Statistical Issues in the Use of Aggregate Market Research Data in Health Communication Planning

William E. Pollard
Research Psychologist (Ret.)
Atlanta, Georgia

Abstract
This paper considers statistical issues entailed in using geographically aggregated data from market research data bases and segmentation systems for public health communication planning. While these issues have been widely discussed in the literature on aggregate data use in epidemiology, sociology, political science, geography, and geostatistics, they have received little attention in the health communication and the social marketing literature. The purpose of this presentation is to highlight these issues and their importance in drawing conclusions from data of this type.

Key Words: Health and market research surveys, health communication planning

1. Introduction

1.1 Health Communication, Social Marketing, and Audience Segmentation
Communication to affect behavior is an important public health tool, and understanding an at-risk target audience is essential for developing effective communication. The field of health communication draws from social psychology and other social sciences, health education, mass communication, and marketing for the design and delivery of effective communication. Aspects of this latter area - marketing, as adapted for public health purposes, will be the focus here. "Social marketing", Grier and Bryant write (2005: 321) "is typically defined as a program planning process that applies commercial marketing concepts and techniques to promote voluntary behavior change" and Kotler, Roberto, and Lee (2002: 5) write "social marketing is the use of marketing principles and techniques to influence a target audience to voluntarily accept, reject, modify, or abandon a behavior for the benefit of individuals, groups, or society as a whole."

A key concept in marketing is that of segmentation – dividing the market or audience into smaller groups that might require different marketing approaches (McDonald and Dunbar, 2004; Meyers, 1996; Weinstein, 1994). Individuals within segments have common needs, wants, lifestyles, behaviors, and values that are likely to make them respond in a similar manner to marketing efforts. Separate messages and approaches can then be tailored for different segments and particular segments may be selected for targeting.

1.2 Geodemographic Segmentation
One widely-used approach to segmentation for commercial marketing is that of *geodemographic segmentation*. Such segmentation is based on classification systems for

small census and postal geographic areas that differentiate such areas with respect to their use of products and services. These area types are the consumer segments that are used in planning direct marketing efforts. Often census demographic data and market research data on product use, media use, and lifestyle are summarized for these segments for these segments and are packaged in commercial data bases along with mapping software by various companies. Some companies and their segmentation systems include CACI – ACORN, Claritas – PRIZM NE, Experian – MOSAIC, and the GIS firms ESRI – COMMUNITY TAPESTRY and MapInfo – PSYTE. Such systems generally partition the census and postal geographies into 50 to 100 different area types or segments. Developments in the US and the UK are described in Harris, Sleight, and Webber (2005) and Sleight (2005).

These segmentation systems are used by businesses to segment the population on probability of response to marketing efforts so that resources can be focused on the most responsive and profitable segments. The segmentation systems help businesses identify and focus on their "best customers" - the audience that will be most likely to perform the desired outcome behavior (i.e., purchase the product). A primary determinant of who is included in the target audience is the probability that they will perform the desired behavior, and improvement in marketing response is brought about by changing or modifying the target audience and refocusing resources on this audience

This use of these systems is based on two ideas. The first is that the people most likely to respond to one's marketing efforts will be like those people who have responded in the past. The second is that one way to find people similar to previous responders is by focusing on the same kinds of geographic areas or segments in which previous responders reside because of some similarities in people who reside in the same area. By analyzing addresses of previous responders in internal company data files, businesses can determine which kinds of areas or segments had the best response rate and, following this, they can than concentrate their resources on people and households with addresses in these segments and avoid expending resources on other segments. See Curry (1993) and Drozdenko and Drake (2002) for a description of this process.

In its most basic form this is essentially finding the best audience to fit the marketing effort and product, or what can be called *audience determination*. Improvement in marketing response and efficiency can be obtained simply through modifying a potential target audience to include those segments with the highest probability of response and to exclude those with low probability of response. In general, this increase in response and efficiency with this type of response-based segmentation is the primary benefit obtained by businesses in using these systems. Direct marketing has an average response rate of around 5% and increasing the response rate by a few percentage points can be quite profitable.

The situation is different in public health efforts where audiences are often fixed, demographically-defined groups dictated by factors such as funding mandates, epidemiological disease prevalence and at-risk considerations, disparities, etc. Furthermore these audiences are the generally the focus of public health attention for the very reason that they have had low probability of response to public health efforts. Focusing on segments that contain these groups does not yield the primary benefit achieved using geodemographic segmentation in commercial marketing, that is, improved response simply through selecting an audience made up of segments with the

highest probability of response. One is simply left with segments with low response rates.

Consequently much public health interest in these segmentation systems has centered on the use of summarized market research data for the segments to draw conclusions about public health target populations that reside in the segments, or what can be called *audience analysis*. In particular, the interest is in how communication-relevant items vary with the concentration of the target population in the segments. For example, what are the most highly used media channels in segments that have high concentration of some demographic at-risk population? However this is not as straightforward as segmentation for *audience determination*. The difficulty is that the market research data for the segments are aggregate summaries for the populations in the areas classified under the segments. The central issue here is that of using associations in aggregate area-level based segments data to draw conclusions about associations in individuals. This type of cross-level inference entails various statistical issues that are the focus of this presentation.

2. Statistical Issues

The data are for areas, yet the goal is to make inferences at the individual level. In spatial terms this is using area level data to make to make inferences at the point level. The *support* for a spatial variable is the shape, size and orientation of the area underlying the measurement. These area and point variables exist on different spatial scales with different support, and changing the support changes the statistical properties of the variables - this is known as the *change of support problem* (COSP). Gotway and Young (2002, p.634) write: "Changing the support of a variable (typically by averaging or aggregation) creates a new variable. This new variable is related to the original one but has different statistical and spatial properties." Associations observed on one spatial scale may differ from those on another scale because the variables have different properties.

2.1 Ecological Inference and the Ecological Fallacy

The effects of the changes in the statistical properties have been discussed from a couple of different, but related, perspectives. The first is that discussed under the headings of *ecological inference* and the *ecological fallacy* in the epidemiological and the social science literature, political science in particular; see for example, Morgenstern (1995), Wakefield (2008), King (1997), and Langbein and Lichtman (1978). In this literature the use of aggregate data to draw conclusions about the characteristics of individuals is known as *ecological inference* or as an *ecological study*. The *ecological fallacy* is the error in inference in assuming that associations at the aggregate level will apply to the individual level. Conclusions drawn from aggregate data may be the same as those that would be obtained from individual data, or they might be exactly the opposite, or anywhere in between, and there is no way to assess the degree of exposure to error from the aggregate data itself.

A highly influential classic critique of ecological inference was published by Robinson (1950). He examined the relationship between foreign vs. native born and illiteracy in the US with 1930 census data. The individual level correlation between being native born and literate was .12. However when calculated with data aggregated by state, the correlation becomes -.53 and with data aggregated by the nine census regions it becomes

-.62, suggesting native born are less literate than foreign born. The ecological correlations are *not only of different magnitude but are opposite in sign*, with greater levels of aggregations yielding greater differences between ecological and individual correlations.

There are two reasons for the differences obtained at the different levels of analysis and the ecological fallacy. The first reason is *aggregation bias* in which the grouping in aggregation alters the relative variance of variables which affects correlation values. Aggregation tends to smooth out variability in the data and increase the size of associations. In the following expression for the correlation r_{xy} between two variables x and y with standard deviations s_x and s_y, respectively, and covariance cov(x,y), it can be seen that decreases in the standard deviations can increase the correlation.

$$r_{xy} = \text{cov}(x,y)/(s_x s_y)$$

The second is *specification bias* or *confounding* in which grouping can crate artifacts or mask true associations due to an excluded variable associated with both variables is left out of the analysis – in the above example, foreign born tended to move to industrialized states for employment and these states had higher levels of literacy.

2.2 The Modifiable Areal Unit Problem

This issue is also discussed under the heading of the *modifiable areal unit problem* (MAUT) in the geographic and geostatistics literature; see for example Haining (2003), Gotway and Young (2004), and Waller and Gotway (2004). The general issue here is with how modifications in the geographic grouping of data affects the relationships among variables. Census and postal geographies are administratively created areas and, as noted above, the statistical properties of variables based on those areas can change if the areas are modified. In a classic study of this problem Openshaw and Taylor (1979) showed that, in analyzing the percentage of elderly and percentage of Republican voters in the 99 counties in Iowa in the 1976 election, correlations between these two variables could range from -0.97 to + 0.99 depending upon how the counties were grouped into larger districts.

The MAUP also involves two effects. The first is a *scale or aggregation effect* in grouping of data into larger areas tends to smooth out variability in the data and increase the size of association. The second is a *zoning or grouping effect* in which grouping can create artifacts or mask true associations due to confounding created by the grouping process. These correspond to the sources of bias identified in the ecological fallacy. Again, this highlights the need for caution in drawing conclusions from associations in aggregate area data about associations in individual level data.

3. Examples and Illustrations

3.1 Examples from market research data

In this section we begin by presenting examples from market research survey data to illustrate differences in conclusions from with health-related communication variables from aggregate and individual data sets. These are based on results reported in Pollard (2009). The aggregate data are data for 66 segments or area types in the Claritas PRIZM NE segmentation system. These data are from the Simmons National Consumer Study 2006 based on questionnaires and interviews from 25,000 respondents annually and are

summarized by Claritas for the segments and packaged in aggregated form with the segmentation system. The individual data are from the national Porter Novelli ConsumerStyles survey 2006 of 12,000 respondents. Table 1 shows the correlation between Hispanic ethnicity status and magazine readership in the last six months, which could be of interest for reaching Hispanic audiences. As can be seen the correlations based on the aggregate data are much larger than those based on individual data reflecting the aggregation effect, and as in the case of results from Robinson (1950) discussed above, the individual level results for the top three magazines are in the reverse direction from the aggregate level results. This reflects confounding due to grouping of Hispanic and African American population within the same segments in the creation of the segmentation system. Segments with higher concentrations of Hispanics have higher levels of readership of these magazines however it is not the Hispanics that are the readers as the individual level data show.

Table 1. Correlations Between Hispanic Status and Magazine Readership

MAGAZINE	AGGREGATE	INDIVIDUAL
ESSENCE	.77	-.05
EBONY	.77	-.06
JET	.72	-.04
PARENT'S MAG.	.54	.04
VOGUE	.51	.03

Table 2 shows the correlation between smoking cigarettes and viewing of television channels which could be of interest for reaching an audience of current smokers. Again it can be seen that the correlations based on aggregate data differ substantially from those based on individual data. In the case of magazine readership, background knowledge of the African American focus of Essence, Ebony, and Jet might make one question the association with Hispanic ethnicity even in the absence of individual level data and avoid reaching erroneous conclusions about choice of magazines to reach a Hispanic audience. However in the absence of individual data in the case of television viewing by cigarette smokers one could reach erroneous conclusions about the reach of messages placed on different channels.

Table 2. Correlations Between Smoking And Viewing of Television Channels

CHANNEL	AGGREGATE	INDIVIDUAL
COURT TV	.57	.10
LIFETIME	.44	.09
NICKLEODEON	.43	.05
SPIKE TV	.39	.12
CARTOON NETWORK	.36	.08

3.2 A Simple Illustration with Hypothetical Data

To emphasize how results from aggregate data can diverge from those that would be obtained with individual data we examine a simple example involving hypothetical data from three areal aggregates and consider the association between a risk factor and a media use variable at the aggregate area level and at the individual level. Values in the tables are percentages.

Table 3 is a cross tabulation of presence or absence of the risk factor and of non-use or use of a particular media channel in the three different areas. With aggregate data only for these areas, the only known values are the marginal values for each area as a whole. The joint values are lost in the aggregation process and are unknown, as indicated by the question marks. It can be seen from the marginal values in bold that the greater the percentage of the population with the risk factor in each area, the greater the use of the media channel.

Table 3. A Three Areal Aggregate Example – Marginal Values

		MEDIA CHANNEL		
AREA 1		No	Yes	
RISK	Yes	?	?	**30**
FACTOR	No	?	?	70
		60	**40**	100
AREA 2				
RISK	Yes	?	?	**20**
FACTOR	No	?	?	80
		70	**30**	100
AREA 3				
RISK	Yes	?	?	**10**
FACTOR	No	?	?	90
		80	**20**	100

A simple scatterplot of these values in Figure 1 shows the strong correlation in the aggregate date with r = 1.00, and this might suggest the use of this media channel for reaching this at-risk population.

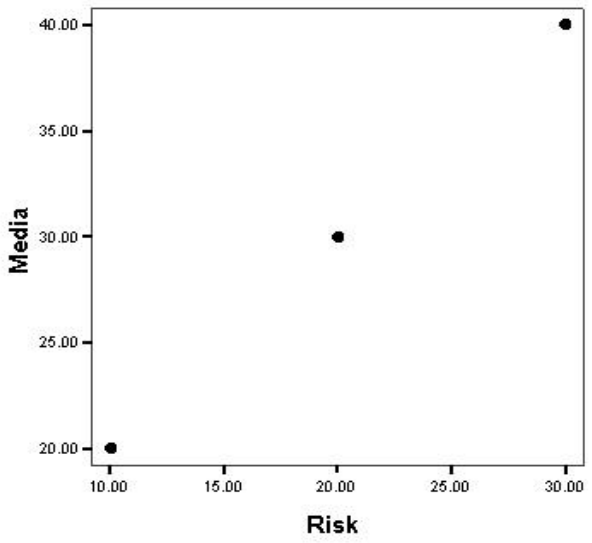

Figure 1. Correlation of Risk Factor and Media Use at the Aggregate Level (r = 1.00)

However, consider Table 4 where individual data from the areas are available and the joint distributions are known and are displayed in the cells. In this case, as indicated by

the zero values highlighted in bold, *not one individual with the risk factor uses the media channel*. Yet the marginal values are the same as in Table 3 for which the association between risk factor and media channel is extremely strong.

Table 4. One Possible Set of Joint Distributions

		MEDIA CHANNEL		
AREA 1		No	Yes	
RISK	Yes	30	**0**	30
FACTOR	No	30	40	70
		60	40	100
AREA 2				
RISK	Yes	20	**0**	20
FACTOR	No	50	30	80
		70	30	100
AREA 3				
RISK	Yes	10	**0**	10
FACTOR	No	70	20	90
		80	20	100

Table 5 shows another set of joint distributions compatible with the same marginal distribution and in this case, as indicated by the values highlighted in bold, *every single individual with the risk factor uses the media channel*.

Table 5. Another Possible Set of Joint Distributions

		MEDIA CHANNEL		
AREA 1		No	Yes	
RISK	Yes	0	**30**	30
FACTOR	No	60	10	70
		60	40	100
AREA 2				
RISK	Yes	0	**20**	20
FACTOR	No	70	10	80
		70	30	100
AREA 3				
RISK	Yes	0	**10**	10
FACTOR	No	80	10	90
		80	20	100

Thus, the correlation of r = 1.00 observed in the aggregate data is *compatible with everything from "no one" to "everyone" with the risk factor having the media channel use at the individual level.* This is just a restatement of the fact that the marginal values do not uniquely determine the joint values and many different joint distributions can be compatible with the same joint distribution. This highlights the loss of information with aggregation. Because the aggregate percentages for the areas are summary statistics for the areas as a whole, it is not possible to determine *which individuals* within the areas do or do not have the characteristics of interest. The information needed to make individual level inferences is lost in the process of aggregation.

4. Conclusion

This problem with using aggregate data from areas to draw conclusions about individual level associations was recognized over 75 years ago: "A relatively high correlation might conceivably occur by census tracts when the traits so studied were completely dissociated in the individuals or families of those traits" Gehlke & Biehl (1934, p. 170). Yet the underlying issues are not widely known outside of the fields dealing with the technical aspects of spatial analysis. It is hoped that this presentation will draw attention to the statistical issues involved.

Results obtained from the analysis of aggregate area-level data are not a simple substitute, or even an approximation for results from the analysis of individual level data. Analysts and users of aggregate data need to have an awareness and understanding of the effects of the issues discussed here and take them into account in the interpretation of relationships observed in area-level aggregate data. These established effects constitute plausible rival hypotheses in conclusions drawn from areal aggregates and need to be acknowledged as potential methodological limitations. Drawing conclusions about audiences using aggregate data is problematic and complex, and the use of individual-level data should be encouraged.

References

Curry, D.J. (1993) *The New Marketing Research Systems: How to Use Strategic Database Information for Better Marketing Decisions*. NY: John Wiley & Sons.

Drozdenko, R.G. & Drake, P. (2002) *Optimal Database Marketing: Strategy, Development, and Data Mining*. Thousand Oaks, CA: Sage Publications, Inc.

Gotway, C.A. & Young, L.J. (2002) Combining incompatible spatial data. *J of the Amer Stat Assoc*, 97: 632-648.

Gotway, C.A. & Young, L.J. (2004) A spatial view of the ecological inference problem. In King, G., Rosen, O., & Tanner, M.A. (eds.) *Ecological Inference: New Methodological Strategies,* pp. 233 – 244. NY: Cambridge Univ. Press.

Gehlke, C.E. & Biehl, K. (1934) Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association*, 29, No. 185, Supplement 169-170.

Grier, S. & Bryant, C.A. (2005) Social marketing and public health. *Ann Rev of Public Health*, 26: 319-339.

Haining, R. (2003) *Spatial Data Analysis: Theory and Practice*. Cambridge, UK: Cambridge Univ. Press.

King, G. (1997) *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data.* Princeton: Princeton Univ. Press.

Kotler, P., Roberto, N. & Lee, N. (2002) *Social Marketing: Improving the Quality of Life.* Thousand Oaks, CA: Sage Publications, Inc.

Langbein, L.I. & Lichtman, A.J. (1978) *Ecological Inference.* Newbury Park, CA: Sage Publications.

McDonald, M. & Dunbar, I. (2004) *Market Segmentation: How To Do It, How to Profit From It*. Amsterdam: Elsevier Butterworth-Heinemann.

Meyers. J.H. (1996) *Segmentation and Positioning for Strategic Marketing Decisions*. Chicago: American Marketing Association.

Morganstern, H. (1995) Ecologic studies in epidemiology: Concepts, principles, and methods. *Ann Rev of Public Health*, 16: 61-81.

Openshaw, S. & Taylor, P.J. (1979) A million or so correlation coefficients: Three experiments on the modifiable areal unit problem. In Wrigley, N. (ed.) *Statistical Applications in the Spatial Sciences*, pp. 127 –144. London, Pion.

Pollard, W.E. (2009) Statistical issues in the use of multi-source market research data for audience analysis in health communication planning. Poster presentation at Twelfth Biennial CDC Symposium on Statistical Methods, April 8, Atlanta, GA

Robinson, W.S. (1950) Ecological correlations and the behavior of individuals. *American Sociological Review*, 15: 351-357.

Sleight, P. (2004) Targeting *Customers: How to Use Geodemographic and Lifestyle Data in Your Business (3rd Ed.)*. Henly-on-Thames, UK: World Advertising Research Center, Ltd.

Sleight P. (2004) An introductory review of geodemographic information systems. *Journal of Targeting, Measurement and Analysis for Marketing*, 12: 379-388.

Wakefield, J. (2008) Ecologic Studies Revisited. *Ann Rev of Public Health*, 29: 75 -90.

Waller, L.S. & Gotway, C.A. (2004) *Applied Spatial Statistics for Public Health Data.* Hoboken, NJ: Wiley.

Weinstein, A. (1994) *Market Segmentation: Using Demographics, Psychographics and Other Niche Marketing Techniques to Predict Customer Behavior (Rev. ed.)*. Chicago: Probus Publishing Co.

Note: For further information the author can be contacted at wep11@comcast.net