8-15-1997

# Using Zip-Code as an Attributein Direct Marketing Research

Raja Sengupta
*Southern Illinois University*, sarojsen@siu.edu

Siddhartha Bhattacharya
*Southern Illinois University*, sidb@siu.edu

# Using Zip-Code as an Attribute in Direct Marketing Research


**Raja Sengupta**
**Siddhartha Bhattacharya**
Department of Geography
Department of Management
Southern Illinois University
Carbondale, IL 62901
sarojsen@siu.edu 618-453-7252
sidb@siu.edu 618-453-7884


## Introduction

The importance a locational indicator such as zip-codes in identifying potential customers is readily recognized by the overwhelming interest of marketers in the field of "geodemographics". According to Goss (1995), "geodemographics is an information technology that enables marketers to predict behavioral responses of customers based on statistical models of identity and residential location". Therefore, the rationale behind all geodemographic systems is that individuals can be divided into groups based on certain characteristics. These characteristics can include income, education, product preference and household type. Further, geodemographic systems also follow the premise that "birds of a feather flock together", or that individuals who fall into the same group live in similar neighborhoods.

All geodemographic systems start with some basic unit of geography that can be used to divide the U.S. into segments for which relevant information such as income, education etc. is available. These units can be census blocks, zip-codes or individual households. The geodemographic systems then groups the basic units based on aggregates of psychographic (personality traits), consumer behavior (product loyalty etc.) and demographic information (age, income, race etc.) for the region. Each group is termed as a "segment" or "cluster". Mitchell (1995) describes a "cluster" as "a class of households with common demographic and lifestyle characteristics, designated by a label". The clusters are developed from the data attached to the basic geographic units using multivariate regression analysis. Once developed, these clusters can then be used to predict the location of potential buyers.

Given the fact that location can have a direct bearing on response rate, it would only seem sensible that zip-codes should be successful in developing DM models that predict customer response to direct marketing campaigns. However, although customer's zip-codes are also found in the database, they are often not used as an attribute in developing the DM model because they do not appear to have any information about the customer's buying habits. Part of the reason for this paradox may be because a nationwide mailing may not elicit response from more than a single customer in a certain zip-code. Each zip-code then would yield very little information. An alternative is to cluster the zip-codes into groups based on spatial adjacency (i.e. zip-codes that are adjacent to one another fall into the same category) and then use the combined zip-codes as an attribute in the DM model. The assumption here is that adjacent zip-codes will have similar response rates. However, as consecutive zip-codes are not spatially adjacent to one another, traditional databases cannot be used to group the zip-codes in this fashion. This problem can be overcome using a Geographic Information System(GIS). In this paper, we present one approach to clustering zip-codes using a GIS, and present some preliminary results obtained following the clustering.

## Methodology

Silk (1979) provides two methods to cluster point patterns into groups. The first technique is termed "quadrat analysis". Using a GIS with both raster and vector capability (such as UNIX ARC/INFO), a grid can be overlain over the points distributed in space. Silk also proposed a rule of thumb for selecting the cell size for the grid , where he suggested that the cell size should approximately equal twice the mean area per point. Once a grid has been overlain over the points, all the points that fall in the same cell of the grid can be cluster together as one group. We used this technique to cluster the zip-codes into groups based on spatial adjacency.

A second technique provided by Silk (1979) to cluster points uses the nearest neighbor algorithm. The nearest neighbor algorithm identifies the closest neighbors to each point in a distribution of points. Thus, a zip-code can be clustered with it's nearest neighboring zip-codes. Future work done by us will focus on the effectiveness of using the nearest neighbor algorithm instead of quadrat analysis to cluster zip-codes.

To test the hypothesis that grouping spatially adjacent zip-codes will allow them to be used as an attribute in developing DM models, we used the clustered zip-codes and a real world dataset of 26,178 customers. Our objective was to develop a score indicating response rate for each cell of the grid, and use decile analysis to determine the benefits of concentrating sales to customers living in cells with a high score. The real world dataset contained information about customers who had responded (responders) or not responded (non-responders) to a direct marketing (DM) campaign. The first step in the process was to sum the responders and non-responders for each zip-code. Based on the aggregate total of responders and non-responders per zip-code in a cell of the grid, two models were developed that generated a score for each cell. The scores ($WG_i$) was calculated as follows:

1) For Model 1

$$W_{G_i} = \frac{\sum_{z \in G} W_z}{N_{G_i}} \quad \text{where} \quad W_z = \frac{r_z}{r_z + n_z}$$

2) For Model 2

$$W_{G_i} = \frac{\sum_{z \in G_i} r_z}{\sum_{z \in G_i} r_z + \sum_{z \in G_i} n_z}$$

In both the equations above, $W_z$ denotes the score of an individual zip-code $Z$ in a cell, $r_z$ is the number of responders in a zip-code, $n_z$ is the number of non-responders in a zip-code, and $NG_i$ is the total number of zip-codes in cell $G_i$. The scores obtained from the models above were then associated with the individuals in the main database using their zip-codes. In keeping with DM industry practice, performance of the models was examined through decile analysis (David Shepard Associates, 1995) of the scores assigned to the individual customers. The results obtained are shown in Table 1 and 2.

Further, another model was developed that examined the premise that the information content of the zip-codes are adequate by themselves, and aggregating adjacent zip-codes spatially using quadrat analysis yields no added benefit. The formula for this model is given below (the symbols are as defined earlier):

$$W_z = \frac{r_z}{r_z + n_z}$$

As before, the scores $W_z$ obtained above (for each zip-code, in this case) were associated with the individuals in the main dataset using zip-codes, and model performance examined through decile analysis(Table 3).

## Results and Discussion

A decile analysis shows individuals ranked by their respective model scores - higher scores indicating better performance -- and separated into 10 equal groups. In table 1, a typical decile analysis, the first row (top decile) indicates performance for the best 10% of the individuals as identified by the model. The Cumulative Lift (CL) provides a measure of improvement over a random mailing, and is calculated as follows:

$$CL = \frac{\text{cumulative response rate for decile}}{\text{overall response rate}} * 100$$

Thus, in Table 1, a cumulative lift of 208 in the top decile indicates that the first model is expected to identify 2.08 times more responders than a random mailing to 10% of the file. Similarly, if 20% of the file is mailed to, the first model is expected to perform 1.75 times better than a random mailing (no model). In Table 2, a cumulative lift of 220 and 185 was achieved for the top two decile by the second model.

On the other hand, the third model (Table 3) identified only 1.27 and 1.19 times more responders than a random mailing survey for the top two deciles. These results indicate that spatial clustering of the 5-digit zip-codes from the real world dataset using quadrat analysis improves the information content of the zip-codes, and thereby allowing them to be used as an attribute along with other variables in developing DM models. Future work will focus on using nearest neighbor analysis to cluster zip-codes, as well as testing the model developed on other real world datasets.

**References**: available on request from the authors.

| Decile | Number of Customers | Number of Responders | Cum. Response Lift |
|---|---|---|---|
| top | 2,618 | 1,572 | 208 |
| 2 | 2,619 | 1,081 | 175 |
| 3 | 2,619 | 953 | 159 |
| 4 | 2,619 | 901 | 149 |
| 5 | 2,618 | 813 | 140 |
| 6 | 2,619 | 740 | 133 |
| 7 | 2,619 | 645 | 126 |
| 8 | 2,619 | 560 | 120 |
| 9 | 2,619 | 310 | 111 |
| bottom | 2,618 | 0 | 100 |
| Total | 26,187 | 7,575 | |

Table 1: Model 1 decile analysis

| Decile | Number of Customers | Number of Responders | Cum. Response Lift |
|---|---|---|---|
| top | 2,618 | 1,665 | 220 |
| 2 | 2,619 | 1,144 | 185 |
| 3 | 2,619 | 987 | 167 |
| 4 | 2,619 | 892 | 155 |
| 5 | 2,618 | 800 | 145 |
| 6 | 2,619 | 706 | 136 |
| 7 | 2,619 | 616 | 128 |

| | | | |
|---|---|---|---|
| 8 | 2,619 | 499 | 121 |
| 9 | 2,619 | 266 | 111 |
| bottom | 2,618 | 0 | 100 |
| Total | 26,187 | 7,575 | |

Table 2: Model 2 decile analysis

| Decile | Number of Customers | Number of Responders | Cum. Response Lift |
|---|---|---|---|
| top | 2,618 | 962 | 127 |
| 2 | 2,619 | 843 | 119 |
| 3 | 2,619 | 880 | 118 |
| 4 | 2,619 | 821 | 116 |
| 5 | 2,618 | 862 | 115 |
| 6 | 2,619 | 804 | 114 |
| 7 | 2,619 | 844 | 113 |
| 8 | 2,619 | 1,004 | 116 |
| 9 | 2,619 | 555 | 111 |
| bottom | 2,618 | 0 | 100 |
| Total | 26,187 | 7,575 | |

Table 3: Model 3 decile analysis