# **IBM Research Report**

## KMAP - A Visualizer for Kohonen Self-organizing Map Clustering Results

## George S. Almasi, Richard D. Lawrence

IBM Research Division Thomas J. Watson Research Center P.O. Box 218 Yorktown Heights, NY 10598



Research Division Almaden - Austin - Beijing - Delhi - Haifa - India - T. J. Watson - Tokyo - Zurich

LIMITED DISTRIBUTION NOTICE: This report has been submitted for publication outside of IBM and will probably be copyrighted if accepted for publication. It has been issued as a Research Report for early dissemination of its contents. In view of the transfer of copyright to the outside publisher, its distribution outside of IBM prior to publication should be limited to peer communications and specific requests. After outside publication, requests should be filled only by reprints or legally obtained copies of the article (e.g., payment of royalties). Copies may be requested from IBM T. J. Watson Research Center , P. O. Box 218, Yorktown Heights, NY 10598 USA (email: reports@us.ibm.com). Some reports are available on the internet at <a href="http://domino.watson.ibm.com/library/CyberDig.nsf/home">http://domino.watson.ibm.com/library/CyberDig.nsf/home</a>.

## KMAP - A Visualizer for Kohonen Self-organizing Map Clustering Results

George S. Almasi and Richard D. Lawrence IBM T.J. Watson Research Center P.O. Box 218 Yorktown Heights, New York 10598 almasi@us.ibm.com, ricklawr@us.ibm.com

August 17, 2002

#### Abstract

We describe a Java-based visualizer that preserves the topological information inherent in the results of cluster data mining that used the Kohonen self-organizing map approach. The visualizer provides a tool with a number of views and drilldowns that an analyst can use to circumvent the data interpretation bottleneck and quickly evaluate the results of the map produced by the datamining run. Being written in Java, it is portable across a large number of computing platforms.

**Keywords:** visualizer, data mining, Kohonen, self-organizing map, clustering, Java.

## 1 Introduction

The Kmap visualizer described in this paper was designed to allow rapid evaluation of the results of data mining performed using the Kohonen self-organizing map [8] clustering approach. A key feature of this approach is that, in addition to grouping similar data records into clusters, it places the clusters on a two-dimensional map in such a way that similar clusters are near each other. The nature of a cluster in a particular neighborhood of the map says something about that neighborhood. In other words, the contents of individual clusters and the topography of the cluster map are both of interest.

The current authors wrote a highly scalable parallel version of the self-organizing map algorithm [9] which is used by IBM's Intelligent Miner for Data [7]. This program was recently put to the test in a demonstration of mining very large databases [4], and the parallel speedup that was obtained was so good (over 14X for a 16-processor SP2) that the normal ratio of time needed to perform the datamining vs. interpreting its results became inverted – data analysis became the new bottleneck. The reason was that the existing visualizer was good at conveying the nature of individual clusters, but the map had to be constructed basically by hand. Kmap was designed to remedy this problem.

This paper is organized as follows: we first describe the design of the Kmap visualizer, its modes of operation and its drilldown capabilities, and its implementation in Java. This is followed by four case studies that show examples of Kmap's features. We then present illustrative examples of Kmap's features in the form of four case studies using stock market, retail supermarket, credit bureau, and census data.

## 2 Implementation

The main Java classes in Kmap reflect the program's functionality:

- **Kmap** collects the clustering results from the input file (generic neural weights file or Intelligent Miner results file) and creates and manipulates the rectangular grid map of clusters.
- Kdetail handles the drilldowns into an individual cluster
- **Kplot** creates the contents of the rectangles for Kmap and Kdetail, using the JClass chart and table java components provided by KL Group Inc. [5].
- **Kmenubar** handles user interactions via the menu bar that is used for both Kmap and Kdetail.
- **NNClustRes** provides an API to Intelligent Miner results files in the form of a dozen "get" methods for obtaining
  - 1. the number of clusters (neural nodes) and data attributes (fields),
  - 2. the neural weights (mean clustering attributes for each node), and
  - 3. several sets of attribute statistics, used for drilldowns and described further below.

There are three different "modes" for viewing clustering results in Kmap and its drill-downs:

- "weights" mode, in which a plot of all the neural weights (i.e., attribute mean value) is shown for each cluster. This mode is useful when the attributes consist of a logical sequence like month-by-month prices or spending. Figure 1 shows such an example.
- "topten" mode, in which the names of the attributes with the top 3 or top 10 weights are shown for each cluster. This mode can be very useful when the number of attributes is large. Figure 3 shows a comparison of results viewed in weights and in topten modes, and Figure 4 gives a detailed example of topten mode in action.
- "sizes" mode, which simply shows the size of each cluster as a filled circle. Figure 3 contains an example of a "sizes" view.

The mode can be selected from the "View" menu, or by specifying w, t, or s after the filename when Kmap is called from the command line.

The cluster size is indicated by the background color in weights mode, by the grayness of the frame in topten mode, and by the size of the circle in sizes mode.

The available supplemental attributes, such as Total Spending or Age, can also be used to color the inner rectangle in topten mode or the background in weights or sizes mode. This is done via the "Color By" pulldown on the menu bar. The functions provided by the other menu bar pulldowns are as follows:

- "File" handles the opening of files and the printing of the displays.
- "View" provides a large number of methods for controlling the appearance of the visualization.
- "Sort By" allows sorting the attributes in topten mode by weights or by six other quantities.

The attributes used in neural clustering are often normalized first, and so the neural weights (the means of the normalized attributes) for each cluster can be different from the means of the raw values. The un-normalized values of the attributes can be included as supplemental fields, and Intelligent Miner will compute their statistics along with those of the active fields. Kmap has provision to read these supplemental field statistics and make them available for alternative ways to sort the attributes in topten mode, using the "Sort By" menu. These quantities available for each cluster are:

- 1. "mean", the mean value of the un-normalized attribute
- 2. "partcp", the fraction of records with non-zero values of this attribute
- 3. "meanNZ", the mean non-zero value of the un-normalized attribute, computed from the ratio of the two quantities above
- 4. "Rmean", the mean value of the un-normalized attribute for this cluster divided by the mean value of this attribute for the entire data set.
- 5. "Rpartcp", the quantity "partcp" normalized to the entire data set as above
- 6. "RmeanNZ", the quantity "meanNZ" normalized to the entire data set as above

There are also popup menus and other panels available for drilldown and other purposes; these are shown in the examples that follow.

### 3 Case Studies

This section describes the use of Kmap to visualize several different kinds of datamining results.

- Standard and Poor 500 monthly closings (weights, drilldowns)
- Safeway (topten mode, more drilldowns)
- Credit Bureau(bankruptcy colorby)
- Census (categorical variables)

#### 3.1 Stock Prices

For our first example, Figure 1 shows Kmap in "weights" mode, displaying the results of a 64-node clustering performed on the monthly closing prices of the Standard & Poor's stocks for the period September 1995 - August 1996. Each stock price is normalized by its maximum value during this period. The action of the Kohonen algorithm in placing similar clusters near each other on the map shows clearly in the gradual transition from gainers on the left to decliners on the right.

Kmap offers two special drilldowns for relatively small datasets like this one. It can read a file listing the cluster for each record and display the membership of each cluster on a popup menu, as shown in Figure 1. In addition, the "View" menu has an entry that leads to a "Record Finder" panel that can find which cluster a particular record is in and highlight that cluster, as shown for the case of Western Atlas and cluster 25 in the figure. A hypothetical scenario for using these capabilities would be to check the membership of a gainer cluster like node 32, find that it contains two petroleum industry members (Mobil and Texaco), wonder where Western Atlas (also a petroleum industry member) had been placed, use the Record Finder, and discover it in cluster 25, a direct neighbor of cluster 32. There is a substantial number of such intuitively satisfying relationships in the map.

If the cluster membership file mentioned above also contains the attribute values for the individual records, Kmap can provide an additional drilldown that allows comparing the behavior of individual records in a cluster, as shown in Figure 2.

#### 3.2 Supermarket Shoppers

The example in this section explains the motivation for "topten" mode and describes some of the features of that mode. The clustering in this case was performed on a month's spending data by some 36,000 supermarket customers in about 100 product categories (dairy, confectionery, baby products, etc.). As shown in Figure 3, the Kmap "weights" mode used in the previous example can be used to view these clustering results as well. Kmap provides point labels to help identify each point even in a crowded plot, and this feature can be used to identify the three attributes with the largest weights in cluster 3 as Sugar, Tea, and Desserts/Puddings, for example. But the large number of attributes makes this a slow process. Furthermore, unlike the months of the previous example, there is no sequential relationship among the attributes that makes a plot view particularly useful. A much faster grasp of the clusters' nature can be obtained in this case from an initial map that shows what the top 3 attributes are for each cluster, with a rich set of available drilldowns for further investigation. This is what "topten" mode does.



Figure 1: Monthly closing prices for the Standard & Poor's 500 stocks over a one-year period



Figure 2: Drilldowns available in weights mode (for example in Figure 1)

In topten mode, each rectangle in the map conveys the following information about its corresponding cluster: the outer frame's degree of grayness is proportional to its population. the inner rectangle's color indicates the value of the supplemental attribute chosen from the "Color By" menu. In Figure 3, for example, blue indicates above-average customer Age, while pink denotes below-average attribute values. The numerical values of the population and Age are also shown for each cluster.

By default, the top 3 attributes are sorted by their neural weights, but this can be changed (for the main map as well as the drilldowns) via the "Sort By" menu, which allows re-sorting by the attribute mean value or the five other statistics mentioned earlier. The color of the bullets in this example correspond to a rainbow spectrum matched to the range of category numbers, so attributes with similar colors have similar category numbers. The bullet shapes carry no information in this example. However, for the wine recommender described in [3], we used a different scheme in which the bullet color indicated the wine color and the bullet shape indicated price (squares cost more than circles, etc.).

Clicking on a cluster's rectangle pops up a menu that can be used for further drilldown, as shown in Figure 3. The first item on the menu allows getting more information on a particular attribute (field), such as the distribution of its revenue among the clusters. For example, the 11% of the total customer population represented by cluster 3 accounted for 24% of the spending on Tea, for a "wallet share enrichment" of 2.18 - i.e., chances of finding a tea buyer in this cluster are more than twice that for the overall customer population.

The second item on the drilldown menu allows getting more detail on a particular cluster. Clicking it brings up a scrollable table showing the neural weights and other field statistics for all the attributes. For convenience, the distribution map for a given attribute can be brought up by clicking the attribute in this table.

The third item on the drilldown menu brings up an enlarged plot of the neural weights for that cluster, as shown in the figure.

The third Kmap view, "sizes", is shown at the top left of Figure 3 for reference. In this mode, the cluster population is conveyed by the size of the corresponding circle (the populations are very close in this case), and the color convention is the same as for the "topten" view.

The drilldowns shown are available in all 3 modes, and each drilldown figure can be printed individually.



Figure 3: Comparison of views and drilldowns. Starting at top left, the "sizes", "weights", and "topten" views of the supermarket shopper clustering results for 9 clusters. The top right view also shows popup menu leading to several drilldowns: into a specific attribute (left center and right), into the weights view of a cluster (right center), and into the topten view of a cluster (bottom right). Clicking an attribute's button in the view at bottom right is a second way to obtain the attribute map at bottom left.



Figure 4: Supermarket retail data clustering results

One application of this clustering capability is in the generation of personalized recommendations, as we describe in [2]. These recommendations were being generated for participants in a trial program using PDAs(personal digital assistants) for supermarket shopping. Since all participants had above-average spending <sup>1</sup>, we clustered only the above-average spenders and obtained the results shown in Figure 4. Note the emergence of a cluster (4) with strong spending on baby products. As the three drilldown maps of attribute share show, these clusters are sharper than the previous ones in that the "wallet share enrichments" are larger. For example, cluster 4 has 10% of the total population but 56% of the total spending on baby products, for a an enrichment of 5.6. This sort of enrichment can be quite valuable for customer relationship and marketing purposes. It also impacted the generation of recommendations. For example, the most popular chocolates in cluster 4 turn out to be quite different from those most frequently preferred by the population as a whole. [2].

#### 3.3 Credit Data

The example in this section shows the value of adroitness in the selection of the attributes used for clustering. Figure 5 shows clustering results from a study on mining a large database from a credit bureau. Twelve of the 44 attributes that were used were created by categorizing each outstanding credit card balance into a small matrix of several ranges of balance and several ranges of the age of the account. This helped the domain expert to find about a dozen different segments on the map. For example, the top attributes for clusters 18 and 28 are "BANK\_BAL03" and "BANK\_BAL06", which both correspond to large balances in accounts less than 6 months old, hence the analyst's label "BIG CHURN" for people who continually accept bank offers for new credit cards with low initial interest rates, and then transfer to another card at the end of the low interest rate period. This ability to discern various segments of credit usage can be very helpful to a bank's efforts to market its various products.

The second example in this section shows the clustering results on a subset of the population whose credit rating is within a "gray area" that normally makes it difficult to obtain credit. One of the supplemental attributes (not used in the clustering, of course) was whether an individual in this subset actually went bankrupt within the next six months. When this attribute is used to color the cluster map (Figure 6), an interesting pattern appears: far from being randomly distributed, the actual bankruptcies are concentrated on the right edge of the map. The bankruptcy rate for the population as a whole is about 0.5%, and most of the clusters have a rate below this value, while one of the clusters at the right of the map has an 8% bankruptcy rate, which is 16 times higher than the average value. The dominant attributes of this cluster are all old accounts with very large balances, which tend to be prominent in the other clusters with above-average bankruptcies as well. The suggestion is that a number of people are being denied credit without good reason, and that a more refined rating system could free up credit to a group that could probably really use it. Customers and the bank would both benefit.

<sup>&</sup>lt;sup>1</sup>like the children in Lake Wobegon

credit100 Kmap -	grayest = 30% of 10	73960 records, sort	ed by weights, colo	red by BEACON_96_SCF	R (mean=633.494)				_ 🗆 ×
File View Color B	y Sort By Exit	2 693 1 138 (1	2 690 89 22K	727 19 30K(3%)	5 725 88 34K	5 723 64 26K	7 727 16 181	8 701 01 81	677 99 27
BANK_BAL11	HOMEFURN_BA	OCREDITU_BAL	BANK_BAL11	OMORT_BAL	OMORT_BAL	OMORT_BAL	OMORT_BAL	OMORT_BAL	OMORT_BAL
INSTFIN_BAL	HOME_BAL	AUTO_BAL	BANK_BAL10	BANK_BAL11	AUTO_BAL	BANK_BAL11	BANK_BAL11	BANK_BAL11	BANK_BAL11
BANK_BAL12	♦ ELEC_BAL	◆ MORT_BAL	BANK_BAL12	OTHDEPT_BAL	SANK_BAL11	OTHDEPT_BAI	SANK_BAL12	◆AUTO_BAL	♦ AUTO_BAL
10 561.92 1608(	11 575.52 6223 (	12 661.06 27K	13 703.63 14K	14 659.36 2232(0%)	15 666.58 4541	16 694.96 14K	17 697 4349 ((	18 689.5 414:	19 684.04 39
OTHOTIL_BAL	STUDENT BAL	BANK BAL11	BANK_BAL08	BANK BAL11	MORT BAL	AUTO BAL	BANK BAL03	BANK BALOS	BANK BAL12
AUTO_BAL	OTHDEPT_BAL	POPDEPT_BAI	BANK_BAL11	MORT_BAL	BANK_BAL11	BANK_BAL11	♦ MORT_BAL	♦ MORT_BAL	BANK_BAL11
20 593.14 1820(	<b>21</b> 576.49 2296 (	22 712.44 11K	23 692.11 5586	<b>24</b> 733.6 7103(0%)	<b>25</b> 638.43 3162	26 651.77 4613	<b>27</b> 688.73 16	28 687.38 18:	<mark>29</mark> 688.92 43:
GOV_BAL	JEWELRY_BAL	BANK_BAL12	BANK_BAL12	JUMBO_BAL	OREVFINAN_BAI	MORT_BAL	BANK_BAL03	BANK_BALO	BANK_BAL12
MORT_BAL	AUTO_BAL	BANK_BAL11	MORT_BAL	CHARGECARD_BA	BANK_BAL11	BANK_BAL01	MORT_BAL	MORT_BAL	BANK_BAL11
30 647.16 13K(1	<b>31</b> 659.73 9814(	32 636.55 21K	<b>33</b> 587.68 458(	34 643.51 17K(2%)	<b>35</b> 608.55 7118	<b>36</b> 571.51 1916	37 614.82 28	<b>38</b> 640.46 11:	<b>39</b> 622.64 18
STUDENT_BAL	OINSTFIN_BAL	POPDEPT_BAI	OILCARD_BAL	BANK_BAL01	BANK_BAL01	BANK_BAL01	BANK_BAL02	BANK_BALOS	BANK_BAL12
INSTFIN_BAL	MORT_BAL			BANK_BAL04	BANK_BAL04	BANK_BAL04	BANK_BAL01	BANK_BALO	BANK_BAL11
BANK_BALIU	AUTO_BAL	V AUTO_BAL	V POPDEPT_BAI	V DANK_BALUZ	BANK_BALUZ	BANK_BALUS	V DANK_DALU	V DANK_DALU	BANK_BALUC
OSTUDENT BAI	07.79 2162 (	OSECUR BAI	OIL CARD BAL	OIL CARD BAL	BANK BAL04	BANK BAL 07	BANK BAL02	BANK BALOS	BANK BALOS
INSTFIN_BAL	MORT_BAL	INSTFIN_BAL	OTHDEPT_BAI	POPDEPT_BAL	BANK_BAL01	BANK_BAL04	BANK_BAL05	MORT_BAL	BANK_BAL05
BANK_BAL11	AUTO_BAL	AUTO_BAL	BANK_BAL11	BANK_BAL11	BANK_BAL07	BANK_BAL08	♦ MORT_BAL	BANK_BALO	MORT_BAL
<b>50</b> 638.2 5578(0	<b>51</b> 658.7 2911(0	52 673.48 43K	53 592.15 16K	54 585.82 2455 (0%)	55 572.57 6432	56 592.4 11K(1	<b>57</b> 678.22 79:	58 656.61 54	<b>59</b> 621.92 24 <sup>*</sup>
STUDENT_BAL	SECUR_BAL	OAUTO_BAL	BANK_BAL10	BANK_BAL10	BANK_BAL10	BANK_BAL07	BANK_BAL05	BANK_BALO	BANK_BAL02
BANK_BAL11	AUTO_BAL	BANK_BAL11	OTHDEPT_BAI	POPDEPT_BAL	POPDEPT_BAL	BANK_BAL08	BANK_BALO	BANK_BALO	BANK_BALO
<b>60 595.68</b> 3060 (	61 648.65 13K(1	62 599.59 582¢	63 625.28 14K	64 605.9 1443(0%)	65 646.42 17K(	66 643.55 5995	67 666.35 94!	<mark>68</mark> 640 6698((	<mark>69 620.21</mark> 30
OAUTO_BAL	OAUTO_BAL	OAUTO_BAL	OTHDEPT_BAI	CLOTHING_BAL	BANK_BAL11	BANK_BAL11	BANK_BAL05	BANK_BALO	BANK_BAL05
BANK BAL11	MORT BAL	OTHDEPT_BAL	AUTO BAL	BANK BAL11	BANK BAL10	BANK_BAL08	BANK_BAL08		BANK_BAL02
70 635.64 3146(	74 680.76 9648(	<b>72</b> 619.82 16K	78 610.86 9480	<b>74</b> 600,93 4763(0%)	<b>75</b> 608.83 4286	76 620.66 10K	77 639.46 77:	78 625.23 63	79 611.89 364
OCREDITU_BAL	CREDITU_BAL	POPDEPT_BAI	OTHDEPT_BAI	OTHDEPT_BAL	BANK_BAL11	BANK_BAL11	BANK_BALOS	BANK_BALO	BANK_BALO
AUTO_BAL	AUTO_BAL			AUTO_BAL		MORT_BAL	BANK_BAL11	BANK_BALO	BANK_BALOS
MORI_BAL	WORT_BAL	BANK_BAL11	BANK_BAL11	WORT_BAL	VPOPDEPT_BAL		WORT_BAL	BANK_BAL11	BANK_BAL11
630.97 4319(	607.99 7956 (	612.58 10K	S85.21 4044	588.18 3649(0%)	581.69 1657	BANK BAL 11	610.58 44'	BANK BALOS	BANK BALOS
HOMEFURN_BA	HOME_BAL	BANK_BAL11	OTHDEPT_BAI	BANK_BAL11	BANK_BAL11	BANK_BAL12	BANK_BALO	BANK_BAL11	BANK_BALOS
BANK_BAL11	♦ AUTO_BAL	MORT_BAL	BANK_BAL11	POPDEPT_BAL	BANK_BAL10	BANK_BAL10	BANK_BAL05	BANK_BALO	BANK_BAL02
90 581.72 2373 (	91 553.64 1371(	92 604.53 4944	93 597.69 3701	94 568.35 1725(0%)	95 471.49 431(	96 549.62 1553	97 609.34 24	98 592.03 19	99 564.5 116
HOMEFURN_BA	HOMEFURN_BA	POPDEPT_BAI	POPDEPT_BAI	OTHDEPT_BAL	OAUTO_BAL	BANK_BAL11	BANK BAL11	BANK_BALO	BANK_BALOS
OTHDEPT_BAL	HOME_BAL	OTHDEPT_BAL	OTHDEPT_BAI	BANK_BAL11	OTHDEPT_BAL	POPDEPT_BAI	BANK_BALOS	BANK_BALO	BANK_BALIC

LOW			OLD	Stodgy Mortgage	Stodgy Mortgage	Stodgy Mortgage	Stodgy Mortgage	Stodgy Mortgage	Stodgy Mortgage
	STUDENT	Shoppers	Medium 1-2 year	Stodgy Mortgage	Stodgy Mortgage	AUTO	BIG CHURN	Stodgy Mortgage	Stodgy Mortgage
STUDENT				JUMBO	Stodgy Mortgage	Little Churn	BIG CHURN	Small Business	Small Business
GOVMT			OIL	Little Churn	Little Churn	Little Churn	MEDIUM CHURN		
GOVMT		Popular Dept	OIL	OIL	Little Churn		MEDIUM CHURN	MEDIUM CHURN	MEDIUM CHURN
GOVMT		AUTO					MEDIUM CHURN	MEDIUM CHURN	MEDIUM CHURN
AUTO	AUTO	AUTO	Serious Shoppers		Medium OLD	Medium OLD	Medium 6-12 months	Medium 6-12 months	Medium 6-12 months
AUTO	AUTO	Popular Dept	Serious Shoppers	Serious Shoppers	Ser Shop / Medium OLD	Medium OLD	Medium 1-2 year	Medium 1-2 year	Medium 1-2 year
		Popular Dept	Serious Shoppers	Serious Shoppers	Serious Shoppers	Medium OLD	Medium OLD	Medium 1-2 year	Medium 1-2 year
		Popular Dept		Serious Shoppers	AUTO	Medium OLD	Medium OLD	Medium 1-2 year	Medium 6-12 months

Figure 5: Credit usage customer segments

File View Cold	<b>p grayest = 18% of 95</b> or By Sort By Exit	157 records, sorted I	by weights, colored I	by ENH_DAS_SCR	(mean=0.003)				_ 🗆 ×
0 0 17K (18%) OTHDEPT_BA BANK_BAL11 INSTFIN_BAL	0 155 (0%) GOV_BAL AUTO_BAL ♦ BANK_BAL11	2 0 831(0%) OINSTFIN_BAL MORT_BAL ♦ AUTO_BAL	3 0 174 (0%) ●INSTFIN_BAL ■MORT_BAL ◆BANK_BAL11	4 0 310 (0%) STUDENT_BA INSTFIN_BAL BANK_BAL11	5 0 640 (0%) STUDENT_BAL INSTFIN_BAL BANK_BAL11	6 0 307 (0%) BANK_BAL01 BANK_BAL04 BANK_BAL02	7 0 356(0%) BANK_BAL02 BANK_BAL01 BANK_BAL05	8 0 944(0%) BANK_BAL0 MORT_BAL BANK_BAL0	9 0 389(0%) BANK_BAL02 MORT_BAL BANK_BAL05
0 1276(1%) BANK_BAL10 OILCARD_BA BANK_BAL11	0 284 (0%)     OCHARGECARD_BA     MORT_BAL     ♦ BANK_BAL12	12 0 378 (0%) ○REVFINAN_BAL MORT_BAL ◆ BANK_BAL11	13 0 570 (0%) ○REVFINAN_BAI MORT_BAL ◆ BANK_BAL11	14 0 1427 (1%) STUDENT_BA INSTFIN_BAL BANK_BAL11	<ul> <li>0 2132 (2%)</li> <li>BANK_BAL01</li> <li>BANK_BAL04</li> <li>BANK_BAL02</li> </ul>	6 0 1404 (1%) BANK_BAL01 BANK_BAL02 BANK_BAL04	D 1762 (2%) BANK_BAL02 BANK_BAL05 AUTO_BAL	0 1332(1% BANK_BAL0 BANK_BAL0 BANK_BAL1	19 0 684 (0%) BANK_BAL02 BANK_BAL05 MORT_BAL
0 4098(4%) AUTO_BAL BANK_BAL11 MORT_BAL	1892 (2%)     OTHDEPT_BAL     BANK_BAL11     ♦ AUTO_BAL	22 0 895 (0%)     OTHDEPT_BAL     POPDEPT_BAL     ♦ MORT_BAL	23 0 287 (0%) BANK_BAL10 BANK_BAL11 MORT_BAL	24 0 1686 (2%) BANK_BAL07 BANK_BAL04 BANK_BAL08	25 0 144(0%) BANK_BAL04 BANK_BAL07 BANK_BAL08	26 0 807 (0%) BANK_BAL01 BANK_BAL04 BANK_BAL05	0 1064 (1%) BANK_BAL05 BANK_BAL02 BANK_BAL08	23 0.02 492( BANK_BAL0 BANK_BAL0 BANK_BAL0 Comparison	29 0 401(0%) BANK_BAL02 BANK_BAL05 BANK_BAL08
0 1412 (1%) OAUTO_BAL BANK_BAL11 ◆ POPDEPT_BA	0 656 (0%)     SECUR_BAL     AUTO_BAL     MINSTFIN_BAL	0 653 (0%) OTHDEPT_BAL BANK_BAL11 ♦ MORT_BAL	33 0 3050 (3%) ●BANK_BAL11 ■AUTO_BAL ◆ OTHDEPT_BAL	34 0 1393 (1%) BANK_BAL08 BANK_BAL11 BANK_BAL05	0 563(0%)     ●BANK_BAL07     BANK_BAL04     ◆BANK_BAL08	36         0         1436 (2%)           BANK_BAL05         BANK_BAL08           BANK_BAL02	0 594(0%) ●BANK_BAL05 ■MORT_BAL ◆BANK_BAL08	33 0 628(0%) ●BANK_BAL0 ●BANK_BAL0 ◆BANK_BAL0	0.03         309 (0           BANK_BAL02         BANK_BAL05           BANK_BAL08         BANK_BAL08
40 0 365(0%) ○AUTO_BAL MORT_BAL ◆BANK_BAL11	41 0 206 (0%) ○SECUR_BAL ■MORT_BAL ◇AUTO_BAL	42 0 628 (0%) ●BANK_BAL11 ■ AUTO_BAL ♦ MORT_BAL	43 0 1736(2%) ●BANK_BAL11 ■MORT_BAL ◆BANK_BAL10	44 0 697(0%) BANK_BAL11 BANK_BAL05 BANK_BAL08	45 0 951(0%) ●BANK_BAL08 ■ BANK_BAL11 ♦ AUTO_BAL	46 0 877 (0%) ●BANK_BAL08 ■BANK_BAL05 ♦ AUTO_BAL	47 0 860(0%) ●BANK_BAL05 ■BANK_BAL08 ◆BANK_BAL02	43 0 478(0%) BANK_BAL0 BANK_BAL0 BANK_BAL0	49 0.02 347(0 BANK_BAL05 BANK_BAL08 BANK_BAL02
50 0 571 (0%) CREDITU_BAI AUTO_BAL \$ BANK_BAL11	51 0 616 (0%) CREDITU_BAL MORT_BAL AUTO_BAL	52 0 289 (0%) ●ELEC_BAL ■ HOMEFURN_BA ♦ BANK_BAL11	53 0 569 (0%) ●ELEC_BAL HOMEFURN_BA ◆ MORT_BAL	54 0 869(0%) ●BANK_BAL11 ●POPDEPT_BA ◆BANK_BAL10	55 0 430(0%) ●POPDEPT_BAL ■ BANK_BAL08 ◆ BANK_BAL05	56 0.01 579(0 BANK_BAL11 BANK_BAL08 BANK_BAL05	57 0 520(0%) BANK_BAL08 MORT_BAL BANK_BAL05	58 0.02 493( BANK_BAL0 BANK_BAL0 ANK_BAL1	59 0.01 353(0 BANK_BAL08 BANK_BAL05 BANK_BAL11
60 0 1568(2%) ○CREDITU_BAI AUTO_BAL ◆ BANK_BAL11	61 0 509 (0%) HOMEFURN_BAL HOME_BAL MORT_BAL	62 0.02 177(0% HOMEFURN_BA HOME_BAL \$ BANK_BAL11	63 0 896 (0%) ●HOMEFURN_B# ■HOME_BAL ◆MORT_BAL	64 0 2143 (2%) POPDEPT_BA OTHDEPT_BA BANK_BAL11	0   1310 (1%)     POPDEPT_BAL     OTHDEPT_BAL     AUTO_BAL	66 0.02 250(0 POPDEPT_BA BANK_BAL02 BANK_BAL05	0   254 (0%)     POPDEPT_BA     BANK_BAL08     BANK_BAL11	63 0.02 310( BANK_BAL0 BANK_BAL1 BANK_BAL0	69 0.08 208(0 ●BANK_BAL05 ■BANK_BAL08 ♦BANK_BAL11
0 4365(5%) MORT_BAL BANK_BAL11 ♦ AUTO_BAL	0 1377 (1%) MORT_BAL AUTO_BAL BANK_BAL11	<pre> 2 0 517(0%) 3 MORT_BAL 3 BANK_BAL12 4 BANK_BAL11 </pre>	1200 (1%)     BANK_BAL12     BANK_BAL11     MORT_BAL	74 0 637 (0%) ●POPDEPT_BA ■MORT_BAL ◆BANK_BAL11	<ul> <li><sup>15</sup> 0 656 (0%)</li> <li>POPDEPT_BAL</li> <li>OTHDEPT_BAL</li> <li>AUTO_BAL</li> </ul>	76 0 398(0%) ●POPDEPT_BA ■BANK_BAL11 ◆ OTHDEPT_BA	0.02 154(0 POPDEPT_BA BANK_BAL08 BANK_BAL11	78 0.03 308( BANK_BAL1 BANK_BAL0 BANK_BAL0	79 0.04 236(0 BANK_BAL12 BANK_BAL05 BANK_BAL08
0 2237 (2%) MORT_BAL BANK_BAL11 OTHDEPT_BA	81 0 654 (0%) ●BANK_BAL06 ■BANK_BAL09 ♦ MORT_BAL	82 0 345(0%) ●BANK_BAL09 ■BANK_BAL12 ◆MORT_BAL	0 589 (0%)     ●BANK_BAL12     BANK_BAL11     ◆BANK_BAL02	84 0 428(0%) BANK_BAL11 BANK_BAL12 MORT_BAL	0 231(0%) OTHDEPT_BAL POPDEPT_BAL BANK_BAL11	B6 0 534(0%) ●POPDEPT_BA BANK_BAL11 ◆ OTHDEPT_BA	0.02 308(0     BANK_BAL11     POPDEPT_BA     ♦ OTHDEPT_BA	83 0.01 391( BANK_BAL1 BANK_BAL0 BANK_BAL0	89 0.01 261(0 BANK_BAL12 BANK_BAL11 BANK_BAL08
90 0 463 (0%) MORT_BAL BANK_BAL11 ♦ AUTO_BAL	91 0 379 (0%) BANK_BAL03 BANK_BAL06 MORT_BAL	92 0 279(0%) ●BANK_BAL06 ●BANK_BAL09 ◆MORT_BAL	0 576 (0%) ■BANK_BAL12 ■MORT_BAL ◆BANK_BAL11	94 0 112(0%) OTHDEPT_BA BANK_BAL12 ◆BANK_BAL11	95 0 402 (0%) ●OTHDEPT_BAL ■ BANK_BAL11 ◆ CLOTHING_BA	96 0.01 252(0 BANK_BAL11 OTHDEPT_BA BANK_BAL08	97 0 957 (1%) ●BANK_BAL11 ■MORT_BAL ◆BANK_BAL12	93 0.02 458( BANK_BAL1 BANK_BAL1 ABANK_BAL1 ABANK_BAL0	99 0 208 (0%) POPDEPT_BA BANK_BAL11 \$ BANK_BAL12
						<1%			
						>1	1%		
							>8% >2%		

Figure 6: Actual bankruptcies among customers rated as marginal credit risks.

0 389 (0%)	BANK_BAL02 Mort_bal Bank_bal05	0 684(0%) BANK_BAL02 BANK_BAL05 MORT_BAL	0 401(0%) BANK_BAL02 BANK_BAL05 BANK_BAL05	0,03 309(0 BANK BAL02 BANK BAL05 BANK BAL05 BANK BAL08	0.02 347 (0 BANK_BAL05 BANK_BAL08 BANK_BAL08 BANK_BAL02	0,01 353(0 BANK_BAL08 BANK_BAL05 BANK_BAL11	0.08 208(0 BANK_BAL05 BANK_BAL08 BANK_BAL11	0.04 236(0 BANK_BAL12 BANK_BAL05 BANK_BAL05 BANK_BAL08	0.01 261(0 BANK_BAL12 BANK_BAL11 BANK_BAL08	0 208(0%) POPDEPT_BA BANK_BAL11 BANK_BAL12
8 0 944(0%)	BANK_BALD MORT_BAL	1332(1%         19           BANK_BAL0         BANK_BAL0           BANK_BAL0         P	23         0.02         492         492           BANK_BALQ         BANK_BALQ         PANK_BALQ	0         628 (0%)         00           BANK_BALQ         BANK_BALQ         00           BANK_BALQ         00         00	48 0 478 (0%) 49 ● BANK_BALG ● BANK_BALG ● BANK_BALG	38         0.02         493 (39)           BANK_BALQ         BANK_BALQ           BANK_BALQ         Pank_BALQ	BANK_BAL0	78 0.03 308 ( 79 ● BANK BAL1 ● BANK BAL0 ● BANK BAL0	Bank_Ball	38         0.02         458 (         99           BANK_BAL1         BANK_BAL1         9           BANK_BAL1         9         9
7 0 356(0%)	BANK_BAL02 BANK_BAL01 BANK_BAL05	<ul> <li>         T 0 1762 (2%)         ● BANK_BAL02         ■ BANK_BAL05         ◆ AUTO_BAL     </li> </ul>	27         0         1064 (1%)           BANK_BAL05         BANK_BAL02           BANK_BAL02         BANK_BAL02	37         0         594 (0%)           BANK_BAL05         MORT_BAL           MORT_BAL	47 0 860 (0%) ●BANK_BAL05 ●BANK_BAL08 ◆BANK_BAL08	57 0 520(0%) ● BANK_BAL08 ■ MORT_BAL ● BANK_BAL05	Comparison     Comparison	<b>77</b> 0.02 154 (0 <b>POPDEPT_BA</b> <b>BANK_BAL08</b> <b>BANK_BAL11</b>	37 0.02 308(0 ●BANK_BAL11 ● POPDEPT_BA ◆OTHDEPT_BA	97 0 957 (1%) BANK_BAL11 MORT_BAL & BANK_BAL12
6 0 307 (0%)	BANK_BAL01 BANK_BAL04 BANK_BAL02	16         0         1404 (1%)           ● BANK_BAL01         ■         BANK_BAL02           ● BANK_BAL04         ■         BANK_BAL04	28 0 807 (0%) ●BANK_BAL01 ■BANK_BAL04 ◆BANK_BAL04	35 0 1436 (2%) ● BANK_BAL05 ■ BANK_BAL08 ◆ BANK_BAL02	46 0 877 (0%) ● BANK_BAL08 ● BANK_BAL05 ◆ AUTO_BAL	26 0.01 579(0         ■BANK_BAL11         ■BANK_BAL08         ●BANK_BAL08         ●BANK_BAL08         ■BANK_BAL08         ■BANK_BAL08	36 0.02 250(0     ●POPDEPT_BA     ■BANK_BAL02     ●BANK_BAL02	76 0 398 (0%) ◆ POPDEPT_BA ■ BANK_BAL11 ◆ OTHDEPT_BA	36         0         534 (0%)           ● POPDEPT_BA           ■ BANK_BAL11           ◆ OTHDEPT_BA	96         0.01         252 (0           BANK_BAL11         0THDEPT_BA         BANK_BAL08
(mean=0.003) 5 0 640 (0%)	<pre>● STUDENT_BAL ■ INSTFIN_BAL ◆ BANK_BAL11</pre>	<ul> <li>15 0 2132 (2%)</li> <li>● BANK_BAL01</li> <li>■ BANK_BAL04</li> <li>◆ BANK_BAL02</li> </ul>	25 0 144 (0%) ●BANK_BAL04 ■ BANK_BAL07 ◆ BANK_BAL08	<ul> <li>35 0 563 (0%)</li> <li>●BANK_BAL07</li> <li>■BANK_BAL04</li> <li>◆BANK_BAL08</li> </ul>	45 0 951 (0%) BANK_BAL08 BANK_BAL11 AUTO_BAL	55 0 430 (0%) ● POPDEPT_BAI ■ BANK_BAL08 ◆ BANK_BAL05	<ul> <li>55 0 1310 (1%)</li> <li>● POPDEPT_BAI</li> <li>● OTHDEPT_BAI</li> <li>◆ AUTO_BAL</li> </ul>	75 0 656 (0%) ● POPDEPT_BAI ■ OTHDEPT_BAI ◆ AUTO_BAL	<ul> <li>85 0 231 (0%)</li> <li>● OTHDEPT_BAI</li> <li>■ POPDEPT_BAI</li> <li>◆ BANK_BAL11</li> </ul>	95 0 402 (0%) ● OTHDEPT_BAI ■ BANK_BAL11 ◆ CLOTHING_BA
9 ENH_DAS_SCR (	<pre> STUDENT_BA STUDENT_BA INSTFIN_BAL BANK_BAL11 </pre>	14         0         1427 (1%)           OSTUDENT_BA         INSTFIN_BAL           MANK_BAL11	24 0 1686 (2%) ●BANK_BAL07 ●BANK_BAL04 ◆BANK_BAL08	34         0         1393 (1%)           BANK_BAL08         BANK_BAL11           BANK_BAL11         BANK_BAL05	<ul> <li>▲▲ 0 697 (0%)</li> <li>● BANK_BAL11</li> <li>■ BANK_BAL05</li> <li>◆ BANK_BAL08</li> </ul>	<b>54</b> 0 869 (0%) ●BANK_BAL11 ■POPDEPT_BA ◆BANK_BAL10	64 0 2143 (2%) ● POPDEPT_BA ● DTHDEPT_BA ◆ BANK_BAL11	74         0         637 (0%)           ● POPDEPT_BA           ■ MORT_BAL           ◆ BANK_BAL11	84 0 428 (0%) ●BANK_BAL11 ■BANK_BAL12 ◆MORT_BAL	94 0 112 (0%) OTHDEPT_BA BANK_BAL12
y weights, colored k 3 0 174 (0%)	●INSTFIN_BAL ■ MORT_BAL ◆ BANK_BAL11	<ul> <li>18 0 570 (0%)</li> <li>CEVFINAN_BAI</li> <li>MORT_BAL</li> <li>BANK_BAL11</li> </ul>	28 0 287 (0%) BANK BAL10 BANK BAL11 MORT BAL	<ul> <li>30 3050 (3%)</li> <li>BANK_BAL11</li> <li>AUTO_BAL</li> <li>OTHDEPT_BAL</li> </ul>	<ul> <li>48 0 1736(2%)</li> <li>BANK_BAL11</li> <li>MORT_BAL</li> <li>BANK_BAL10</li> </ul>	<ul> <li>569 (0%)</li> <li>● ELEC_BAL</li> <li>■ HOMEFURN_B/</li> <li>◆ MORT_BAL</li> </ul>	0         896 (0%)           OHOMEFURN_B/           HOME_BAL           MORT_BAL	78 0 1200(1%) ● BANK_BAL12 ■ BANK_BAL11 ◆ MORT_BAL	88 0 589 (0%) ●BANK_BAL12 ■BANK_BAL11 ◆BANK_BAL02	08 0 576 (0%) ●BANK_BAL12 ■MORT_BAL ◆BANK_BAL11
157 records, sorted k 2 0 831 (0%)	●INSTFIN_BAL ■ MORT_BAL ◆ AUTO_BAL	12 0 378 (0%) CEVFINAN_BAI MORT_BAL BANK_BAL11	22 0 895 (0%) ● OTHDEPT_BAL ■ POPDEPT_BAL ● MORT_BAL	22 0 653 (0%) ● OTHDEPT_BAL ■ BANK_BAL11 ◆ MORT_BAL	<ul> <li>42 0 628 (0%)</li> <li>● BANK_BAL11</li> <li>■ AUTO_BAL</li> <li>◆ MORT_BAL</li> </ul>	52 0 289 (0%) ●ELEC_BAL ■HOMEFURN_B/ ◆BANK_BAL11	<ul> <li>0.02 177 (0%</li> <li>●HOMEFURN_B/</li> <li>●HOME_BAL</li> <li>◆BANK_BAL11</li> </ul>	72 0 517 (0%) MORT_BAL BANK_BAL12 ◆BANK_BAL11	82 0 345 (0%) ●BANK_BAL09 ■BANK_BAL12 ◆MORT_BAL	22 0 279 (0%) ●BANK_BAL06 ■BANK_BAL09 ♦MORT_BAL
p grayest = 18% of 95 r By Sort By Exit 1 0 155 (0%)	● GOV_BAL ■ AUTO_BAL ◆ BANK_BAL11	11         0         284 (0%)           CHARGECARD_B/           MORT_BAL           BANK_BAL12	21 0 1892 (2%) OTHDEPT_BAL BANK_BAL11 ◆ AUTO_BAL	31         0         556 (0%)           OSECUR_BAL         AUTO_BAL           AUTO_BAL         NISTFIN_BAL	41         0         206 (0%)           SECUR_BAL           MORT_BAL           AUTO_BAL	<b>51</b> 0 616 (0%) ●CREDITU_BAL ■MORT_BAL ◆AUTO_BAL	61 0 509 (0%) ●HOMEFURN_BAI ■HOME_BAL ◆MORT_BAL	71 0 1377 (1%) MORT_BAL AUTO_BAL & BANK_BAL11	81         0         554 (0%)           ●BANK_BAL06           ■BANK_BAL09           ●MORT_BAL	91 0 379(0%) BANK BAL03 BANK BAL06 MORT_BAL
(≝seg650E Kmat File View Colo □ 17⊼(18%)	<ul> <li>OTHDEPT_BA</li> <li>BANK_BAL11</li> <li>INSTFIN_BAL</li> </ul>	10 0 1276(1%) BANK_BAL10 OILCARD_BA BANK_BAL11	20 0 4098 (4%) ● AUTO_BAL ■ BANK_BAL11 ◆ MORT_BAL	30 0 1412 (1%) ● AUTO_BAL ■ BANK_BAL11 ◆ POPDEPT_BA	40 0 365 (0%) ● AUTO_BAL ■ MORT_BAL ◆ BANK_BAL11	50 0 571 (0%) ● CREDITU_BAI ■ AUTO_BAL ◆ BANK_BAL11	30 0 1568 (2%) ● CREDITU_BAI ■ AUTO_BAL ◆ BANK_BAL11	70 0 4365 (5%) ● MORT_BAL ■ BANK_BAL11 ◆ AUTO_BAL	30 0 2237 (2%) ● MORT_BAL ■ BANK_BAL11 ◆ OTHDEPT_BA	0 463 (0%) ● MORT_BAL ■ BANK_BAL11 ◆ AUTO_BAL

Figure 7: Actual bankruptcies among customers rated as marginal credit risks.

#### **3.4** Census Data

The example in this section shows how Kmap can handle categorical values. The clustering was performed using publicly available census data with a mixture of 12 continuous and categorical attributes for each person. The data comes from a municipality located near a National Laboratory, which accounts for the unusually large number of advanced degrees and large diversity of backgrounds.

Categorical attributes pose an extra degree of complexity in neural clustering because they are expanded to a one-to-N mapping: for example, if the original attribute were "animal type" with values "cat, dog, cow", it would be replaced for clustering purposes by three attributes "is\_a\_cat", "is\_a\_dog", "is\_a\_cow" that can each assume values 1 or 0. So the total number of attributes actually used in the clustering is data-dependent.

Kmap handles this by making two passes through the results file, obtaining the number of distinct values for each categorical attribute on the first pass and constructing the effective attributes coming from each original attribute on the second pass. Sample Kmap results are shown in Figure 9 and are compared to the default Intelligent Miner visualizer's view in Figure 8. The two views each present interesting information in different ways, but they are indeed consistent with each other. For example, one of the supplemental attributes (i.e., not used for clustering) is whether household income is above or below \$50,000. Using this attribute to color the Kmap in Figure 9 points to clusters 18 and 24 as being highest in this attribute, with 70% of the households in cluster 24 having income above \$50,000. This agrees with the corresponding pie chart in Figure 8.

One of the original categorical attributes was "education", and one of its values was "doctorate". Drilling down into the effective categorical attribute "Doctorate(education)" (or "has\_a\_PhD" in our earlier terminology) in Figure 9 (comfortingly, perhaps) an overlap between clusters with high PhD populations and clusters with high income. Drilling down into the "educ\_num" attribute, a number proportional to the number of years of education, shows similar results. Both results are consistent with the drilldowns in Figure 8. Kmap does a good job showing the attribute distributions on the map. The default visualizer does a good job showing the categorical attributes in pie charts. Both can be useful.

#### 3.5 Appendix: A Tale of Two Clusters

Figure 10 offers a more detailed look at the kind of information that Kmap can extract from the supermarket customer clustering results mentioned earlier.

### 4 Summary

Kmap has proved to be a versatile visualizer for Kohonen self-organizing map clustering results obtained from a variety of datamining engagements. Analysts value being able to see the map topography, as well as the drilldowns provided, many of which resulted from their suggestions. We provided examples of Kmap's use in four different areas of commerce. Currently we are applying Kmap to technical data generated by the simulators and hardware control systems of massively parallel supercomputers.



[income]



Figure 8: Census data clustering results seen with the Intelligent Miner default visualizer

educ\_num

education

👹 census36 Kmap grayest = 8%	of 32560 records	, sorted by Rmean	n, colored by >50K	(income) 💶 🔼 🗙
File View Color By Sort By Exit				
0 0.32 2602( 1 0.25 326(1	2 0 514 (2%)	3 0 1140(4%)	4 0 1016(3%)	5 0 1482(5%)
🗘 🛡 HS-grad(educ: 🔎 Amer-Indian-E	Black(race)	Own-child(rel;	Never-worked	Own-child(relatio
Transport-ma	Never-worked	Without-pay(w	11th(education	Some-college(ed
Craft-repair(o + Haiti(native_co	Haiti(native_complexity)	Armed-Force:	12th(educatio)	Never-worked(w
6 0.28 1273( 7 0.45 398(1	8 0.02 293(0	9 0.02 393(1	10 0.03 564(2	11 0.04 524(2%)
Assoc-voc(ed 🔎 Trinadad&Tob	Haiti(native_cc	Other-relative	😑 Holand-Nethei	Taiwan(native_co
7th-8th(educa Jamaica(nativ	Other-relative	Hungary(nativ	Other-relative	Own-child(relatio
10th(education  Black(race)	Jamaica(nativ)	Outlying-US(G	Guatemala(na	Bachelors(educat
12 0.44 1513( 13 0.08 379(1	14 0.08 398(1	15 0.04 748(2	<b>16 0.11 1132(</b>	17 0.13 1120(3%)
Some-college Mexico(native	Outlying-US(G	Armed-Force:	븆 Not-in-family(I	Bachelors(educat
Husband(relat st-4th(educa	📕 Thailand(nativ	Ireland(native	Armed-Force:	Hungary(native_c
Married-civ-s	Laos(native_c	Not-in-family()	Never-marrie	Not-in-family(rela
18 0.75 2112( 19 0.42 370(1	20 0.11 418(1	21 0.16 853(3	22 0.1 289(0%	23 0.07 1348(4%)
Bachelors(edi 🔍 Laos(native_c	Separated(ma	Divorced(mari	Married-spou	🖶 Not-in-family(rela
Yugoslavia(na Cambodia(nat	Peru(native_c	Not-in-family(I	Dominican-Re	Never-married(m
Masters(educ + Hong(native_c	Columbia(nati	Transport-mo	Thailand(nativ	Priv-house-serv(
24 0.7 1489(5 25 0.57 300(0	26 0.07 217(0	27 0.14 736(2	28 0.08 817(2	29 0.08 577(2%)
🗣 Prof-school(e 🛛 🐥 Taiwan(native	Other-relative	Divorced(mari	Divorced(mari	Widowed(marital
Self-emp-inc( <sup>,</sup> Greece(native	Priv-house-se	Never-worked	Not-in-family(I	Hungary(native_c
Doctorate(edu     Iran(native_co	El-Salvador(na	+ Federal-gov(w	Scotland(nativ	Priv-house-serv(
30 0.34 2122( 31 0.25 514(2	32 0.48 1514(	33 0.04 327(1	34 0.08 1601(	35 0.05 1141(4%)
Self-emp-not- 🔎 ?(workclass)	Wife(relations	Divorced(mari	븆 Unmarried(rel	Separated(marita
Farming-fishir 📕 ?(occupation)	Married-AF-sp	Own-child(rel:	Divorced(mari	Widowed(marital
Self-emp-inc(* Portugal(nativ	Scotland(nativ	Cuba(native_c	Honduras(nati	Unmarried(relatic

👹 Doctorate(education) Kmap; darkest shade	= 24% market share of 413	
File View Color By Sort By Exit		
0weight 0 1weight 0 2weight 0 3	weight 0 <mark>4weight 0.</mark> 5weight 0	
00% Val 00% Val 00% Val 0	0% Val 00% Val 00% Val	
m 00 m 00 m 00 m	📸 educ_num Kmap; darkest shade = 9% market share of 328,228	_ 🗆 🗙
6weight 0 7weight 0.8weight 0 9	File View Color By Sort By Exit	
m = 0.06 m = 0.079 m = 0.0 m	0weight 0 1weight 0.2weight 0.3weight 0.4weight 0.5w	eight O.
12 weight 0 13 weight 0 14 weight 0 1	07% Val 00% Val 01% Val 03% Val 02% Val 05%	Val
00% Val $00%$ Val $03%$ Val $0$	m. 90.89 m. 90.89 m. 90.91 m. 90.89 m. 80.76 m.	10 0.99
m 0 0 m 0 0.21 m 0 2.58 m	. <mark>6</mark> weight 0. 7weight 0. 8weight 0. <mark>9</mark> weight 0. 10.weight 0. 11.	reight O.
18.weight 0.19.weight 0.20.weight 0.2	03% Val 01% Val 00% Val 01% Val 02% Val 02%	Val
21% Val 04% Val 00% Val 0	m 80.82 m 101 m 90.9 m 90.89 m 90.88 m	13 1.3
m 03.25 m 03.62 m 00.38 m	12.weight 0 13.weight 0.14.weight 0.15.weight 0.16.weight 0.17.w	reight O.
24.weight 0.25.weight 0.26.weight 0.2	05% Val 00% Val 01% Val 02% Val 04% Val 04%	Val
24% Val 09% Val 00% Val 0	m 10 0.99 m 6 0.57 m 9 0.94 m 9 0.89 m 10 1.03 m	13 1.29
m 05.29 m 010.25 m 00.73 m	n <mark>18</mark> .weight 0 <mark>19</mark> .weight 0. <mark>20</mark> .weight 0. <mark>21</mark> .weight 0 <mark>22</mark> .weight 0. <mark>23</mark> .w	reight O.
30.weight 0.31.weight 0.32.weight 0.3	09% Val 01% Val 01% Val 03% Val 00% Val 04%	Val
01% Val 03% Val 05% Val 0	m 13 1.33 m 11 1.11 m 9 0.94 m 10 0.97 m 9 0.9 m	10 1.02
m 00.19 m 01.84 m 01.04 m	24.weight 0.25.weight 0.26.weight 0.27.weight 0.28.weight 0 29.w	reight O.
	06% Val 01% Val 00% Val 02% Val 02% Val 02%	Val
	m 13 1.33 m 12 1.2 m 9 0.86 m 11 1.06 m 10 0.99 m	9 0.9
	30.weight 0.31.weight 0 32.weight 0.33.weight 0.34.weight 0.35.w	reight O.
	06% Val 01% Val 05% Val 00% Val 05% Val 03%	Val
	m 90.89 m 90.94 m 101.04 m 100.95 m 101 m	9 0.92

Figure 9: Census data clustering results seen with Kmap



Figure 10: A Tale of Two Clusters: rather different lifestyles emerge during examination of the supermarket shopper clustering results. Cluster 4 seems to be an older group of serious homebakers likely to be found having a tea party, whereas cluster 9 seems to be a younger, health-conscious group more likely to be found jogging - and not smoking!

## 5 Acknowledgements

We are grateful for many helpful discussions with our colleagues Joseph Bigus and Michael Rothman.

## References

- Kotlyar V., Viveros M.S., Duri S.S., Lawrence R.D., Almasi G.S. "A Case Study in Information Delivery to Mass Retail Markets" In the Proceedings of the 10th International Conference on Database and Expert Systems Applications (DEXA), Florence, Italy, August/September 1999, pp. 842-851. Published by Springer-Verlag as Lecture Notes in Computer Science, vol 1677
- [2] R. D. Lawrence, G. S. Almasi, V. Kotlyar, M. S. Viveros, and S. S. Duri, "Personalization of Product Recommendations in Mass Retail Markets", *Data Mining and Discovery* 3(2): 171-195 (1999).
- [3] G. S. Almasi and A. J. Lee, "PDA-based Personalized Recommender Agent", Proceedings of the 5th International Conference on the Practical Application of Intelligent Agent and Multi Agent Technology (PAAM 2000), Manchester, England, April 2000, pp. 299-309.
- [4] H. Edelstein, "Mining Large Databases: A Case Study", www.twocrows.com/largedb.pdf
- [5] "JClass Java Components", www.klgroup.com
- [6] G. Almasi, "Kmap Demo", www.gsalmasi.com
- [7] Intelligent Miner for Data, www.ibm.com/software/data/iminer/fordata
- [8] T. Kohonen, "Self-Organizing Maps", Springer-Verlag, 1995.
- [9] R. D. Lawrence, G. S. Almasi, and H. E. Rushmeier, "A Scalable Parallel Algorithm for Self-Organizing Maps with Applications to Sparse Data Mining Problems", *Data Mining and Knowledge Discovery*, Vol. 3, 1999, pp. 171-195. (U. S. Patent 6,260,036, July 10, 2001.)
- [10] J. C. Bezdek and N. R. Pal, "Some New Indexes of Cluster Validity", *IEEE Transactions on Systems, Man, and Cybernetics Part B: Cybernetics* Vol. 28, No. 3, June 1998, pp. 301 315.