the scatter plot to match the physical map by the codes below (Figure 4B). Following Darrell and Jeff's method based on PROC GAMP [4], we are able to display the clustering results on the map of Texas (Figure 4C).

```
* Rotate the coordinates;
data _5;
   set _4;
   dim1  = -dim1;
   dim21 = dim2*cos(3.14159*(3/8))-dim1*sin(3.14159*(3/8));
   dim11 = dim2*sin(3.14159*(3/8))+dim1*cos(3.14159*(3/8));
run;
proc sgscatter data = _5;
   plot dim11 * dim21 / group = cluster grid;
run;
```



A

B

C

Figure 4. Clustered ZIP codes scattered in 2D space. (A) ZIP codes plotted by two dimensions with cluster labels. (B) ZIP codes rotated from Figure 4A. (C) Clustered ZIP codes mapped back to Texas map

In addition, we retrieve a total 160 ZIP codes of the Orlando metropolitan area from SASHELP.ZIPCODE by Metro Statistical Area code (MSA). Besides, the ZIP code of Disney Resort where SAS Global Forum 2012 is held is 32830. Following the same process described above, we also obtain five clusters. The Disney Resort, Sanford and Lake Okeechobee are labeled separately (Figure 5).

```
* Extract ZIP codes for Orlando metropolitan area where MSA equals 5960;
data oma;
```

```
   set sashelp.zipcode;
   where statecode = "FL" and msa = 5960;
run;
```

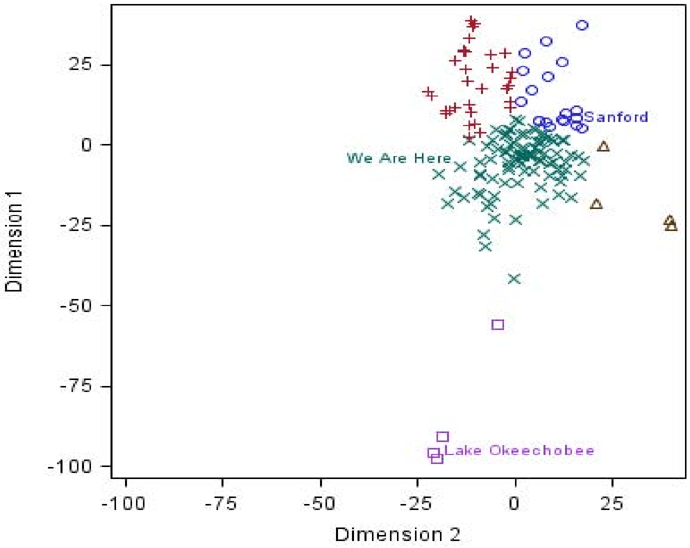


Figure 5. Clustered ZIP Codes for the Orlando Metropolitan Area

USING CLUSTERED ZIPCODE FOR DATA EXPLORATION AND ANALYSIS

Clustering zip code data with MDS coordinates provides an efficient way to reduce the number of zip codes to a manageable number of clusters (regions), which implies we can use the clustered zip codes to identify the regional patterns in a data set for data exploration or data analysis. In the section, we illustrate this idea using the 2010 survey data on occupied housing characteristics downloadable from the US Census Bureau [5]. We cluster the ZIP codes of Texas into twenty regions (Figure 6A), and then compute the mean percentage of the three occupied housing characteristics: percentage of houses occupied by only one person (Figure 6B), percentage of houses occupied by the owner (Figure 6C) and percentage of houses occupied by householder over 65 years old (Figure 6D).

```
proc sql;
   create table _6 as
      select a.zipa as zipcode, a.dim1, a.dim2,
         b.percent_owner, b.percent_1p, b.percent_senior, b.averageSize
      from _3 as a inner join texas_household as b
         on a.zipa eq b.zipcode;
quit;

* Only clustering the zip codes with valid household information;
%let cluster = 20;
proc fastclus data = _6 maxc = &cluster out = _7 noprint;
   var dim:;
run;

* Coordinates reflection and rotation;
data _8;
   set _7;
   dim1 = -dim1;
   dim21 = dim2*cos(3.14159*(3/8))-dim1*sin(3.14159*(3/8));
   dim11 = dim2*sin(3.14159*(3/8))+dim1*cos(3.14159*(3/8));
run;

* Summary statistics;
proc sql;
```

```
   select cluster,
      mean(percent_owner/100)  as mean1 label='Average of Percentages of Owner Occupied',
      mean(percent_1p/100)     as mean2 label='Average of Percentages of 1-person Occupied',
      mean(percent_senior/100) as mean3 label='Average of Percentages of Senior (65+)
Householder'
   from _8
   group by cluster
   order by cluster;
quit;
```

For each housing percentage, we identify the regions with the highest values and the lowest values and plot these regions with different colors from other regions (Figure 6). It is clear that the clustered zip codes help finding some interesting regional characteristics of this housing data set. For instance, in the two biggest city of Texas (Dallas and Houston), houses are less likely to be occupied by the owner, or by seniors (age 65+). In this example, we simply explore the regional pattern by computing the summary statistics in each cluster of zip codes, and it is worth to mention that the clustered zip codes can play a more important role in data analysis, such as using the cluster labels as predictors in regressions or decision trees.

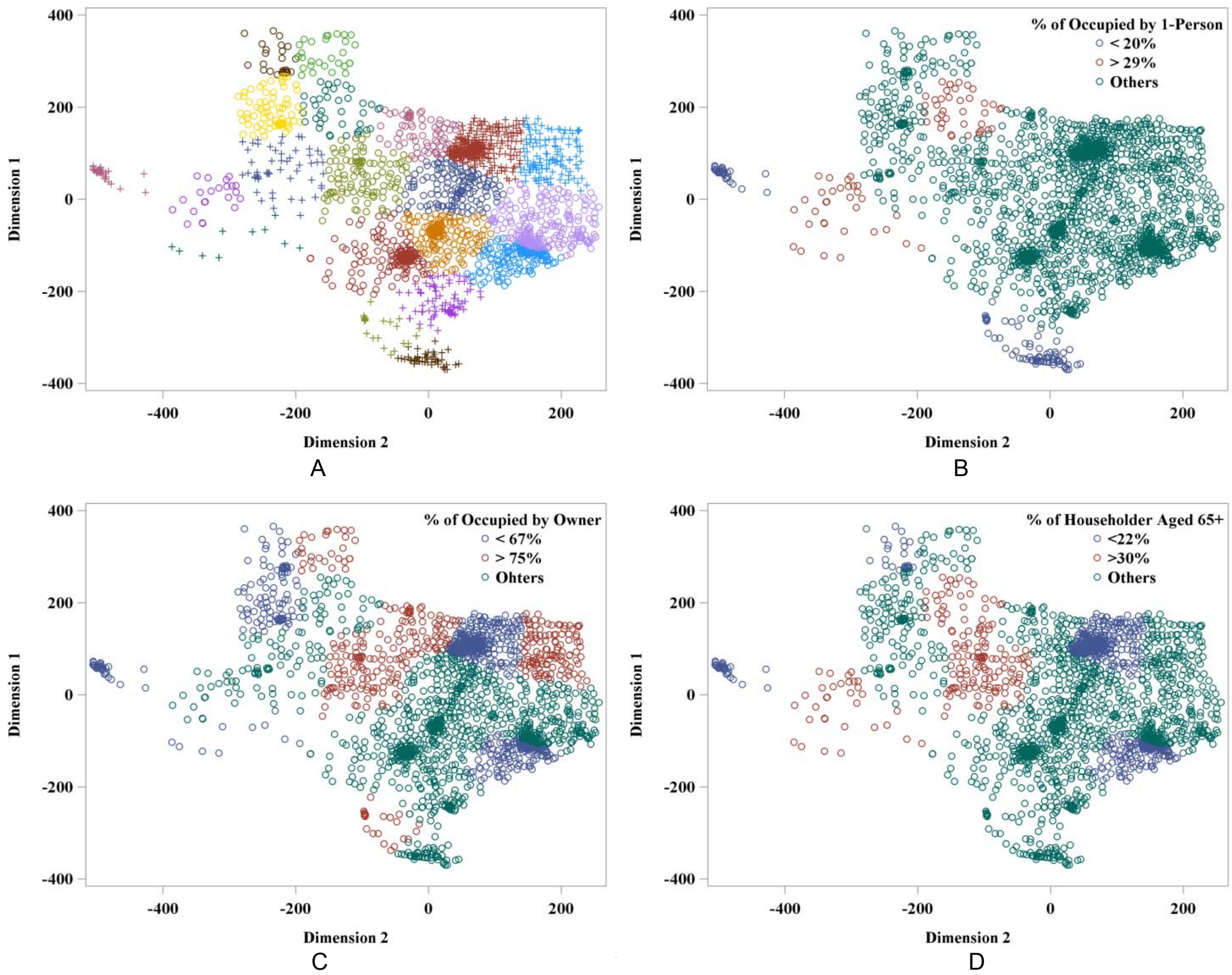


Figure 6. Clustered Texas Zip Codes Combined with the Us Census Bureau Housing Data. (A) Texas's ZIP Codes Clustered as 20 clusters after MDS. (B) Percentage of Houses Occupied by Only One Person Labeled by Clusters. (C) Percentage of Houses Occupied by the Owner Labeled by Clusters. (D) Percentage of Houses Occupied by Householder Over 65 Years Old Labeled by Clusters

CONCLUSION

The MDS procedure in SAS/STAT is a convenient way to interpret ZIP codes of the United States. SAS's clustering and graphical procedures, together with the powerful ZIPCITYDISTANCE function, support SAS to be a productive platform to utilize the information by the ZIP codes.

REFERENCES

1. Chun-houh Chen and Antony Unwin. 'Handbook of Data Visualization'. First edition. Springer, 2008.
2. Larry Hoyle. 'Visualizing Two Social Networks Across Time with SAS'. SAS Global Forum Proceeding 2009. http://support.sas.com/resources/papers/proceedings09/229-2009.pdf
3. The US Census Bureau. 'ZIP Code® Tabulation Areas'. http://www.census.gov/geo/ZCTA/zcta.html
4. Darrell Massengill and Jeff Phillips. 'Tips and Tricks IV: More SAS/GRAPH Map Secrets'. SAS Global Forum Proceeding 2009. http://support.sas.com/resources/papers/sgf09/230-2009.pdf
5. The US Census Bureau. 'Occupied Housing Characteristics: 2010'. http://factfinder2.census.gov/faces/nav/jsf/pages/searchresults.xhtml?ref=top&refresh=t

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Chao Huang
Office of Institution Research and Information Management
221 PIO Building
Stillwater, OK. 74075
Email: hchao8@gmail.com
Web: www.sasanalysis.com

Xiangxiang Meng
SAS Institute Inc,
Cary, NC. 27513
Email: Xiangxiang.Meng@sas.com