Graphical Analysis: Concurrently Visualizing Correlationship and Interrelationship

J. Y. Choe

1. INTRODUCTION and PREVIOUS APPROACHES

Information is a key factor in the world today. Given the wide range of statistical information available to people through a variety of means, the visual representation of information is becoming increasingly important. To this end, Wallgren et al. (1996) said, "Society has become more and more dependent on statistics and other numerical information. However, all this numerical information is meaningless if it cannot be presented in a proper and easily accessible way. Charts and maps are effective aids to those who need to illustrate numerical data."

In addition to performing sound statistical analysis, creating diagrams and other visual aids to assist the reader in comprehending such analysis is critically important. Fisher (1990) said, "The preliminary examination of data is facilitated by the use of diagrams. Diagrams prove nothing, but bring outstanding features readily to the eye: they are therefore no substitute for such critical tests as may be applied to data, but are valuable is suggesting such tests, and in explaining conclusions founded upon them."

Many researchers have proposed methods and techniques of visualizing statistical graphs consisting of complex ideas communicated with clarity, precision, and efficiency. Such studies have been conducted by Barton (1999), Basford and Tukey (1999), Wallgren et al. (1996), Henry (1995), Tufte (1984), Chambers et al. (1983), Everitt (1978), etc.

Furthermore, as shown by Pardoe and Cook (2002), Fisher and Switzer (2001), Hanley et al. (2001), Lee et al. (2000), Almond et al. (2000), Doane and Tracy (2000), Cook and Weisberg (1999), Rousseeuw et al. (1999), Korn and Graubard (1998), Dawson et al. (1997), Murdoch and Chow (1996), Dallal and Finseth (1977), Rybak (1975), Hettmansperger and

McKean (1974), Anscombe (1973), Hsiao (1972), and Weiss (1970), many papers have been put forward new conceptions of displaying informative graphs based on survey data, statistical data, and analysis results.

Among informative means of representing survey and analytical data, recently, Structural Equation Modeling (SEM), often utilizing graphical path diagrams, has become a widely known and convenient framework for statistical factor analysis. This particular technique incorporates several traditional multivariate procedures as special cases, such as factor analysis, regression analysis, discriminant analysis, and canonical correlation analysis.

To further clarify and articulate statistical results, as well as to streamline reporting procedures, a number of software programs have been developed. Programs with advanced functions such as LISREL, EQS, and AMOS have also become popular among many researchers. Quite similar in their range of functions, reviews of these packages by Miles (1998), Hox (1995), and Waller (1993) state that, although each program has its own strengths and weaknesses, for standard analyses, any package can be used.

In studies that investigate the interrelationship between the dependent variable and independent variables of a particular phenomenon, it is assumed that the analysis process is very simple. First, a correlation analysis is used to investigate the relationship between variables and to choose significant variables for the following regression-analysis; then, the interrelationship is clarified by regression analysis.

The above techniques and programs are effective in displaying analytical results for each process. However, in order to further clarify statistical information, such as simultaneously visualizing the results of two or more analyses, new techniques are necessary. If the outputs of the correlation analysis and the regression analysis can be visualized concurrently in one diagram, readers can intuitively comprehend the overall image of analyzing the interrelationship between cause and effect, and researchers can make efficient use of limited report space and

presentation time.

Therefore, this paper proposes Graphical Analysis (GA) as a method of concurrently visualizing the results of more than one analysis when examining the interrelationship between cause and effect. Section 2 of this paper describes the GA concept and provides two practical examples of using GA. Further implications for analysis are given in the conclusion.

2. AN IDEA: GRAPHICAL ANALYSIS

GA is a visualization technique by which the results of more than one analysis are jointly presented in one diagram. This technique does not require the installation of specialized mathematical models, and only the results of the analysis are consolidated by combining statistical analysis. Using this technique allows us to easily and inclusively read specific characteristics from one diagram. This in turn simplifies reporting and presenting multiple types of information.

For example, in the case of correlation and regression analyses, correlationship and interrelationship can be grasped and explained simultaneously with one diagram. The relative spatial-position, correlationship, and interrelationship, as well as other relationships among variables can be analyzed intuitively from one diagram by using the GA technique.

My first example illustrates the following detailed analytical process. Initially, I arbitrarily selected the task "Analysis of influence factors of travel demand on evacuee-trip production at the aggregate level in an earthquake" from Choe (2002) (see Table 1). As shown in Table 1, the analysis variables consist of three items: demand characteristics, zone characteristics, and damage characteristics. For the purposes of simplicity, I will omit detailed explanations of each variable.

The aim of this task is to investigate factors that influence the production of travel demand on evacuee-trips.

Table 1. Analysis variables

Feature	Variables	Labels for diagram
Dependent variables	A. Demand characteristics	
	a. Number of evacuee trips per person as daily demand (6:00 to 24:00)	Daily demand
	b. Number of evacuee trips per person in period I (6:00 to 9:00)	Demand I
	c. Number of evacuee trips per person in period I (10:00 to 12:00)	Demand II
	d. Number of evacuee trips per person in period III (13:00 to 24:00)	Demand III
Independent variables	B. Zone characteristics	
	e. Population density	Pop. density
	f. Number of average family members	Family members
	g. Proportion of infants	Prop. of infants
	h. Average distance between house	Dist. house
	C. Damage characteristics	
	i. Rate of collapsed all buildings	Rate bldg.(c)
	j. Rate of collapsed detached houses	Rate house(c)
	k. Rate of collapsed apartment buildings	Rate apt.(c)
	l. Rate of partially destroyed all buildings	Rate bldg.(d)
	m. Rate of partially destroyed detached-houses	Rate house(d)
	n. Rate of partially destroyed apartment buildings	Rate apt.(d)
	o. Rate of fire loss to all buildings	Rate bldg.(f)
	p. Rate of fire loss to detached-houses	Rate house(f)
	q. Rate of fire loss to apartment buildings	Rate apt.(f)

Source: Choe (2002).

At this point, I conducted a simple analytical process consisting of two steps (correlation analysis and regression analysis) so as to make the results more understandable.

In this case, analyses generally produce two results, namely correlation coefficients and regression coefficients, however regression coefficients consist of four dependent variables: Daily demand, Demand I, Demand II, and Demand III. Tables 2 and 3 illustrate matrices of correlation coefficients and summarized regression coefficients, respectively. While these methods and techniques are very useful for visualizing the results of such analyses, these are unable to project the results of two or more analyses onto one diagram. Furthermore, readers may find it difficult to comprehend the correlationship and interrelationship of this statistical information and to compare the relationships of the four dependent variables.

Here, I suggest GA can be utilized as a technique to project the results

Table 2. Correlation Matrix for 17 Variables

	Daily demand	Demand I	Demand II	Demand III	Pop. density	Family members	Prop. of infants	Dis. house	Rate house(c)	Rate house(d)	Rate house(f)	Rate apt.(c)	Rate apt.(d)	Rate apt.(f)	Rate bldg.(c)	Rate bldg.(d)	Rate bldg.(f)
Daily demand	1.00	0.72	0.45	0.70	0.11	-0.14	-0.08	-0.17	0.54	0.29	-0.22	0.58	0.23	0.02	0.51	0.53	-0.20
Demand I	0.72	1.00	0.22	0.11	0.08	-0.17	-0.01	-0.04	0.45	0.18	-0.14	0.47	0.17	0.01	0.44	0.35	-0.14
Demand II	0.45	0.22	1.00	0.01	0.12	-0.17	0.05	-0.07	0.28	0.27	-0.11	0.25	0.16	-0.00	0.24	0.28	-0.05
Demand III	0.70	0.11	0.01	1.00	0.04	0.02	-0.15	-0.20	0.30	0.17	-0.17	0.36	0.14	0.03	0.29	0.38	-0.16
Pop. density	0.11	0.08	0.12	0.04	1.00	-0.17	-0.06	0.12	0.26	0.25	-0.02	0.18	0.15	0.06	0.23	0.27	-0.17
Family members	-0.14	-0.17	-0.17	0.02	-0.17	1.00	0.28	-0.02	-0.17	-0.18	-0.04	-0.14	-0.07	-0.14	-0.17	-0.16	-0.03
Prop. of infants	-0.08	-0.01	0.05	-0.15	-0.06	0.28	1.00	0.17	0.08	-0.08	0.08	-0.02	0.03	0.03	0.05	0.00	0.02
Dis. house	-0.17	-0.04	-0.07	-0.20	0.12	-0.02	0.17	1.00	0.05	0.01	0.05	-0.09	-0.06	-0.02	0.03	-0.04	0.15
Rate house(c)	0.54	0.45	0.28	0.30	0.26	-0.17	0.08	0.05	1.00	0.27	-0.25	0.84	0.29	0.06	0.71	0.71	-0.24
Rate house(d)	0.29	0.18	0.27	0.17	0.25	-0.18	-0.08	0.01	0.27	1.00	-0.14	0.33	0.53	0.13	0.55	0.46	-0.11
Rate house(f)	-0.22	-0.14	-0.11	-0.17	-0.02	-0.04	0.08	0.05	-0.25	-0.14	1.00	-0.24	-0.25	0.62	-0.26	-0.30	0.09
Rate apt.(c)	0.58	0.47	0.25	0.36	0.18	-0.14	-0.02	-0.09	0.84	0.33	-0.24	1.00	0.33	0.11	0.70	0.72	-0.22
Rate apt.(d)	0.23	0.17	0.16	0.14	0.15	-0.07	0.03	-0.06	0.29	0.53	-0.25	0.33	1.00	0.01	0.40	0.49	-0.18
Rate apt.(f)	0.02	0.01	-0.00	0.03	0.06	-0.14	0.03	-0.02	0.06	0.13	0.62	0.11	0.01	1.00	0.07	-0.05	0.20
Rate bldg.(c)	0.51	0.44	0.24	0.29	0.23	-0.17	0.05	0.03	0.71	0.55	-0.26	0.70	0.40	0.07	1.00	0.57	-0.23
Rate bldg.(d)	0.53	0.35	0.28	0.38	0.27	-0.16	0.00	-0.04	0.71	0.46	-0.30	0.72	0.49	-0.05	0.57	1.00	-0.30
Rate bldg.(f)	-0.20	-0.14	-0.05	-0.16	-0.17	-0.03	0.02	0.15	-0.24	-0.11	0.09	-0.22	-0.18	0.20	-0.23	-0.30	1.00

Source: Choe (2002). Note: Significant level is omitted.

Table 3. Regression Analysis of Travel Demand

Independent variables	Dependent variables			
	Daily demand	Demand I	Demand II	Demand III
Family members	0.36	-	-	0.91
Prop. of infants	-	-	-	-0.37*
Dist. house	-0.19*	-	-	-0.28*
Rate bldg.(c)	0.24	0.35	-	0.49
Rate apt.(c)	0.25	0.41	0.33	-
Rate bldg.(d)	0.25	-	-	-
Rate house(d)	-	-	0.27	-
r^2	0.76	0.53	0.32	0.58

Source: Choe (2002). Note: 1. The level of significance is $P<0.01$, but yet "*" indicates significance to be at the level of $P<0.05$. 2. Regression analyses used in the source have been formatted by equations without constants.

onto one diagram that, in turn, will greatly benefit researchers in reading multiple results. It should be emphasized that GA, shown in Figure 1, is different from the general process for Multivariate Statistical Analysis (MSA). For example, usually the correlation analysis is displayed before the regression analysis; nevertheless, as shown in Figure 1, they have been plotted together on same process line. GA, as a graphical technique for displaying results, makes them easier to read and intuitively understand. Yet, prior to constructing the GA illustration, we need to prepare the results of correlation analysis and regression analysis. The process is basically as follows:

1. Construct a correlation matrix (M) such as Table 2 by correlation analysis and regression coefficients (β) by regression analyses. It is necessary to recalculate M because the matrix will eventually be used

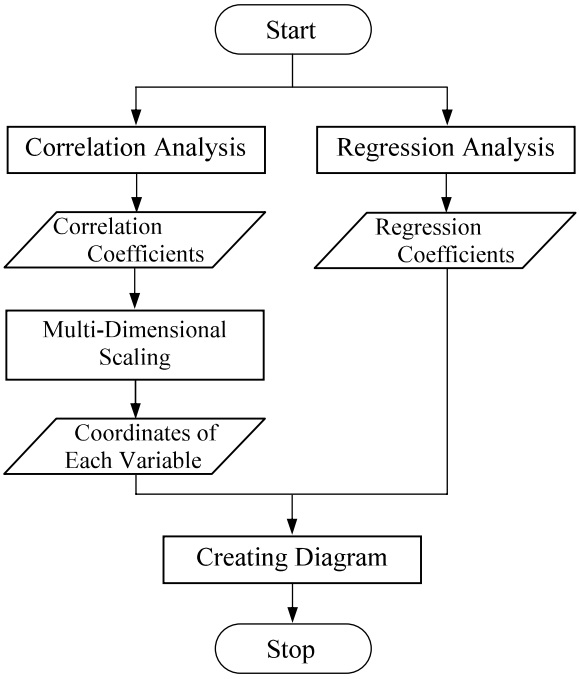


Figure 1. The Graphical Analysis (GA) Process

for inputting to Multi-Dimensional Scaling (MDS).

The recalculated correlation matrix M_r is

$$M_r = A - M, \quad (1)$$

where M is the square matrix of correlation coefficient and A is a matrix of ones. Here, A minus M is used to set the value 1 of correlation for MDS to 0.

2. Obtain coordinates (S_{xy}) of each variable by MDS using the recalculated matrix. S_{xy}, which will in turn define relative space as a conceptual distance among variables.

The conceptual distance shows notional position relations by the correlation coefficient. Generally, notional position relations are two points that are approached if similarity exists between events. If they are dissimilar, they are two points that are distantly separated and expressed by the Euclid distance.

3. Draw a diagram using S_{xy} and β. I plot the space location of the variables by S_{xy} as shown in Figure 2. To distinguish between dependent and independent variables, dependent variables are illustrated as an ellipse with a screen tone background. When the distance between variables is close, the correlationship between the variables is high and positive. Similarly, the correlationship is low or negative when the distance between variables is great. Moreover, to clarify correlationship between dependent and independent variables, independent variables with negative correlationship are depicted as ellipses with a broken line.

Next, I indicate the interrelationship by connecting the lines between the dependent variables and each independent variable that has a significant regression coefficient β calculated from regression analysis to be at the level of $P<0.01$ or 0.05. Significance levels of $P<0.01$ and

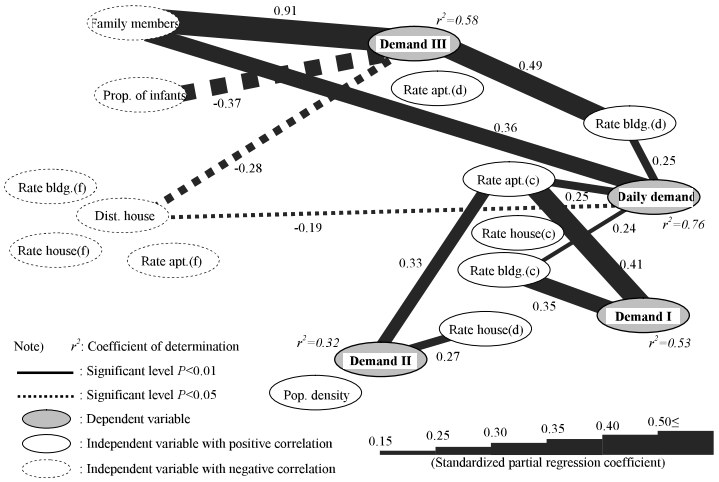


Figure 2. Relationship of Travel Demand on Evacuee-Trip Production with Factors by Graphical Analysis

0.05 are symbolized as such by solid lines and broken lines, respectively. The influence magnitude is displayed by line thickness (the thicker the line, the greater the influence), as shown in Figure 2, and indicates the influence of the independent variable on the dependent variable.

Finally, all regression analysis statistics are specified by writing the value of regression coefficients beside connected lines to easily read the thickness of such lines, the coefficient of determination r^2, and the legend.

To further illustrate the GA process, I am including a further example using data from Tsujinaka (1999). This data was collected to reveal the structure of civil society in a given nation or region, how it is related to the government, the market, and the political system, and how it is constrained by the government, the market, and the political system. Detailed information regarding this research can be obtained from CSC (2008).

This second example of GA is an analysis for grasping correlationship and interrelationship between factors of self-assessment influence regarding policy issues in areas in which civil-society interest groups

conduct activities, and comparing the relationship structure of Japan and Korea. Though the contents of data are indicated in Table 4, I have omitted the detailed explanations of the statistics of the first example and only demonstrate the final result by GA (see Figure 3).

The number of variables (30) set for this example is higher than that used in the previous example. Despite this, the process of investigating correlationship between variables and interrelationship between the

Table 4. Analysis variables

Feature	Variables	Labels for diagram
Dependent variable	A. Influence (self-reported)	Influence
Independent variables	B. Organization characteristics	
	a. Organization type	Org. type
	b. Influence area	Infl. area
	c. Political aim	Pol. aim
	d. Ideology (organization)	Ideo. org.
	e. Ideology (members)	Ideo. mem.
	f. Establishment year	Est. year
	C. Actor relationship	
	g. National-level political relationships	Nat. rel.
	h. Local-level political relationships	Loc. rel.
	i. Policy coordination	Pol. coord.
	j. Politician relationships	Polit. rel.
	k. Contact with government agencies	Govt. cont.
	l. Contact with ruling party (parties)	Rul. cont.
	m. Contact with opposition party (parties)	Opp. cont.
	n. Provision of information to mass media	Mass media
	D. Activities	
	o. Budgeting	Budgeting
	p. Election-related	Electing
	q. Lobbying (general)	Lobbying
	E. Influence on specific events	
	r. Participation	Participation
	s. Policy determination	Pol. det.
	F. Results of successful of influence on policy decisions	
	t. Policy implementation	Pol. imp.
	u. Policy cancellation/revision	Pol. ccl./rev.
	G. Organization resources	
	v. Number of individual members	Num. ind.
	w. Number of organizational members	Num. org.
	x. Number of affiliated members	Num. aff.
	y. Number of full-time employees	Num. full
	z. Number of part-time employees	Num. part
	aa. Organization income	Org. inc.
	ab. Subsidy from the national government	Nat. sub.
	ac. Subsidy from the local government	Loc. sub.

Source: Tsujinaka (1999).

dependent variable (such as self-assessment influence) and independent variables is the same.

Given the high number of variables, the information that would have to be contained in tables such as those used in the first example (Tables 2 and 3) is far more complicated and difficult for the reader to intuitively comprehend. Furthermore it is not easy to compare concurrently correlationship and interrelationship in two or more cases, such as the cross-national comparison of Japan and Korea in this example.

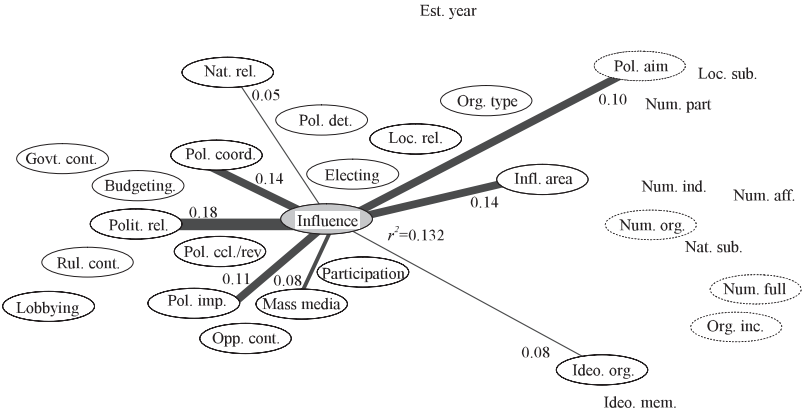
Initially, I performed an interrelationship analysis by multiple analysis of variance (MANOVA) to reveal the main and interaction effects of categorical variables on multiple dependent interval variables. At this point, to effect the GA process as shown in Figure 1, I substituted Regression Analysis and Regression Coefficients for MANOVA and Covariate Coefficients, respectively. The process, which is basically the same as the previous example, is as follows:

1. Construct a correlation matrix (M) for the appropriate number of variables (in this case, 30) by correlation analysis, and covariate coefficients (β) by MANOVA. It is necessary to recalculate M as shown in equation (1).

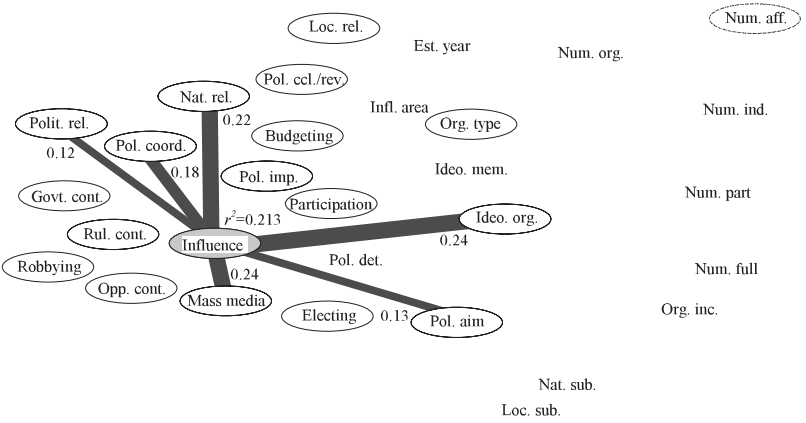
2. Obtain coordinates (S_{xy}) of each variable by MDS using the recalculated matrix.

3. Draw a diagram using S_{xy} and β. I plot the spatial location of the variables by S_{xy} as shown in Figure 3. To distinguish between dependent and independent variables, dependent variable is illustrated as an ellipse with a screen tone background. Furthermore, to clarify correlationship between dependent and independent variables, independent variables with negative correlationship are depicted as ellipses with broken lines. The variables symbolized as ellipses are at $P<0.05$.

The process here is similar to the first three steps described in the first example above. The procedure is almost identical except that it



(a) Japan



(b) Korea

(Standardized partial covariate coefficient)

Note)　r^2: Coefficient of determination
: Significant level P<0.05
: Dependent variable at P<0.05　: Independent variable with positive correlation at P<0.05
: Independent variable with negative correlation at P<0.05

Figure 3. Relationship of Self Assessment Influence in Policy Issue with Factors by Graphical Analysis

allows for visualization of covariate coefficients instead of regression coefficients. At this point, the lines indicating the interrelationship between the dependent variable and each independent variable are connected, and the value of covariate coefficients, the coefficient of determination r^2 are written in the legend. Here, the significance level is set to $P<0.05$ to fix a significant covariate coefficient β for connecting the line, creating the solid line. The final diagram clearly shows the relationship among variables and can intuitively compare the two cases. In this manner, this example proves the flexible utilization and the expansiveness of GA in comparison analysis.

The above processes involved in creating the diagram are not fixed but adaptable in response to the desired analysis. Certainly different contexts and cases require modifications in the content of the diagrams that investigate interrelationships among variables, such as regression analyses with constants, correlation analyses with positive and negative correlationships between dependent and independent variables, and structure models including MANOVA and discriminant analysis.

3. CONCLUSION

In this paper, I have proposed GA as a visualization technique to allow us to easily and inclusively read multiple statistical information from one diagram. GA does not entail a mathematical model, but rather, is inherently composed of present multivariate statistical analysis.

Due to the simplicity and flexibility of the GA process from the viewpoint of utilization, GA has a number of advantages over other visualization methods. When there are multiple levels of information (such as correlation coefficients, positive and negative relationship, regression or covariate coefficients, significant level, and other statistics) to explain, GA can be used as a feature by which only pertinent information is transmitted, allowing for data analysis to be presented in a comprehensible, simple, and efficient manner.

These examples have clearly shown the utility of GA for exhibiting correlationship and interrelationship between independent and dependent variables.

REFERENCES

Almond, R. G., Lewis, C., Tukey, J. W., and Yan, D. (2000), "Displays for Comparing a Given State to Many Others," *The American Statistician,* 54, 89–93.

Anscombe, F. J. (1973), "Graphs in Statistical Analysis," *The American Statistician,* 27, 17–21.

Barton, R. R. (1999), *Lecture Notes in Statistics,* NY: Springer.

Basford, K. E., and Tukey, J. W. (1999), *Graphical Analysis of Multiresponse Data Illustrated With a Plant Breeding Trial,* NY: Chapman & Hall/CRC.

CSC (2008), *The Special Project on Comparative Civil Society, the State and Culture: New Values in the Age of Globalization and Reconstruction of the Public,* Tsukuba: University of Tsukuba. http://csc.tsukuba.ac.jp/csc/index.html. Access date: December 12, 2008.

Chambers, H. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983), *Graphical Methods for Data Analysis,* BT: Duxbury Press.

Choe, J. Y. (2002), A Fuzzy-Neural Network Model for Travel Demand on Evacuee-Trip Production at the beginning of the Great Hanshin-Awaji Earthquake, *Journal of Social Safety Science,* 4, 31–40 (in Japanese with English abstract).

Cook, R. D., and Weisberg, S. (1999), "Graphs in Statistical Analysis: Is the Medium the Message?" *The American Statistician,* 53, 29–37.

Dallal, G., and Finseth, K. (1977), "Double Dual Histograms," *The American Statistician,* 31, 39–41.

Dawson, K. S., Gennings, C., and Carter, W. H. (1997), "Two Graphical Techniques Useful in Detecting Correlation Structure in Repeated Measures Data," *The American Statistician,* 51, 275–283.

Doane, D. P., and Tracy, R. L. (2000), "Using Beam and Fulcrum Displays

to Explore Data," *The American Statistician,* 54, 289–290.

Everitt, B. S. (1978), *Graphical Techniques for Multivariate Data,* London: Heinemann Educational Books Ltd.

Fisher, N. I., and Switzer, P. (2001), "Graphical Assessment of Dependence: Is a Picture Worth 100 Tests?" *The American Statistician,* 55, 233–239.

Fisher, R. A. (1990), *Statistical Methods, Experimental Design, and Scientific Inference,* Oxford: Oxford University Press.

Hanley, J. A., Joseph, L., Platt, R. W., Chung, M. K., and Bélisle P. (2001), "Visualizing the Median as the Minimum-Deviation Location," *The American Statistician,* 55, 150–152.

Henry, G. T. (1995), *Graphing Data: Techniques for Display and Analysis,* London: Sage Publications.

Hettmansperger T. P., and McKean, J. W. (1974), "A Graphical Representation for Non-Parametric Inference," *The American Statistician,* 28, 100–102.

Hox, J. J. (1995), AMOS, Eqs and Lisrel for Window: A comparative review, *Structural Equation Modeling,* 2, 79–91.

Hsiao, F. S. T. (1972), "The Diagrammatical Representation of Confidence-Interval Estimation and Hypothesis Testing," *The American Statistician,* 26, 28–29.

Korn, E. L., and Graubard, B. I. (1998), "Scatterplots With Survey Data," *The American Statistician,* 52, 58–69.

Lee, J. J., Hess, K. R., and Dubin, J. A. (2000), "Extensions and Applications of Event Charts," *The American Statistician,* 54, 63–70.

Lee, J. J., and Tu, Z. N. (1997), "A Versatile One-Dimensional Distribution Plot: The BLiP Plot," *The American Statistician,* 51, 353–358.

Miles, J. (1998), Review type: Statistical analysis/structural equation modeling, *Psychology Software News,* 8, 2, 58–65.

Murdoch, D. J., and Chow, E. D. (1996), "A Graphical Display of Large Correlation Matrices," *The American Statistician,* 50, 178–180.

Pardoe, I., and Cook, R. D. (2002), "A Graphical Method for Assessing the Fit of a Logistic Regression Model," *The American Statistician,*

57, 263–272.
Rousseeuw, P., Ruts, I., and Tukey, J. W. (1999), "The Bagplot: A Bivariate Boxplot," *The American Statistician,* 53, 382–387.
Rybak, J. (1975), "Diagrams for Set Theory & Probability Problems of Four or More Variables," *The American Statistician,* 29, 91–93.
Tsujinaka, Y. (1999), *Cross-national Survey on Civil Society Organizations and Interest Groups (JAPAN) J-JIGS Code Book,* Tokyo: LDB (in Japanese).
Tsujinaka, Y. (1999), *Cross-national Survey on Civil Society Organizations and Interest Groups (KOREA) K-JIGS Code Book,* Tokyo: LDB (in Japanese).
Tufte, E. R. (1984), *The Visual Display of Quantitative Information,* Cheshire: Graphics Press.
Waller, N. G. (1993), Software review. Seven CFA programs: EQS, EzPATH, LICS, LISCOMP, SIMPLIS, and CALIS. *Applied Psychological Measurement,* 14, 73–100.
Wallgren, A., Wallgren, B., Persson, R., Jorner, U., and Haaland, J. A. (1996), *Graphing Statistics & Data,* CA: Sage Publications.
Weiss, N. S. (1970), "A Graphical Representation of the Relationships Between Multiple Regression and Multiple Correlation," *The American Statistician,* 24, 25–29.