

Automate the Process of Image Recognizing a Scatter Plot: an Application of Non-parametric Cluster Analysis in Capturing Data from Graphical Output

Zhaojie Wang, Celgene Corporation, Summit, NJ

ABSTRACT

A fundamental method of pharmaceutical research is to compare the indication between a new drug and an existing competitor drug. Sometimes, the original data of the competitor drug may not always be available. What a researcher can access is only a plot from a published paper. To make the comparison, the researcher has to capture tabular data from this legacy graphical output.

Previously, several methods^{[1][13]} were introduced to digitalize a graphical output into a set of coordinates of pixels, which are the minimum components of a graph. With the pixel coordinates, the graphical output of a competitor drug can be copied and relayed on the graph of the new drug for comparison purpose. These methods are more helpful for reverse-engineering the trend of a curve plot than recognizing a scatter plot, because it is more challenging to recognize a scatter plot. An individual spot, even in the smallest size and the simplest pattern, is composed by multiple pixels. To recognize a scatter plot, we have to figure out a way to derive the coordinate of each individual spot from the pixel coordinates.

This article introduces a non-parametric cluster analysis method to facilitate recognizing a scatter plot. PROC MODECLUS in SAS 9.2 is used to implement this statistical method. It processes the pixel coordinates obtained from the original scatter plot, and derives the coordinates of scatter spots automatically. This automation strategy is able to improve not only the efficiency but also the accuracy of tabular data capture. To illustrate this approach, example plots and SAS codes are presented.

As an important area of computer vision and artificial intelligence, automated image recognition has drawn more and more attention in pharmaceutical research. The application introduced in this article illustrates that SAS can be a powerful and convenient tool for recognizing and processing image data. Enhanced with this potential, SAS is able to play a more significant role in pharmaceutical research.

INTRODUCTION

For the purpose of new drug development and marketing, pharmaceutical researchers have to compare the indications between the new drug and existing competitor drugs. Ideally, researchers can obtain the original indication data of both the new drug and the competitor drugs. Then they are able to compare the indications by analyzing the data obtained. However, sometimes, the original data of the competitor drug may not be accessed. The information of the competitor drug can only be accessed from a published paper, which usually presents the indication with graphs instead of original tabular data. The graphs can be scatter plots, survival curves, regression curves, and so on. Then, researches have to figure out a reverse-engineering method to obtain the tabular data from the graphical output.

A scatter plot visualizes data with scatter spots. The scatter spots are patterns, which are built by many pixels arranged in a specified order. To automate the process of recognizing a scatter spot, following two step strategy is proposed.

1. Recording the pixel coordinates using GetData graph digitalizer.
2. Recognize the scatter plot using a non-parametric cluster analysis method, which is proposed to derive the spot coordinates from the pixel coordinates. PROC MODECLUS in SAS 9.2 is used to implement this statistical method.

METHOD

1. RECORD THE PIXEL COORDINATES USING GetData GRAPH DIGITALIZER

The first step of recognizing a scatter plot is to digitalize the plot into a set of pixel coordinates.

Although graph digitalization software are widely available, choosing a right one is still important for better efficiency and accuracy of an image recognition. Because of the sophistication of pharmaceutical research, it is not uncommon that a scatter plot contains dozens or even hundreds of spots. Considering each spot can be composed by dozens of

pixels, it is challenging to track all the pixels by a manual point-and-pick software, like what we can do with Windows Paint^[1] and GraphPirate^[1]. Ideally, a software should be able to record the coordinates of all the pixels automatically. This requirement can be satisfied by many software, including GetData^[2], WinDig^[3], Engauge Digitizer^[4], Digitizer for Origin^[5], and DataThief^[6]. Because GetData Graph Digitalizer^[2] has a free full-functional trial version, I use it to illustrate how to digitalize a scatter plot.

GetData is supported by most Windows operating system (2000, XP, Vista, Win7). It is a straightforward four step process^[2] to digitalize a plot with GetData.

1. Open a graph in format BMP, JPG, TIF, or PCX.
2. Set the scale by defining a X-Y coordinate system.
3. Digitize by selecting the digitize area.
4. Export the pixel coordinate data to TXT, XLS, XML, DXF or EPS file.

Following scatter plot is used as an illustration. A JPG format file of this plot is obtained by scanning.

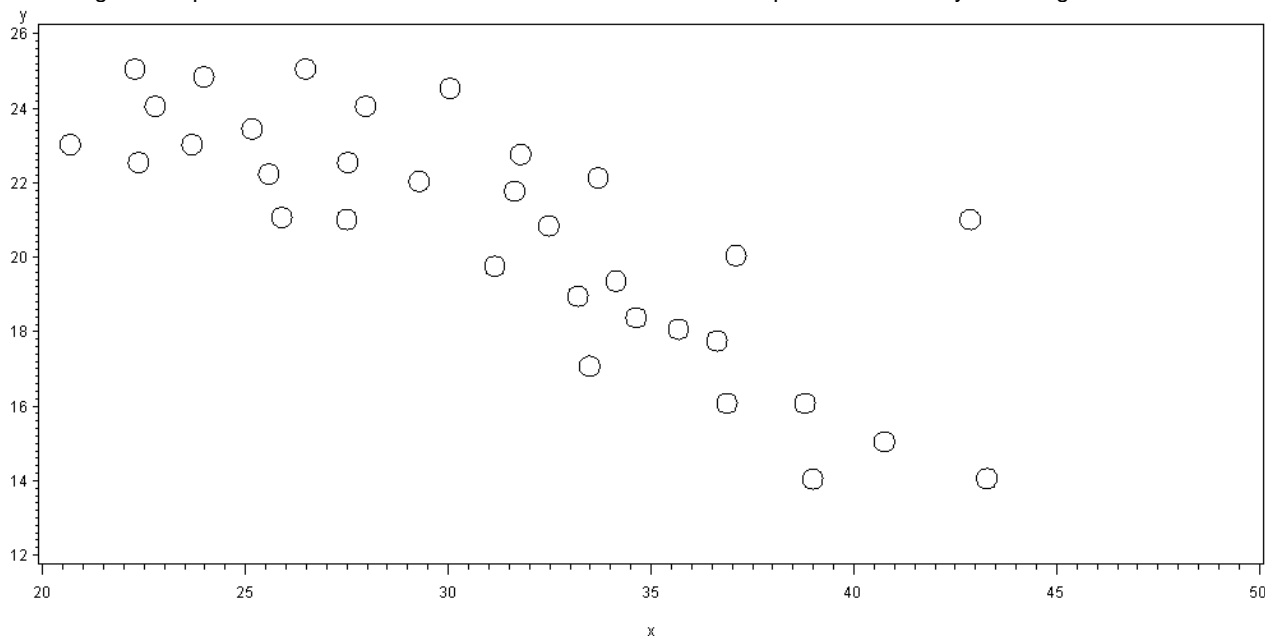
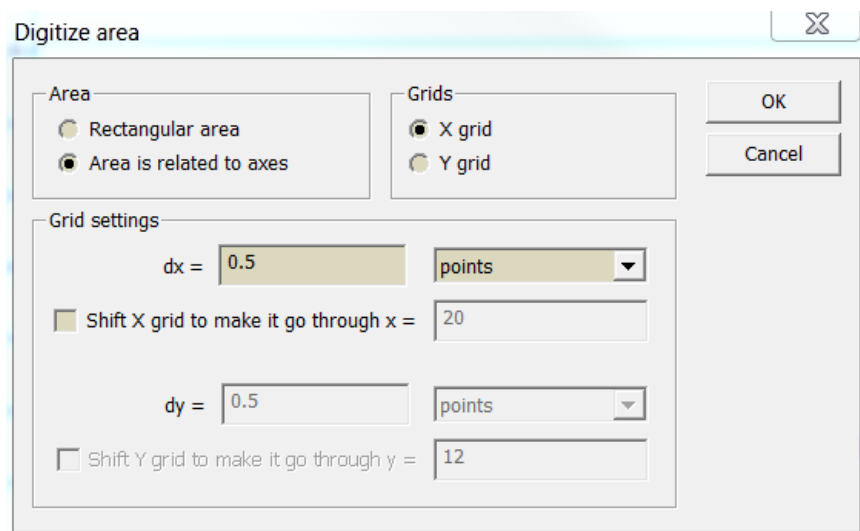


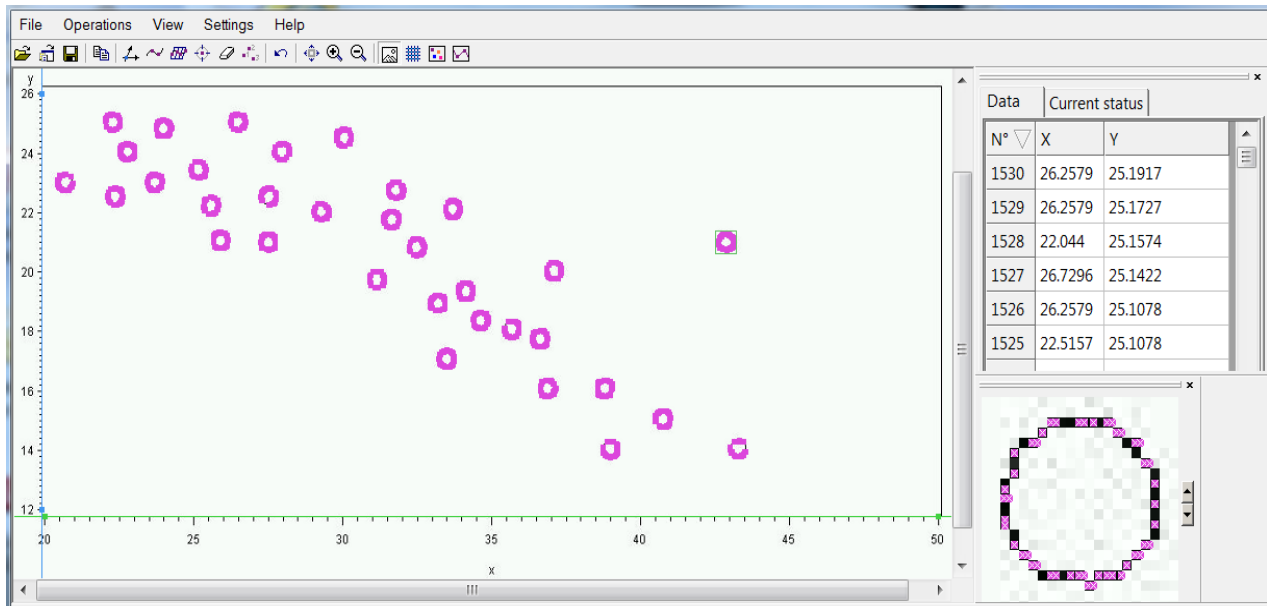
Figure 1. Scatter Plot Scanned in JPG Format

After open this image file by GetData, the X-Y scale can be set by selecting 2 reference points on each of X-axis and Y-axis. Then, set up both X grid and Y grid in following window to define the density of pixel sampling.



Display 1. The GetData Window of Setting Up Sampling Grids

After then, drag the mouse to select the digitizing area on the plot. For this case, totally 1530 pixels are picked from 33 spots on the scatter plot. And 1530 pairs of x-y coordinates are recorded automatically. Figure3 and 4 show how the pixels are picked up and recorded.



Display 2. GetData Window of Recording Pixels

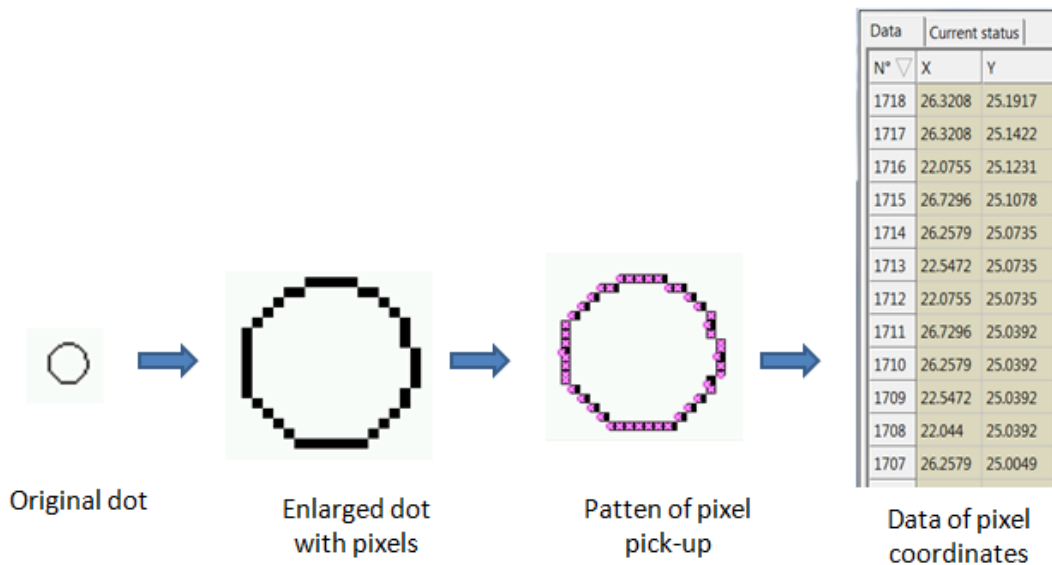


Figure 2. Illustration of Digitalizing A Single Spot into Pixel Coordinate Data

2. RECOGNIZE THE SCATTER SPOTS USING A NON-PARAMETRIC CLUSTER ANALYSIS METHOD

The key step of recognizing a scatter plot is to derive the spot coordinates from the pixel coordinates. Without further processing the pixel coordinates, the pixel coordinate data can only be used to 'copy' and 'paste' the reference scatter plot on the plot of the trial drug. But, no regression or other further analyses can be done. It works like a camera, which can only take a picture, but can never understand what the picture means.

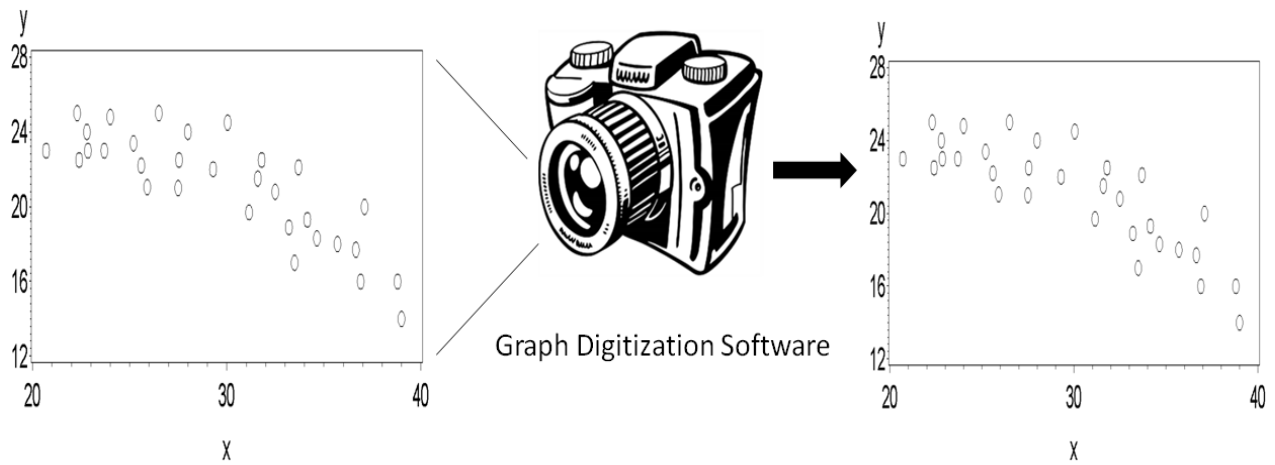


Figure 3. Without the Process of Image Recognition, Graph Digitization Software Can Only Work Like a Camera to Copy and Paste the Scatter Plot.

A human is able to recognize an image, because he/she has not only eyes to see the image, but also a brain to process the image information to understand it. In this proposed strategy of scatter plot recognition, GetData works as the 'eyes' to record the pixel coordinates. And the non-parametric cluster analysis method, which is implemented by SAS PROC MODECLUS, takes the role of 'brain' to derive the spot coordinates by processing the pixel coordinates.

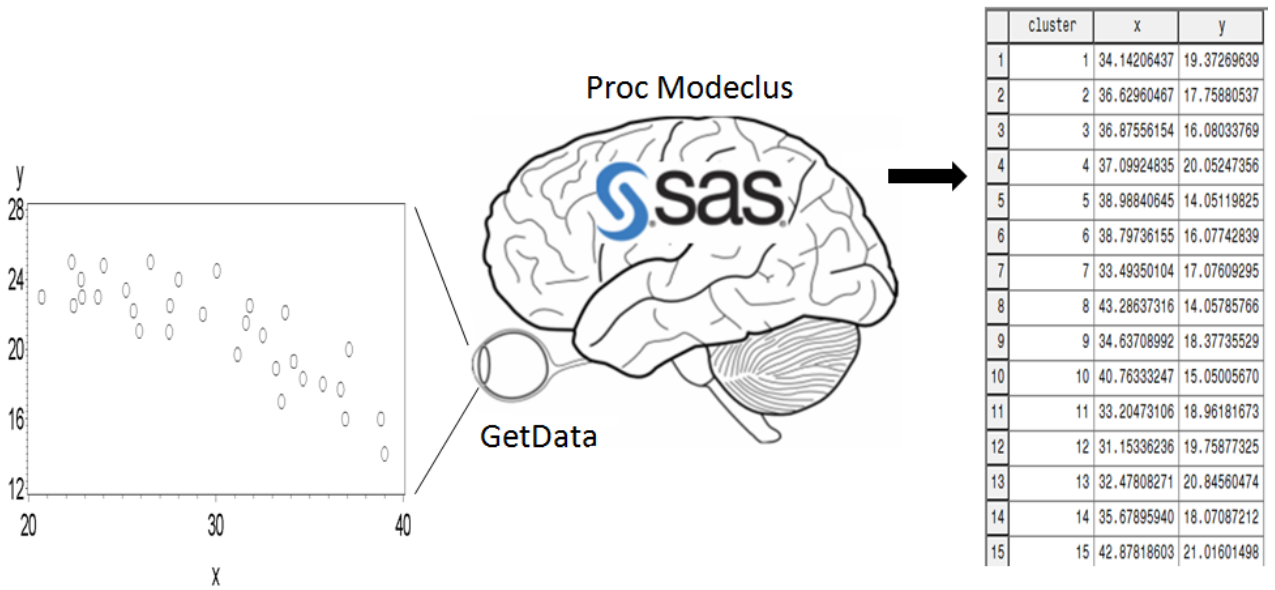


Figure 4. In the Process of Image Recognition, SAS Works as a 'Brain' to Implement the Non-parametric Cluster Analysis to 'Understand' the Scatter Plot.

It is a challenging task to identify scatter spot coordinates. It is the reason why software are rarely able to identify scatter spots automatically, although many of them enable user trace curves.

The pixels of the same scatter spot are relatively close to each other in a near neighbor. From a statistics point of view, they belong to the same cluster. Enlightened by this property of scatter spots, I have the idea to use the cluster analysis method for the scatter plot recognition.

As a statistical method, cluster analysis was originally introduced in anthropology and psychology in 1930's.^{[7][8]} Nowadays, with the need of processing big data, cluster analysis method is increasingly used in many fields, including image analysis, big data mining, machine learning, bioinformatics, and so on.

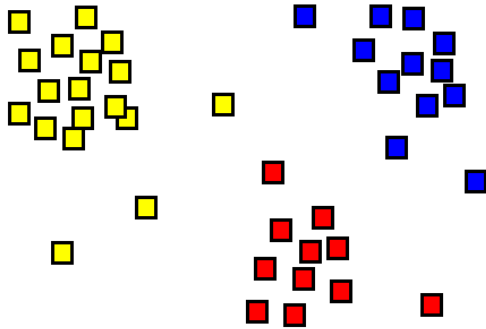


Figure 5. The Result of a Cluster Analysis Shown as the Coloring of the Squares into Three Clusters.^[9]

A cluster is a group of objects, which are more similar to each other than to those in other group. Cluster analysis is a number of statistical algorithms and methods for grouping multiple objects into clusters according to their similarity. It aims at sorting different objects into groups in a way that the degree of association between two objects is maximal if they belong to the same group and minimal otherwise. There are three major types of clusters possible^[10]:

- **Disjoint clusters** place each object in one and only one cluster.
- **Hierarchical clusters** are organized so that one cluster can be entirely contained within another cluster, but no other kind of overlap between clusters is allowed.
- **Overlapping clusters** can be constrained to limit the number of objects that belong simultaneously to two clusters, or they can be unconstrained, allowing any degree of overlap in cluster membership.

Because of the variety of the data properties and cluster types, cluster analysis contains a number of various algorithms and methods. There are five SAS procedures designed for clustering with different algorithms^[10].

- **PROC MODECLUS** finds disjoint clusters of observations with coordinate or distance data by using nonparametric density estimation.
- **PROC FASTCLUS** finds disjoint clusters of observations by using a k-means method applied to coordinate data.
- **PROC CLUSTER** performs hierarchical clustering of observations by using eleven agglomerative methods.
- **PROC VARCLUS** performs both hierarchical and disjoint clustering of variables by using oblique multiple-group component analysis.
- **PROC TREE** draws tree diagrams, also called dendrograms or phenograms by using output from the CLUSTER or VARCLUS procedure.

The appropriate clustering algorithm and parameter settings depend on the individual data set and intended use of the results. For recognizing the scatter plot, I propose to use nonparametric clustering methods, which base on nonparametric density estimates^[12]. Given a sufficiently large sample of pixels, nonparametric clustering methods are capable of detecting clusters/spots of unequal number of pixels, and with highly irregular shapes. Therefore, it provides more flexibility and reliability of recognizing scatter spots.

SAS MODECLUS procedure is designed to implement the nonparametric cluster analysis, in which the cluster is defined as a local maxima of the probability density. To calculate the density estimates, PROC MODECLUS uses hyper spherical uniform kernels of fixed or variable radius^[11]. The density estimate at a point X_i is computed by dividing n_i the number of observations within a sphere centered at the point (including point X_i itself) by the product of n the total sample size and V_i the volume of the sphere.

$$\hat{f}_i = \frac{n_i}{nV_i}$$

In above formula, both n_i and V_i are determined by the sphere radius, which is required to be specified by users with the R= option or K= option in the statement of PROC MODECLUS. The radius determine the number of clusters. A

smaller radius yields a larger number of clusters. It is usually necessary to try several different radius values. The value of R= option can be tried starting from following empirical formula ^[12]

$$\left[\frac{2^{v+2}(v+2)\Gamma(0.5v+1)}{nv^2} \right]^{1/(v+4)} \sqrt{\sum_{l=1}^v s_l^2}$$

where μ is the number of variables or the dimensionality, and S_l is the standard deviation of the l th variable. It can be derived in a DATA step by using the GAMMA function $\Gamma(t)$. Or the value of R= option can be estimated by following data-based Maximum Likelihood Estimation (MLE) method.

$$\sum_{i=1}^n \log \hat{f}_i^- \quad \text{where} \quad \hat{f}_i^- = \frac{n_i^-}{nv_i}$$

and n_i^- is the number of observations within the neighborhood of x_i , not including x_i itself.

After trying several different radius to obtain the best clustering results of recognizing the scatter plot above, the R= option is specified as 0.41. And the METHOD= option is specified as METHOD=1, which is recommended for most purposes.

```
PROC MODECLUS Data=scatter Method=1 R=0.41 Out=scatter_out;
Run;
```

PROC MODECLUS outputs a table of cluster statistics, including the cluster number, the number of observations in each cluster, and the maximum estimated density within the cluster.

The MODECLUS Procedure				
R=0.41 METHOD=1				
Cluster Statistics				
Cluster	Frequency	Maximum Estimated Density	Boundary Frequency	Estimated Saddle Density
1	52	0.04328877	0	
2	56	0.04328877	0	
3	51	0.04328877	0	
4	49	0.04328877	0	
5	51	0.04328877	0	
6	52	0.04081513	0	
7	46	0.04081513	0	
8	44	0.04081513	0	
9	48	0.04081513	0	
10	49	0.03957831	0	
11	43	0.03957831	0	
12	47	0.03957831	0	
13	48	0.03957831	0	
14	45	0.03957831	0	
15	47	0.03957831	0	
16	47	0.03957831	0	
17	48	0.03957831	0	
18	48	0.03834148	0	
19	44	0.03834148	0	
20	47	0.03834148	0	
21	45	0.03834148	0	
22	41	0.03710466	0	
23	46	0.03710466	0	

24	46	0.03710466	0
25	48	0.03710466	0
26	46	0.03710466	0
27	47	0.03710466	0
28	42	0.03586784	0
29	43	0.03463102	0
30	45	0.03463102	0
31	41	0.03463102	0
32	40	0.0333942	0
33	39	0.0333942	0
Cluster Summary			
	Number of	Frequency of	
R	Clusters	Unclassified	Objects
<hr/>			
0.41	33		0

Output 1. A Table of Output Statistics Created by PROC MODECLUS

The output dataset of PROC MODECLUS assigns a cluster membership to each of the pixels.

	cluster	x	y	density	_R_	_obs_	_method_
1	1	25.65688	20.99019	0.035867	0.41	291	1
2	1	25.65688	21.09313	0.034631	0.41	290	1
3	1	25.66037	21.05501	0.037104	0.41	1382	1
4	1	25.66037	21.07407	0.035867	0.41	1385	1
5	1	25.66037	21.12363	0.034631	0.41	1389	1
6	1	25.66037	21.14270	0.034631	0.41	1390	1
7	1	25.67435	20.99019	0.035867	0.41	294	1
8	1	25.67435	21.09313	0.035867	0.41	293	1
9	1	25.68832	21.23039	0.033394	0.41	297	1
10	1	25.70580	21.23039	0.034631	0.41	300	1
50	1	26.14605	21.07788	0.039578	0.41	340	1
51	1	26.16002	21.05882	0.038341	0.41	341	1
52	1	26.17749	21.05882	0.034631	0.41	342	1
53	2	33.23549	17.05936	0.038341	0.41	717	1
54	2	33.23899	16.93736	0.034631	0.41	1277	1
55	2	33.23899	16.95642	0.037104	0.41	1279	1
56	2	33.23899	17.00599	0.035867	0.41	1281	1
57	2	33.23899	17.02505	0.035867	0.41	1283	1
58	2	33.23899	17.09368	0.038341	0.41	1285	1
59	2	33.23899	17.12799	0.038341	0.41	1286	1

Output 2. Output Dataset of PROC MODECLUS

Because each cluster corresponds to one scatter spot, the mean of the all the pixel coordinates within a cluster is the best estimate of the spot coordinate. The means can be calculated by following PROC MEANS.

```

PROC MEANS Data=scatter_out ;
  var x y;
  by cluster ;
  output out=cluster_mean (where=( _stat_='MEAN' ));
Run;

```

PROC MEANS derives following coordinates of all the 33 scatter sports.

	cluster	_stat_	_freq_	x	y
1	1	MEAN	52	25.916586035	21.073634155
2	2	MEAN	56	33.495495158	17.061614924
3	3	MEAN	51	34.139228018	19.360705284
4	4	MEAN	49	39.012821061	13.99929972
5	5	MEAN	51	40.777736671	15.041661327
6	6	MEAN	52	22.786580121	24.038691973
7	7	MEAN	46	27.53532039	22.564814815
8	8	MEAN	44	27.50865574	21.048165478
9	9	MEAN	48	34.612086536	18.355108479
10	10	MEAN	49	22.387298735	22.525287893
24	24	MEAN	46	31.622307295	21.771786492
25	25	MEAN	48	31.801260773	22.749171659
26	26	MEAN	46	36.613875976	17.745891352
27	27	MEAN	47	42.865798356	20.980623928
28	28	MEAN	42	22.291521081	25.059005084
29	29	MEAN	43	29.305088327	22.010209252
30	30	MEAN	45	31.150632813	19.745073832
31	31	MEAN	41	35.694891855	18.055024178
32	32	MEAN	40	25.596436059	22.216530501
33	33	MEAN	39	43.275726137	14.004566784

Output 3. Output Dataset of PROC MEANS Contains the Derived Coordinates of All 33 Scatter Spots

Following figure shows the 33 scatter spots identified by PROC MODECLUS.

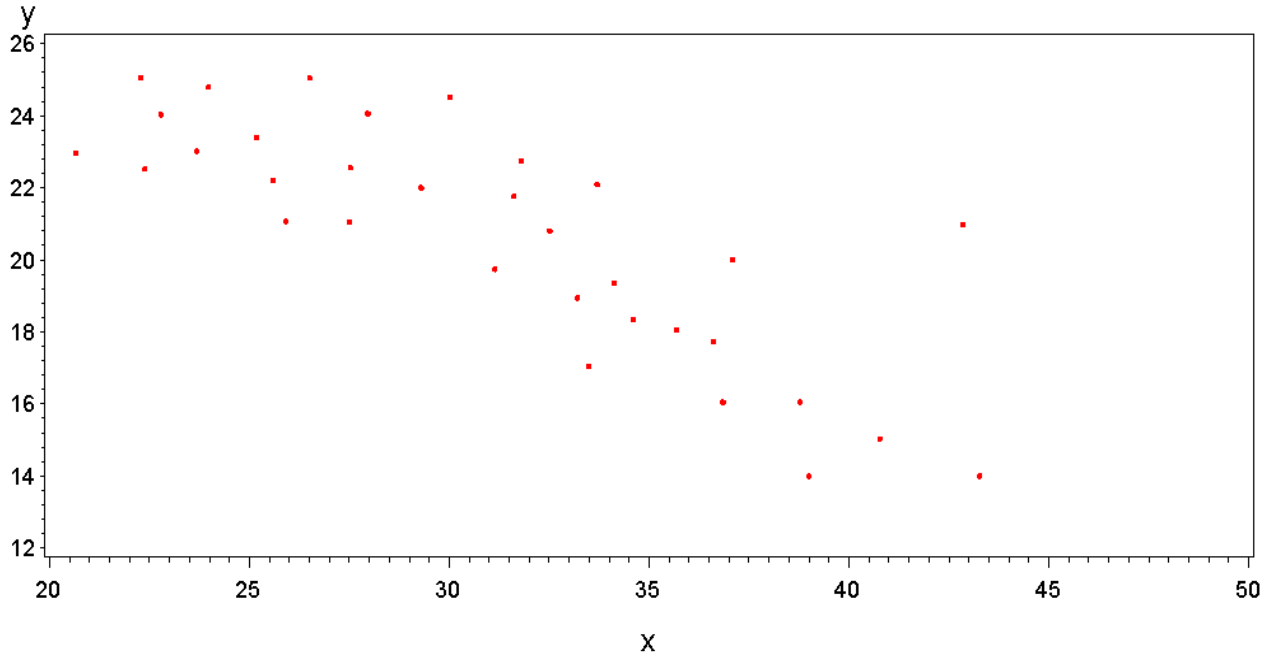


Figure 6. The Dark Dots Show the 33 Scatter Sport Identified by PRCOC MODECLUS

To visualize the precision of the scatter sport recognition, following figure overlaps the original scatter plots (circle) and the derived clusters (dot) together. All the dots fall right on the center of the circles. It means that all the spots on the scatter plot were captured precisely.

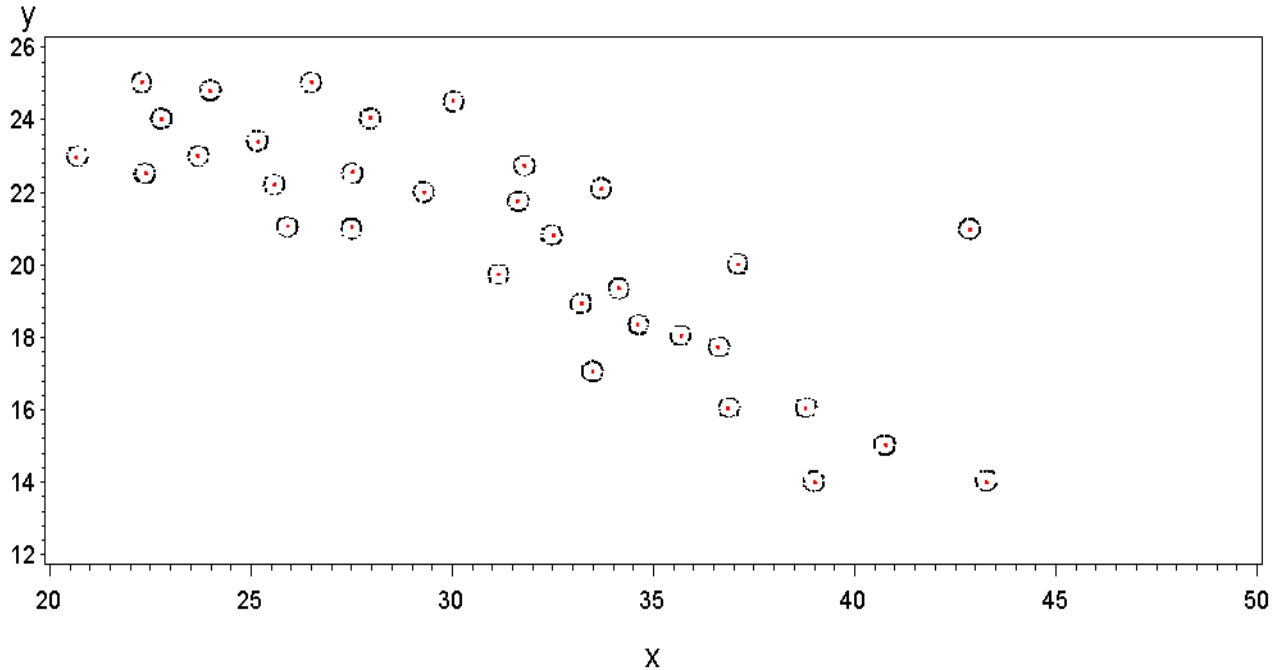


Figure 7. Overlap of the Original Scatter Spots (Circle) and Scatter Spots (Dot) Identified by Proc Modeclus.

DISCUSSION

In the real world, the symbol of the scatter spot is not necessary to be a circle. It can be dot, star, square, triangle, heart, and so on. This proposed non-parametric clustering method is able to detect all the symbols because of its capability of detecting clusters in highly irregular shapes. Figures 8 and 9 show the precision of this method in recognizing the star symbol and heart symbol.

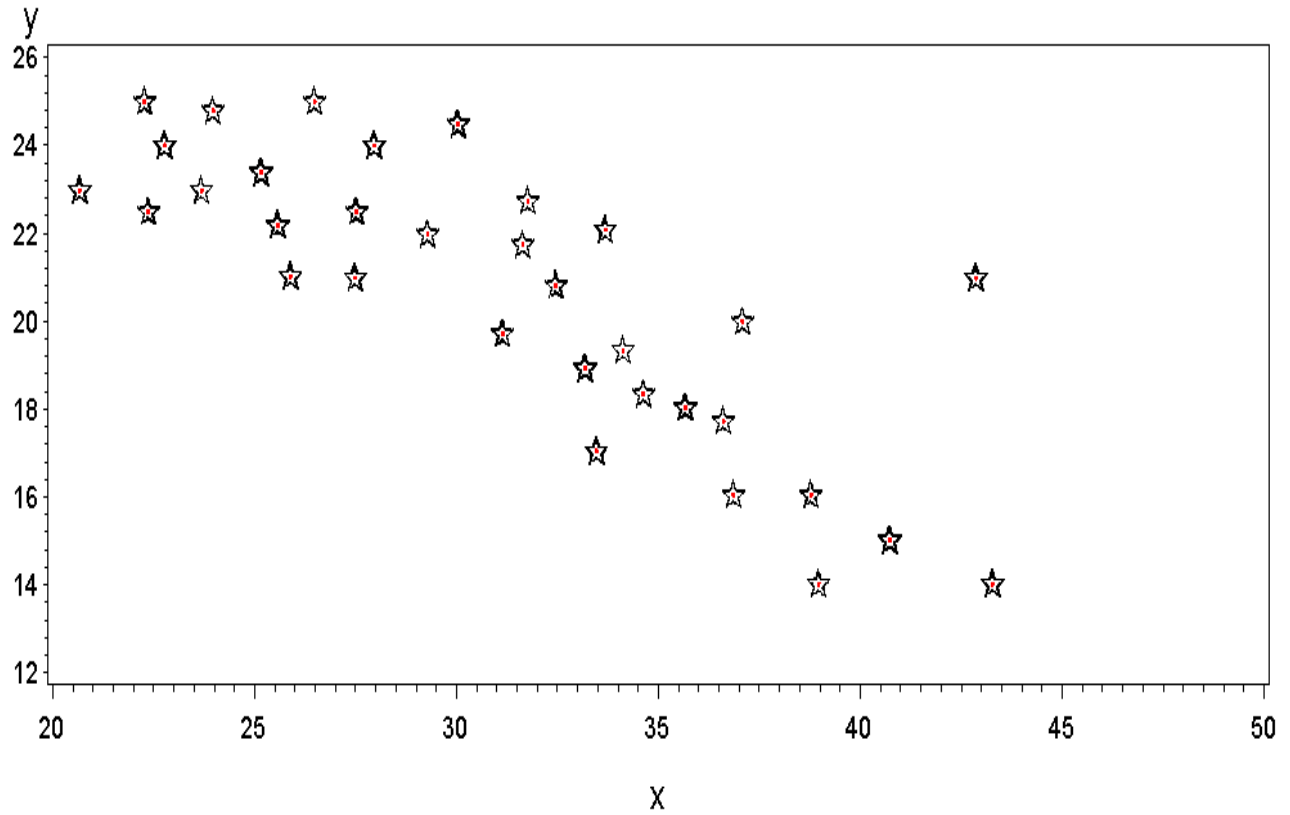


Figure 8. Overlap of the Original Scatter Spots (Star) and Scatter Spots (Dot) Identified by Proc Modeclus.

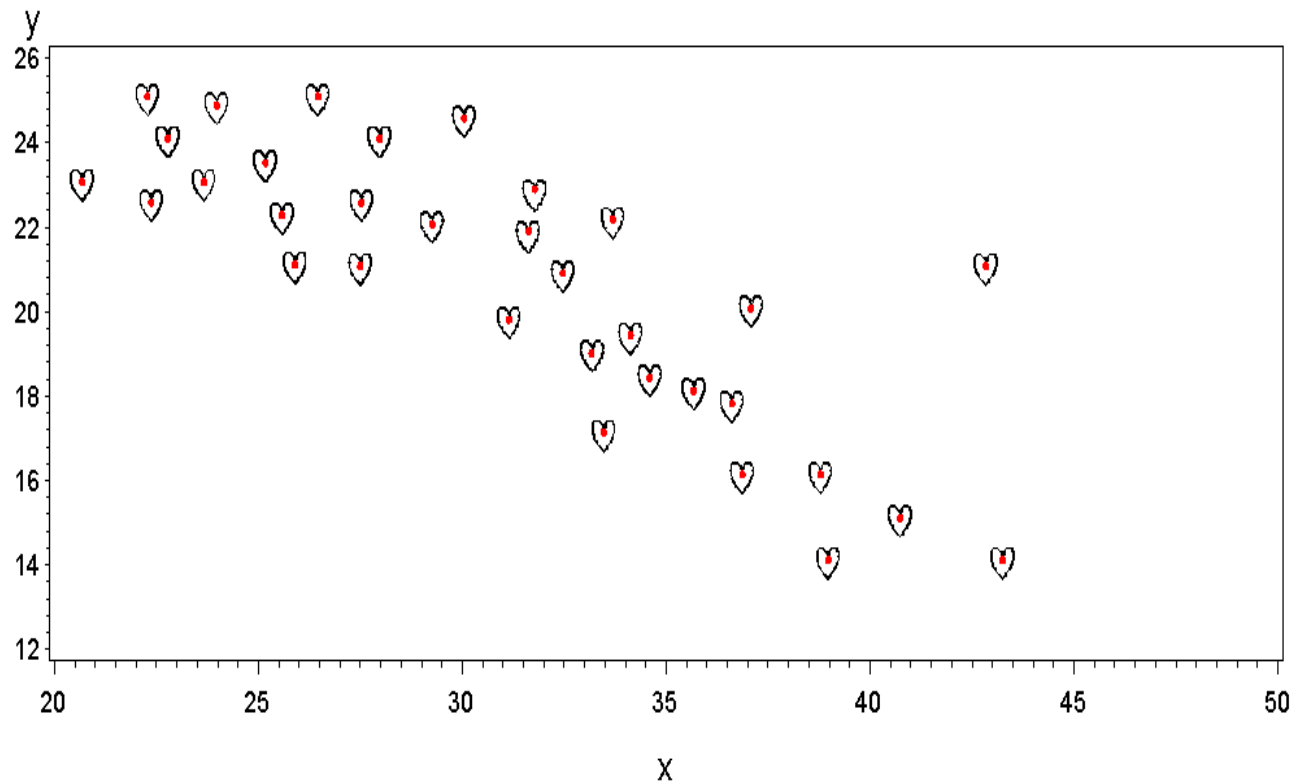


Figure 9. Overlap of the Original Scatter Spots (Heart) and Scatter Spots (Dot) Identified by Proc Modeclus.

A major challenge of a recognizing scatter plot is the situation of spot overlap. Like other methods of image recognition, this non-parametric cluster analysis method has limited capability to identify the spots when they are overlapped together completely or partially. Starting from this exploration method, more advanced techniques or semi-manual methods can be developed to handle this situation.

CONCLUSION

The high efficiency and consistency make this proposed image recognition method especially helpful for capturing tabular data from a scatter plot containing large number of spots. As an automated method, it not only reduces tedious manual work, but also leaves few space for human error.

In pharmaceutical industry, SAS is traditionally used for data analysis and report creation of clinical trials. Beyond these applications, statistical analysis is more and more widely used in pharmaceutical industry, including image recognition in microarray analysis, data mining in big genetic data, and so on. The application and strategy introduced in this article illustrates that, as a sophisticated statistical software, SAS can be a powerful and convenient tool for recognizing and processing image data. Enhanced with this potential, SAS is able to play a more significant role in pharmaceutical and clinical research.

REFERENCES

1. Brian Fairfield-Carter (2009). Reverse-engineer a Reference Curve: Capturing Tabular Data from Graphical Output. Proceedings of the PharmaSUG 2009, Paper CC23.
2. Web site of GetData, <http://www.getdata-graph-digitizer.com/>
3. Web site of WinDig, <http://www.unige.ch/sciences/chifi/cpb/windig.html>
4. Web site of Engauge Digitizer, <http://digitizer.sourceforge.net/>
5. Web site of Digitizer for Origin, <http://www.originlab.com/fileexchange/details.aspx?fid=8>
6. Web site of DataThief, <http://datathief.org/>
7. Tryon, Robert C. (1939). *Cluster Analysis: Correlation Profile and Orthometric (factor) Analysis for the Isolation of Unities in Mind and Personality*. Edwards Brothers.
8. Bailey, Ken (1994). "Numerical Taxonomy and Cluster Analysis". *Typologies and Taxonomies*. p. 34. ISBN 9780803952591.
9. Cluster analysis, Wikipedia, http://en.wikipedia.org/wiki/Cluster_analysis
10. SAS/STAT® 9.2 User's Guide Introduction to Clustering Procedures, <http://support.sas.com/documentation/cdl/en/statugclustering/61759/PDF/default/statugclustering.pdf>
11. Silverman, B. W. (1986). *Density Estimation*, New York: Chapman & Hall.
12. SAS/STAT® 9.2 User's Guide The MODECLUS Procedure, <http://support.sas.com/documentation/cdl/en/statugmodeclus/61808/PDF/default/statugmodeclus.pdf>
13. Brian Fairfield-Carter (2010). Capturing Tabular Data from Graphical Output Part 2: Web and Windows-Based Tools. Proceedings of the PharmaSUG 2010, Paper AD12.

ACKNOWLEDGMENTS

I would like to thank Celgene for consistently encouraging innovation and supporting conference participation.

CONTACT INFORMATION

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

Zhaojie Wang
Celgene Corporation
300 Connell Dr., Berkeley Heights, NJ 07922
E-mail: ZJWANG2006@GMAIL.COM

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.