Marketing Data Analysis with SAS Software

Zdzislaw Piasta - Kielce University of Technology (Poland)

Introduction

The main goal of data analysis in market research is to gather information which links an organization or firm to its market. In market research data from both primary and secondary sources are analysed. Primary data are collected and recorded for the specific research needs of a particular firm or organization. Surveing using questionnaires is the most effective method for collecting primary marketing data. Secondary marketing data are collected for general purposes (census, government reports, statistical yearbooks).

The author has participated in market research that had been performed recently in some sectors of the Polish market. Some of the surveys have been conducted as part of a Polish/American extension project. SAS software has been used for discovering market information, which is an important aid in decision making.

Analysis of data obtainable with surveys

Selected methods of primary marketing data analysis will be illustrated with data from a consumer survey for an American peanut butter. The survey has been conducted in two regions of Poland to determine the market need for this new product on the Polish market. The survey has been done around shopping areas. A sample of 113 shoppers has been interviewed. Respondents have been answering for 32 questions. The answers have been coded and used to create a SAS data set *survey*.

The SAS System offers a variety of procedures helpful in creating reports from surveys. The following statements explain how the TABULATE procedure has been used to produce reports from the peanut butter survey:

```
proc format;
value gender  1='female'
               2='male';
value educat  1='primary'
               2='secondary'
               3='university;
value region   1='GDANSK'
               2='OLSZTYN';
proc TABULATE data=survey format=7.;
table gender educat, region all*(n, pctn*f=4.1)
(income absrate relrate)*mean*f=8.2;
class gender educat region;
var income absrate relrate;
```

```
format gender gender. educat educat. region region.;
label educat='LEVEL OF EDUCATION'
      income='INCOME PER PERSON (in thous. zloties)'
      absrate='ABSOLUTE RATE ON SCALE FROM 1 TO 4'
      relrate='RELATIVE RATE ON SCALE FROM 1 TO 5';
keylabel pctn='%';
run;
```

The variable RELRATE corresponds to the question:
How would you rate the American peanut butter in comparison to other peanut butter on the market?
Five possible responses have been included with this question:
much better (5), better (4), same (3), worse (2), much worse (1).
The variable RELRATE corresponds to the question:
How do you rate the American peanut butter?
which has five possible responses:
like very much (4), like moderately (3), dislike slightly (2), dislike very much (1), no opinion (.).
The results of simple analysis performed with the TABULATE procedure are shown in Output 1.

Output 1. Analysis of answers to selected questions in the peanut butter survey

	REGION		ALL		INCOME PER PERSON (in thous. zloties)	ABSOLUTE RATE ON SCALE FROM 1 TO 4	RELATIVE RATE ON SCALE FROM 1 TO 5
	GDANSK	OLSZTYN	ALL				
	N	N	N	%	MEAN	MEAN	MEAN
GENDER							
female	30	42	72	63.7	2374.47	3.04	3.85
male	13	28	41	36.3	2486.26	3.15	3.73
LEVEL OF EDUCATION							
primary	7	2	9	8.0	1875.93	3.00	4.11
secondary	22	13	35	31.0	2218.50	3.36	3.86
university	14	55	69	61.1	2585.04	2.95	3.74

The results in Output 1 indicate that the American peanut butter is high rated by all groups of consumers specified in the table.

An important part of an analysis of data obtainable with surveys is a graphical presentation of results. Three-dimensional block charts are useful for illustrating similarities and differences among data. The blocks in a chart can be categorized by grouping and subgrouping variables. The following statements create a graph which shows mean relative rates of the American peanut butter in groups of consumers of the same gender and with the same level of education. The categories of the subgroup variable REGION identify segments which represent in each block proportions of respondents from Gdansk and Olsztyn, respectively.

```
goptions hsize=18 cm vsize=12 cm htext=0.42 cm hpos=80 vpos=45
device=amigl /*a user's modified divice similar to cgmmwwa */
gaccess='sasgastd>c:\seugi94\chart.cgm'
gsfmode=replace ;
pattern1 c=black v=empty;
pattern2 c=black v=x2;
legend frame label=('REGION:');
proc gchart data=sasuser.masloam format=4.2;
block educat    / type=mean
                  sumvar=relrate
                  group=gender
                  subgroup=region
discrete legend=legend noheading;
format educat educat. gender gender. region region.;
label educat='LEVEL OF EDUCATION';
run; quit;
```

The produced graph is presented in Figure 1.

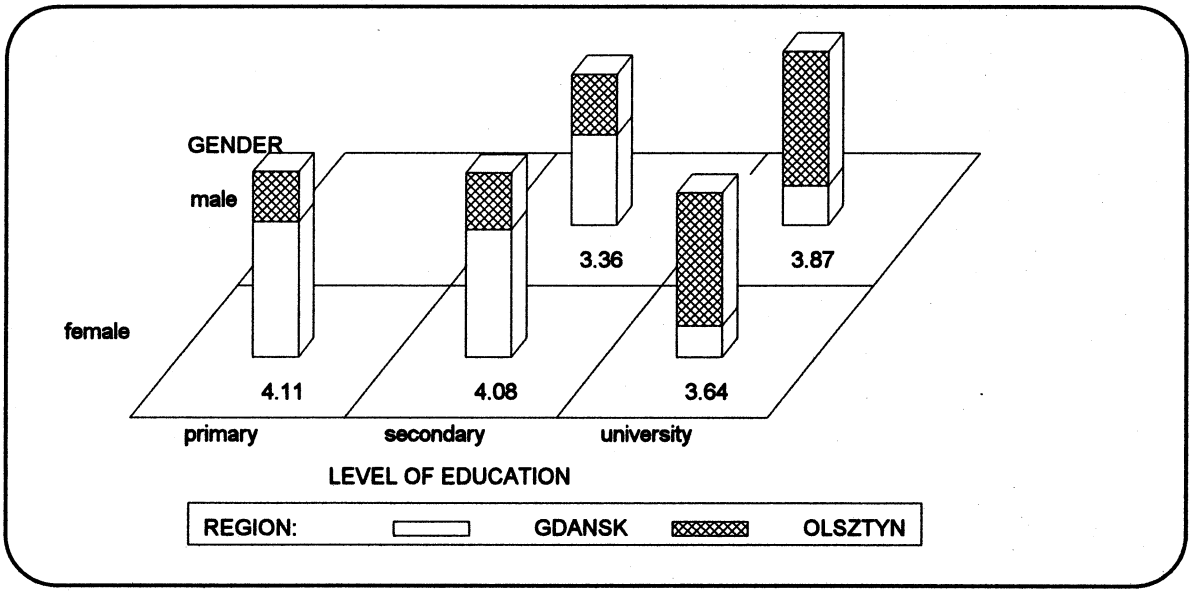

Figure 1. Mean relative rates of the American peanut butter

The statements:

```
proc gchart data=survey;
            block relrate
                 /type=percent
                  group=gender
                  subgroup=region
                  discrete
                  legend=legend
                  noheading;
                  run;
```

produce the graph shown in Figure 2.

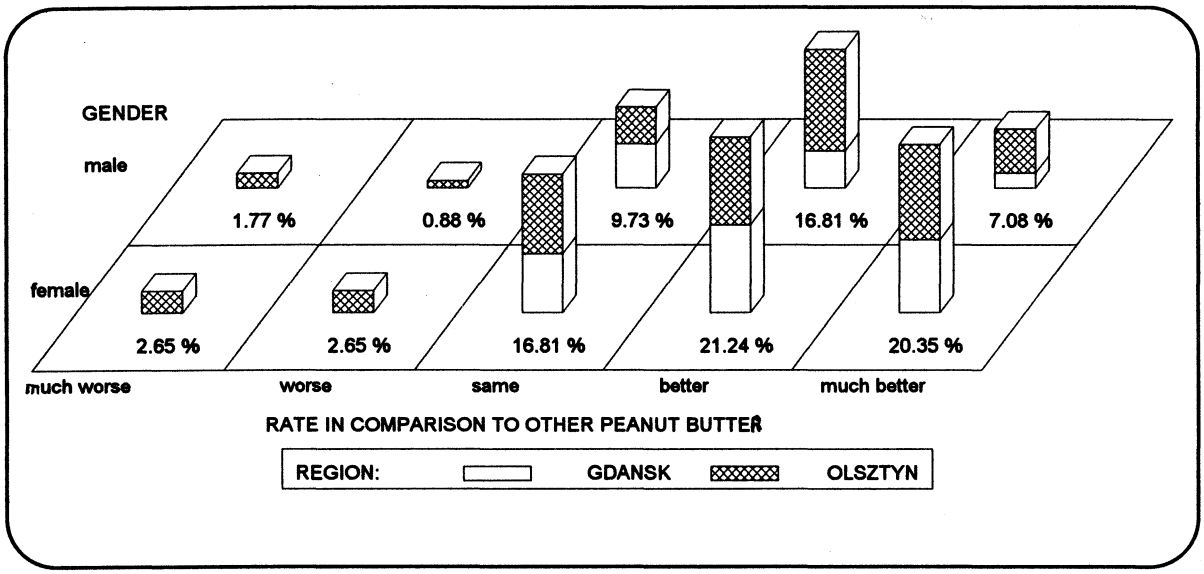

Figure 2. A percentage of different rate categories for male and female in the peanut butter survey

The block chart in Figure 2 indicates that female rates of the American peanut butter are more dispersed in comparison to male rates. All negative rates are related with respondents from Olsztyn.

For successful promotion of a new product it is important to know what is the main source of consumer's information about new food products. The variable SOURCE takes on the value 1 when a consumer gets his information while shopping. Otherwise, if the main source of consumer's information is television, radio, national and local newspapers then the variable takes on the value 0.

Figure 3 illustrates how the age of respondents, their income per person, and the accepted price for a smaller (340 g) jar of the peanut butter correlates with the variable SOURCE. The graph is created using the G3D procedure:

```
goptions hsize=20 cm vsize=15 cm htext=0.42 cm
hpos=80 vpos=45
device=cgmmwwa gaccess='sasgastd>c:\seugi94\g3d.cgm'
gsfmode=replace ;
data survey;
length shapeval $8.;
set survey;
if SOURCE=1 then shapeval='pyramid';
if SOURCE=0 then shapeval='diamond';
proc g3d data=survey;
scatter AGE*INCOME=PRICE /xticknum=6 yticknum=6 zticknum=6
shape=shapeval grid caxis=black;
run;
```

The graph in Figure 3 shows that consumers obtaining their information about new food products from television, radio, national and local newspapers are ready to accept higher price for a jar of the American peanut butter. Also older respondents are ready to pay more for the peanut butter.

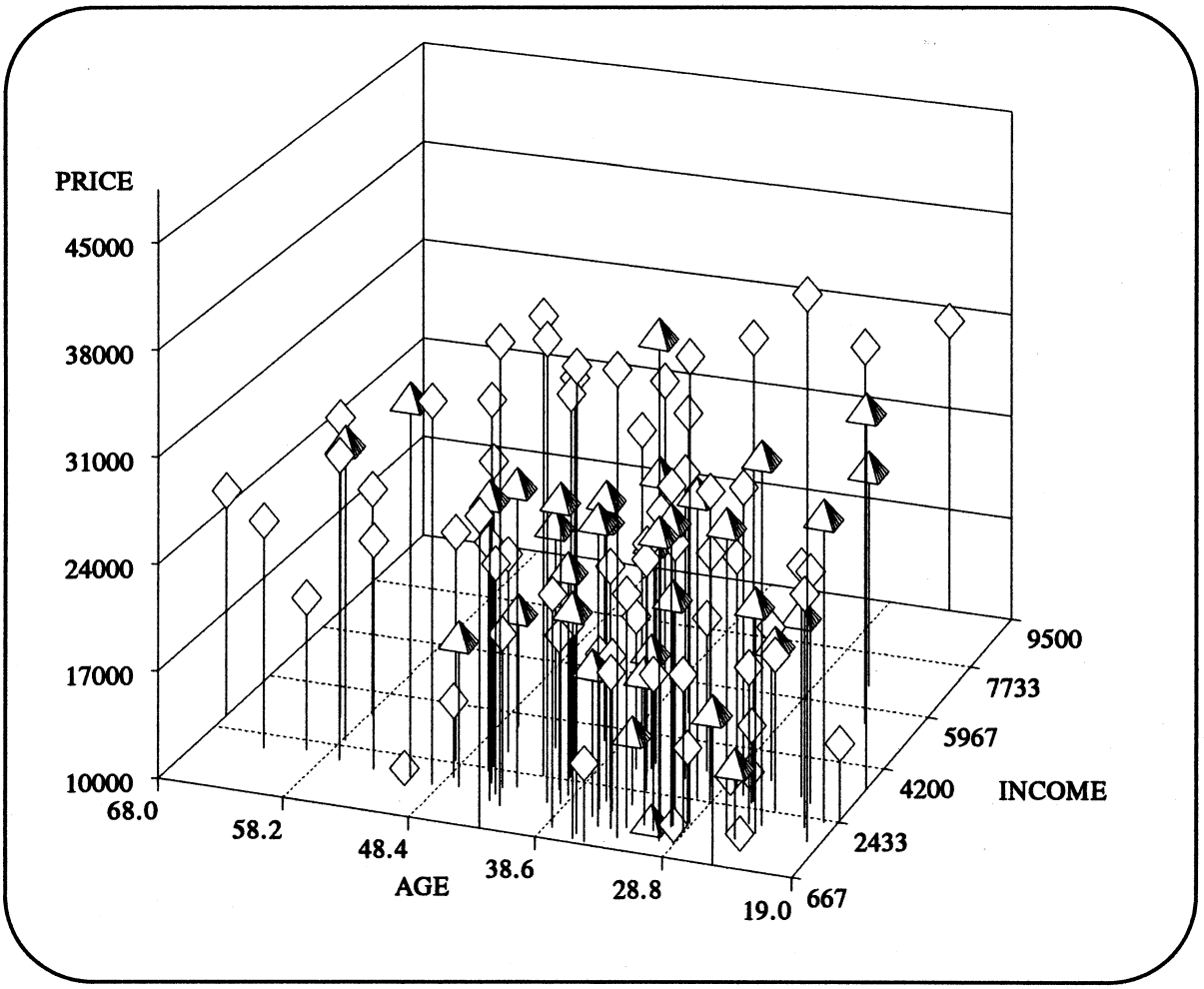

Figure 3. Relationships between the age of respondents, their income per person (in thousand zloties), the accepted price (in zloties) of the peanut butter jar, and the main source of consumer's information about new food products ('while shopping' - pyramid; 'otherwise' - diamond)

The presented examples of the analysis of data obtainable with surveys illustrate only small part of the SAS System possibilities in the domain of a market information delivery. In the next section some other tools of SAS software which are useful in an analysis of marketing data are shown.

Analysis of data obtainable from secondary sources

The role of market researcher is to obtain and analyze data from both primary and secondary sources to provide a decision maker with the information necessary to make a decision. Secondary data are important in market research since they are almost always more readily available and less expensive to obtain than primary data. Secondary market information may help the researcher to gain a background on the problem.

In Poland one of the major sources of secondary data is a statistical yearbook. In the yearbook there are a lot of data characterizing demographic and socioeconomic situation in all 49 provinces of Poland.

An example of an analysis of selected variables determining potential "richness" or "poverty" of a particular province is presented. Six variables are used in the analysis:
URODZ - birth rate,
BEZROB - unemployment rate,
EMERENT - rate of pensioners,
DZIETN - average number of children in a family,
LOPSPOL - social welfare rate,
STARSI - rate of persons, age 65 and more.

To explore relationships among these variables principal component analysis has been performed. Principal component analysis can be used for creating new agregated and uncorrelated variables. Each new variable accounts for the greatest possible variance in primary data. These synthetic variables are frequently useful in other analyses.

The input variables have been standardized across all elements of each particular variable X using the range of the variable as the divisor. The values s_i of the new variable S are computed as (see [1]):

$$s_i=(x_i- MIN(X)) / (MAX(X)-MIN(X)).$$

The variables obtained after standardization are denoted as SURODZ, SBEZROB, SEMERENT, SDZIETN, SLOPSPOL, SSTARSI, respectively. The covariance matrix has been used to compute the principal components, because the standardized variables take on the values on the same scale bounded by 0 and 1. Figure 4 shows the results obtained with SAS/INSIGHT software. The results indicate that the first two principal components generated from the covariance matrix explain over 70 percent of variance. Figure 4 shows also the first three eigenvectors

Eigenvalues (COV)

Component	Eigenvalue	Difference	Proportion	Cumulative
PCV1	0.1366	0.0322	0.3998	0.3998
PCV2	0.1043	0.0406	0.3055	0.7053
PCV3	0.0638	0.0364	0.1867	0.8920
PCV4	0.0273	0.0185	0.0800	0.9720
PCV5	0.0088	0.0081	0.0258	0.9978
PCV6	0.0008	.	0.0022	1.0000

Eigenvectors (COV)

Variable	PCV1	PCV2	PCV3
SURODZ	0.3328	0.4895	-0.2003
SBEZROB	0.4932	0.0837	0.3269
SEMERENT	-0.4540	0.2814	0.3351
SDZIETN	0.2421	0.6829	-0.2122
SLOPSPOL	0.3380	-0.0189	0.7788
SSTARSI	-0.5168	0.4555	0.2986

Figure 4. Results of principal component analysis

of the covariance matrix. The eigenvectors are used to form principal components as linear combinations of the standardized and centered to mean 0 variables. The new aggregated variable PCV1 has high positive loadings with the variables SBEZROB and SLOPSPOL, which reflect real poverty. The variable PCV2 takes on the highest values for the regions with a high proportion of the youngsters and children, but also with a high proportion of the elders.

The first two principal components can be used for visualization of the potential "richness" or "poverty" of provinces. Figure 5 shows the positions of the Polish provinces (identified by their codes) on the plane spanned by PCV1 and PCV2. This plane gives the best possible fit to the primary data points as measured by the sum of squared perpendicular distances from each data point to the plane.

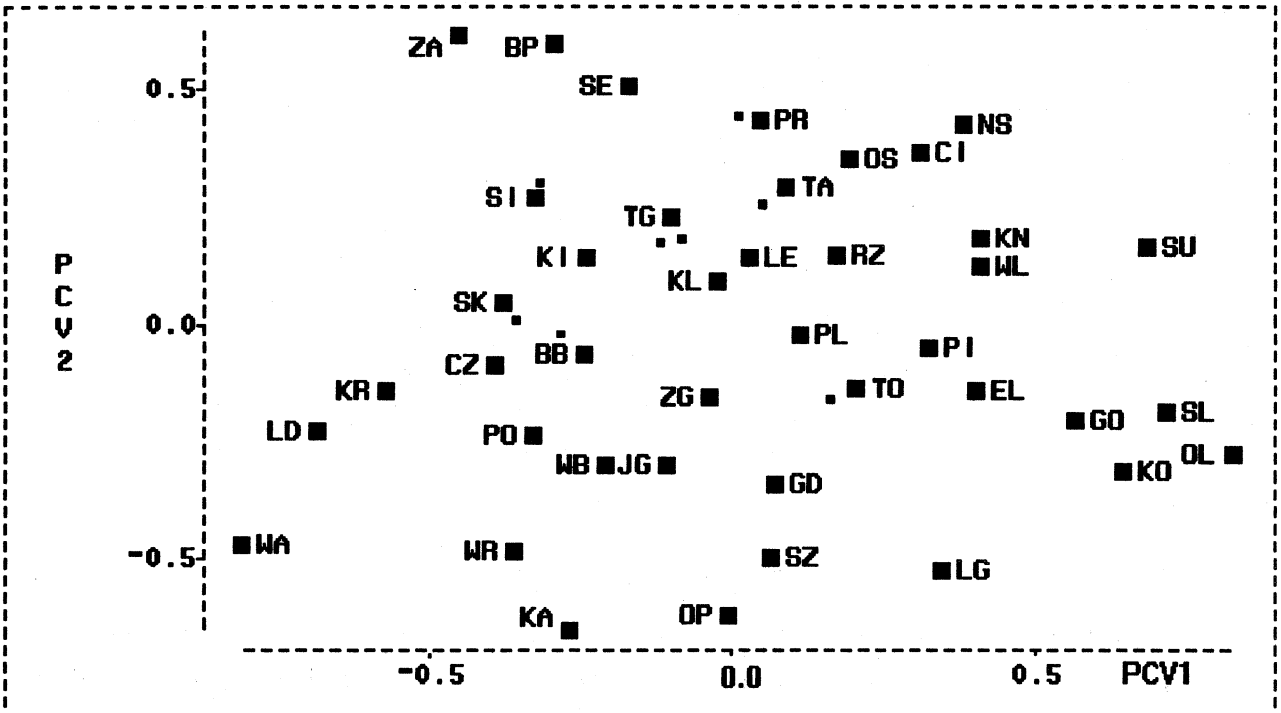

Figure 5. Scatter plot of the first two principal components

An alternative form of visulization of multivariate data is a star plot. It plots each province as a star figure with one ray for each variable. Ray lengths are proportional to the values of that variable (see [2]). The star plot has been created with the macro STARS [3]. The shape of each star reflects a situation in the respective province. Provinces with a higher level of potential "poverty" are identified by greater stars. The horizontal right rays correspond to the variable BEZROB. The variables LOPSPOL, EMERENT, STARSI, URODZ, DZIETN are ordered around the star starting from the ray corresponding to BEZROB in opposite to clockwise direction. The star plot for 49 Polish provinces is presented in Figure 6.

SAS/GRAPH software offers a variety of forms for visulization of variables on two- and three-dimensional geographic maps. The block map has been used to show how the values of the variable AVINCOME (average monthly income before taxes) relate to the Polish provinces. The following SAS statements produce the plot shown in Figure 7.

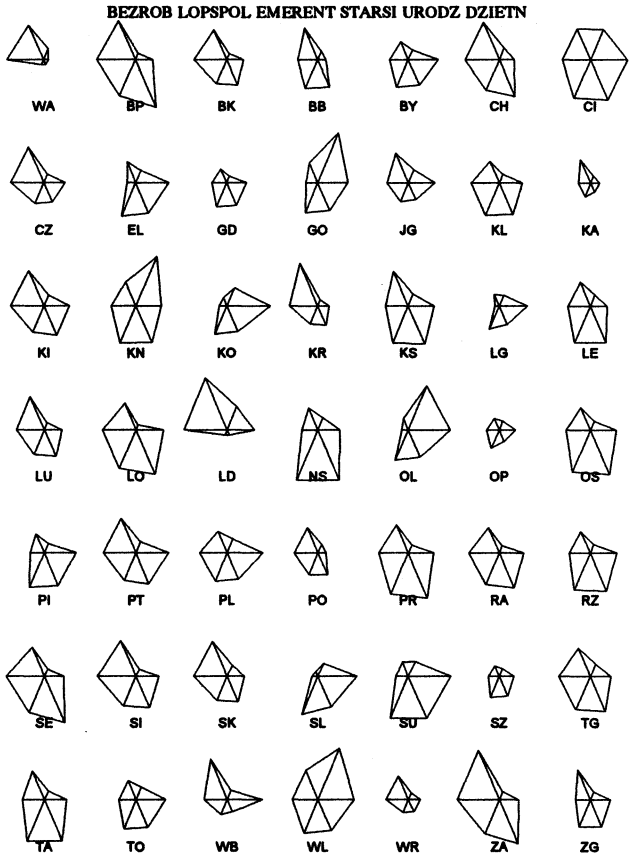

Figure 6. Star plot for the Polish provinces

```
pattern1 c=black v=mempty;
pattern2 c=black v=x2 repeat=49;
goptions hsize=18 cm vsize=16 cm hpos=80 vpos=45
device=amigl gaccess='sasgastd>c:\seugi94\mapblock.cgm'
gsfmode=replace ;
proc gmap data=yearbook map=sasuser.POLAND;
id id;
block AVSALARY /discrete zview=6 nolegend;
run; quit;
```

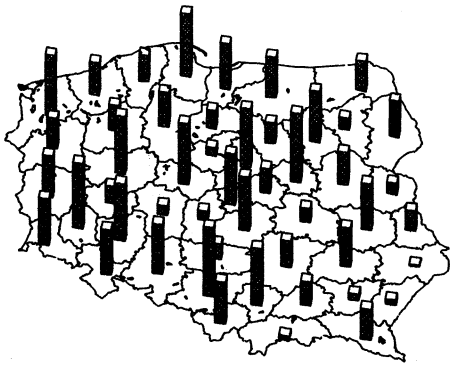

Figure 7. Average monthly income before taxes in the Polish provinces

The choropleth map has been used to show how the values of the variable AVRELPNS (relative average retiring pension before taxes) are associated with the Polish provinces. The following SAS statements produce the map shown in Figure 8.

```
pattern1 c=black v=m3n0;
pattern2 c=black v=m1n0;
pattern3 c=black v=e;
pattern4 c=black v=m1n90;
pattern5 c=black v=m3n90;
pattern6 c=black v=m5n90;
goptions hsize=15 cm vsize=18 cm htext=0.42 cm hpos=80 vpos=45
device=amigl gaccess='sasgastd>c:\seugi94\mapchoro.cgm'
gsfmode=replace ;
proc gmap data=yearbook map=sasuser.POLAND;
id id;
choro AVRELPNS  /levels=5;
run; quit;
```

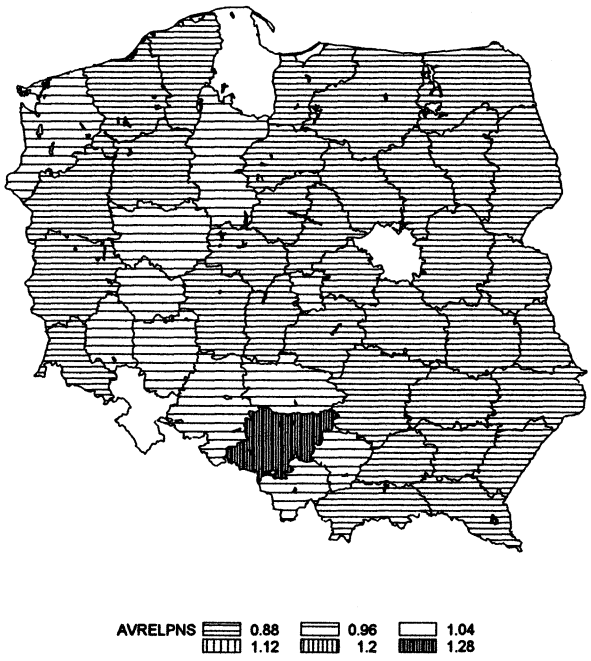

Figure 8. Average relative retiring pension before taxes in the Polish provinces

Summary

SAS software offers a variety of powerful analytical and graphical tools for tranforming primary and secondary marketing data into market information. Results of analyses of data obtainable with surveys are helpful in discovering behaviour patterns of clients and consumers. Results of analyses of secondary source data are useful in determining the strategy for a firm or organization. The SAS System delivers information in an efficient fashion and in a form which is relevant to market decisions.

Acknowledgements

I am grateful to Professor Bill Miller from the University of Georgia and to Dr. DeeVon Bailey for inspiration. Thanks are due to Dr. Stanislaw Pilarski from Olsztyn and Dr. Daniel Roszak from Gdansk for the supply of the data from the peanut butter survey.

This work was partially supported by the State Committee for Scientific Research (grant no. 8 S503 033 06) and by the internal grant of the Kielce University of Technology.

References

1. G. W. Milligan and M. C. Cooper, A study of variable standardization, Working Paper Series 87-63, College of Business, The Ohio State University, May 1987.
2. J. M. Chambers, W. S. Cleveland, B.Kleiner and P. A. Tukey, Graphical Methods for Data Analysis, Wadsworth, Belmont, 1983.
3. M. Friendly, SAS System for Statistical Graphics, First Edition, SAS Institute Inc., Cary, NC, USA, 1991.

Trademark citations

SAS, SAS/GRAPH and SAS/INSIGHT are registered trademarks of SAS Institute Inc., Cary, NC, USA.

Any comments will be appreciated:
Dr. Zdzislaw Piasta
Kielce University of Technology
25-314 Kielce, Al. 1000-lecia P. P. 5
POLAND

e-mail: mat-zp@srv1.tu.kielce.pl