VALORES EXTREMOS EN GEOGRAFIA: EDA, ESTADISTICA ROBUSTA Y GRAFICOS

Roy P. Bradshaw **Vicente Rodríguez**

Universidad de Nottingham C.S.I.C., MADRID

INTRODUCCION

La segunda mitad de los años 80 marca una frontera que separa los contenidos de unos libros de Estadística, basados en ejemplos explicados en el texto, de otros cuya base son casos semejantes, pero ejecutados mediante programas de ordenador. Esta transición es el resultado de la presencia total del ordenador en los ámbitos académicos y científicos, una realidad imposible de sortear además de no ser deseable hacerlo.

Expresado esto así, parece como si el ordenador fuera más un peligro que un poderoso instrumento. El individuo es capaz de pensar en una organización científica del trabajo, pero la ejecuta lentamente y muchas veces con errores. La máquina, en cambio, realiza estas tareas de manera exacta y rápida, aunque no sea capaz de pensar en la lógica de los procesos. Estamos, pues, en una época en la que la conjunción del hombre y el ordenador puede conseguir enormes progresos, si se equilibra bien esta unión. Los peligros, estando latentes, no serían destacables.

Mientras cada día los costes del *hardware* son más bajos para productos de mayor calidad y capacidad de memoria y cálculo y los programas son mas abundantes y diversos, también aumenta la dimensión que han adquirido las bases de datos. Estas son las condiciones básicas para el uso generalizado de ordenadores.

De esta manera, la investigación científica discurre hoy por unos derroteros, en los que parece que el "hombre" va a remolque de las "máquinas". Como ha señalado Fox (1990) parece como si los investigadores hubieran sido desplazados de sus datos, ignorando la información que manejan, siendo hoy más fácil detenerse en un análisis multivariante que en un simple histograma. Ante los resultados obtenidos en el primero, fácilmente calculados, tiende a ocultarse la importancia del segundo como instrumento de análisis. Este es el camino que lleva a la pérdida de

una "cierta subjetividad" que debe primar en el análisis estadístico ante el automatismo tecnológico del procesado de los datos (Berger y Berry, 1988).

Es necesario, pues, recuperar el análisis de datos orientado hacia la investigación científica, separándola claramente del nivel popular, el de los medios de comunicación, y del nivel de los profesionales de la Estadística como ciencia (McPherson, 1988). El camino para ello es el análisis de datos, como una de las fases esenciales del proceso de investigación en Ciencias Sociales (fig. 1). Como también sucede con otras ciencias aplicadas al conocimiento de realidades humanas o territoriales, el análisis estadístico es "relevante para aquellos proyectos en los que la información recogida se presenta por números" (Healey, 1990).

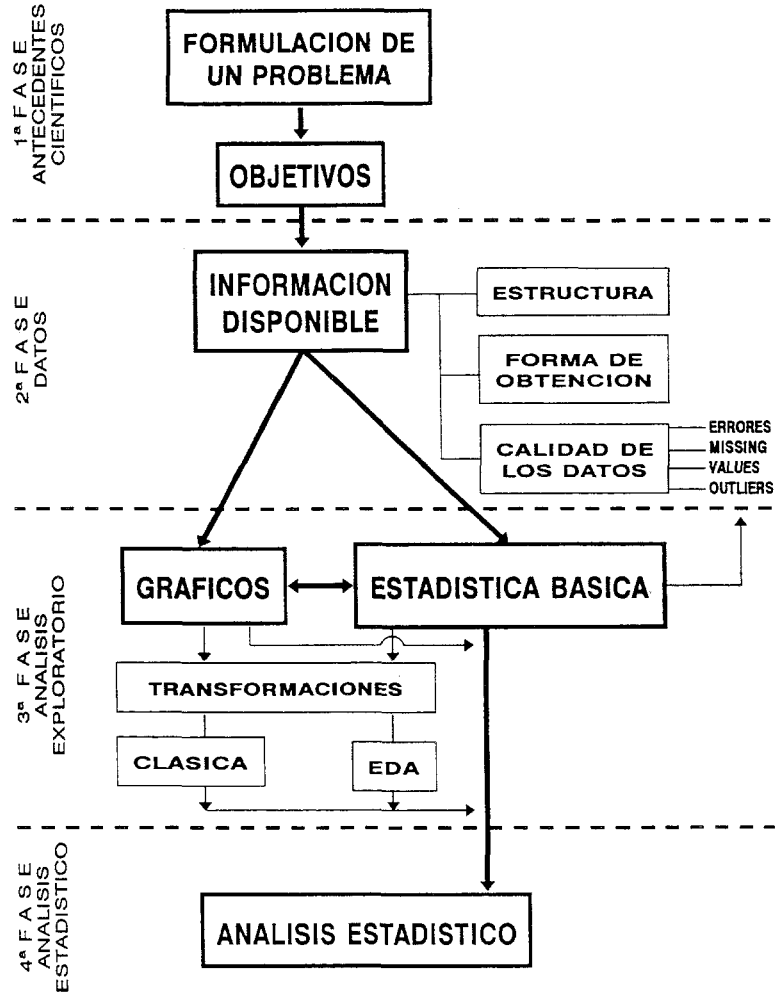

Figura 1.- Esquema de investigación

El objeto de esta ponencia es insistir en un argumento, el análisis de datos, no necesariamente nuevo, pero que no termina de despegar entre las Ciencias Sociales, y en concreto en la Geografía. Se dividirá en su doble vertiente, numérica y gráfica. Uno de los hechos que hacen deseable este objetivo es el de las posibles deficiencias existentes en cualquier información numérica, especialmente en las más voluminosas, expresadas en forma de errores, y los valores extremos integrados en el conjunto de datos. A continuación se efectuarán algunas reflexiones sobre análisis exploratorio, estadística robusta y análisis gráfico de datos, con la exposición de algunas de sus ventajas a través de ejemplos prácticos ejecutados manualmente y con referencias a algunos programas estadísticos *standard* que incorporan estas técnicas.

LOS DATOS Y EL ANALISIS ESTADISTICO Y GRAFICO

De los elementos que básicamente componen el proceso de investigación (fig. 1) quizás sea el análisis de datos el que menos interés despierta, a pesar de la importancia que objetivamente tiene.

Es conocido el hecho de que "producir" datos no es una cuestión sencilla, tanto para el productor a gran escala (oficinas estadísticas) como para el que genera sus propios datos con fines específicos. Por ello, las técnicas de encuestas, muestreos y tratamiento de errores tienen un desarrollo amplio en la Estadística como ciencia.

La forma de obtener la información condiciona la estructura y calidad de los datos. Sin embargo, la transición que se produce entre el acotamiento de cualquier realidad, especialmente la social, económica o geográfica, y su representación mediante números, está sujeta a manipulaciones y ello conlleva inevitablemente errores. Para el proceso de investigación es imprescindible conocer su existencia, primero, y llevar a cabo una evaluación y tratamiento, después.

Medir implica asignar valores cuantitativos a los fenómenos según unas determinadas convenciones. En este proceso el error es consustancial con la actividad humana, lo que no significa que esta realidad se deba ignorar, sino todo lo contrario, detectar, corregir y explicar. Para ello se cuenta habitualmente con técnicas estadísticas y gráficas de diverso tipo.

Existen a grandes rasgos dos tipos de errores (fig. 2) que tienen también soluciones distintas según el estadio de la investigación en que se inserten.

ERRORES EN ANALISIS DE DATOS			
TIPOS	SUBTIPOS	ERRORES	TRATAMIENTO
MUESTRAL		Variabilidad muestral	Error muestral
		Tamaño de la muestra	Error standard de la muestra
		Representatividad	
		Procedimiento	
NO MUESTRAL	Científicos	Aplicaciones incorrectas de modelos	Revisión problema científico
	Fuente	Conceptos imposibles	Est.básica EDA Gráficos Transform. --- Contraste fuentes Rechazo
		No respuesta	
		Mala respuesta	
		Información falsa	
	Tipográficos	Inversiones de cifras	Est.básica EDA Gráficos Transform. --- Contraste fuentes Rechazo
		Repeticiones	
		Edición datos	
		Transmisión datos	
	Cálculo	Valores cero	Est.básica EDA Gráficos Transform.
		Valores negativos	
		Valores imposibles	
		Redondeo	
		Truncado	

Figura 2.- Errores en análisis de datos

Los primeros son los *errores muestrales*. Tienen que ver con el proceso estadístico de diseño de la muestra a partir de la que se obtienen los datos y pueden ser tratados por métodos cuantitativos. Si el productor es una oficina pública o ha realizado la muestra con pretensiones muy generales, los parámetros muestrales suelen aparecer explicados en la metodología, de manera que el uso de dichos datos está bien documentado en el sentido científico del término (tamaño, procedimiento de obtención y representatividad de la muestra). Cualquier empleo de datos fuera de esas condiciones puede afectar a la bondad de los mismos y a cualquier explicación (para una valoración de lo que pueden significar este tipo de errores y su significado gráfico; vease Schmid, 1983).

El interés de este trabajo no es profundizar en este campo sino reflexionar en las implicaciones que tienen los *errores no muestrales* para el análisis de datos.

Antes de entrar en los 3 tipos fundamentales de errores, es conveniente detenerse brevemente en el concepto de *error científico*, que no está directamente relacionado con la calidad material de la información. Se puede definir como la aplicación incorrecta de un/os dato/s a un problema científico (modelo) en la idea de que tal información se adecúa a sus presupuestos. Suele ser consecuencia, primero, de una trasposición mecánica de un modelo desde un ámbito geográfico o científico a otro sin que se haya valorado adecuadamente tal posibilidad y, segundo, de una falta de análisis previo de los datos. Esta situación puede producirse porque no se ha reflexionado suficientemente sobre el problema científico a explicar y su solución puede estar en una correcta valoración de las dos primeras fases del proceso de investigación.

Otro conjunto de errores proceden del diseño de la fuente estadística, sobre todo cuando se trata de registros que implican una declaración voluntaria de su contenido por parte de la persona que se registra. En buena medida los datos sociales, económicos, demográficos, etc. son recogidos de esta manera y ello no esta exento de dificultades que tienden a modificar el sentido de la información. Así, no es difícil encontrarse con datos encuadrados en capítulos "sin respuesta" en cuestiones como la nacionalidad o situación de residencia en registros poblacionales, o informaciones en "conceptos imposibles", como número de hijos entre mujeres de más de 65 años, o simplemente "informaciones falsas". En algunos ejemplos, sacados de estadísticas oficiales, se detectan situaciones sorprendentes por lo disparatadas, como la de Senegal que en 1979 tenía 10 habitantes por médico cuando lo normal en países africanos en ese momento eran 5.000 habitantes.

La solución a estos problemas, cuando se detectan a través de un análisis de datos, se orienta hacia el contraste con otras fuentes o simplemente a su rechazo suponiendo que el resto de la información es correcta.

Los *errores tipográficos* tienden a ser los más comunes como consecuencia de la frecuente manipulación de los datos. Además de los errores fácilmente detectables en el procesado humano de la información (copia o introducción

manual en el ordenador que dan lugar a inversiones o a repeticiones de cifras), existen otros derivados de la edición de las bases de datos o de la transferencia de ficheros entre programas u ordenadores, no siempre controlados en el 100% de los casos por la persona. Si estos errores no trastocan esencialmente la distribución de los datos suelen pasar desapercibidos.

Algo semejante sucede con los *errores de cálculo.* A pesar de que el ordenador es capaz de trabajar, en muchos casos, con grandes cifras, existe una tendencia en muchos programas a redondear o truncar las cifras con objeto de reducir el volúmen de cálculo y ello puede alterar, posiblemente en escasa cuantía, la distribución original de la información. Ahora bien, cuando se trata de operaciones entre variables, a veces se producen errores en algunos resultados (valores imposibles como consecuencia de dividir una cantidad por cero, valores negativos en variables en que este valor es irreal, etc.) o simplemente no introducir en el cálculo los casos con valores omitidos *(missing values).* Esta última limitación es bastante frecuente en los programas estadísticos standard.

En el caso de los errores tipográficos y de cálculo el análisis de datos, previo a cualquier otra manipulación, se manifiesta como una herramienta tremendamente eficaz para su detección y corrección, si son modificables, o para su explicación cuando la varibilidad sea un rasgo propio de la variable analizada. En este caso es donde más relevancia adquieren los valores extremos *(outliers).*

Pero algunas de esas funciones también pueden ser desarrolladas por medio de la representación gráfica, en algunos casos con mayor nitidez.

ANALISIS EXPLOTATORIO DE DATOS

El análisis de datos, en cualquiera de sus filosofías y modalidades, es el instrumento adecuado para su valoración con el objetivo de identificar "estructuras subyacentes" en las grandes masas de datos (Deville y Malinvaud, 1983; Bradshaw y Phillips, 1992), y de comprobar su calidad para detectar posibles errores (McPherson, 1989).

Quizás la discusión entre estadística clásica y análisis exploratorio de datos pueda parecer académica por una mera cuestión de hacer valer la bondad de un planteamiento, el clásico, opuesto a una "nueva" forma de analizar la información, el EDA. Sin embargo ello no es así. Este último tiene, en teoría, una vía directa con la intuición comprensiva de los fenómenos analizados, aunque esta cualidad también sea considerada propia de otras formas de hacer el tratamiento de la información como el Análisis Inicial de Datos (IDA) o Preliminar (PDA) (Chatfield, 1985; Chatfield y Schimek, 1987), o el análisis secundario.

Durante los últimos veinte años ha habido un enorme interés en los métodos de *análisis exploratorio de datos* (EDA). Este crecimiento debe mucho al trabajo pionero del estadístico americano John Tukey, que ha publicado muchos artículos y

al menos dos libros definitivos (Tukey, 1972; Mosteller y Tukey, 1977). El objetivo de Tukey ha sido crear nuevas técnicas de análisis que no estén afectadas por las mismas limitaciones que los métodos tradicionales de la Estadística "clásica", al mismo tiempo que sean fáciles de usar y de comprender. Tukey cree que sus métodos se adaptan particularmente a análisis exploratorios donde el investigador esté interesado en obtener una rápida visión de las rasgos principales de los datos.

Aunque los métodos de EDA se usan frecuentemente para estudios previos no deberían ser vistos como algo inferior, sino que producen resultados más precisos que las técnicas clásicas cuando se usan con muchos de los datos disponibles en Geografía Física y Humana. Estas técnicas tienden a estar disponibles en los principales paquetes estadísticos. Ya desde prinicipios de los años 80 existen repertorios (Francis, 1981; Woodward et al., 1988) en los aparecen programas estadísticos, que incluyen EDA, aunque todavía a un nivel de desarrollo reducido (Rodriguez, 1990). Sólo algunos grandes paquetes estadísticos (CSS, MINITAB, STATPAK, STATGRAPHICS, SYSTAT, SPSS/PC+, etc.) lo ofrecen como un conjunto de técnicas.

De acuerdo con McNeil (1977) el análisis de datos puede ser dividido en tres fases: análisis exploratorio, inferencia estadística y modelos. En la primera el investigador intenta conseguir una mejor comprensión de los datos a partir de diferentes puntos de vista y generar ideas acerca de los mismos. En el segundo paso el investigador intenta comprobar hipótesis por medio de tests, mientras que en la tercera se trata de crear modelos o teorías que expliquen las hipótesis.

Hasta ahora la Estadística ha puesto más atención en la segunda y tercera fase, quedando el EDA relegado. La razón de este énfasis en el análisis confirmatorio desarrollado en ciencias físicas y humanas se debe a que tienen abundantes teorías formales de las que deducir hipótesis, hecho que no es tan fácil en Ciencias Sociales. Más que hipótesis se trata de "intuiciones" sobre cómo se comportan los datos.

Las técnicas de Tukey permiten explorar los datos, a veces en forma tentativa, pero siempre de una forma creativa. Sin embargo, EDA no es sólo un conjunto de nuevas técnicas, sino más bien un punto de vista, un espíritu de búsqueda y creatividad. No es un "empirismo ciego", sino un uso inteligente de los datos, donde el investigador juega un papel principal. Se basan en el principio de que "el investigador sabe más que el ordenador". Combinando el conocimiento de la materia con las conjeturas inteligentes el investigador puede revelar más de los datos que los propios tests estadísticos.

Para Tukey, EDA es un análisis numérico propio de detectives. El investigador debería ser como Sherlock Holmes y tener un buen conocimiento profesional sobre los datos y usarlo para juzgar el valor y significación de la evidencia. Debe seguir cualquier pista, resultado inesperado, etc. para conocer porqué suceden los hechos. A menudo las pistas no llevan a ninguna parte, pero lo más importante es el espíritu de búsqueda. A diferencia de Sherlock Holmes, el investigador no busca contestaciones únicas, sino que intenta buscar respuestas razonables. En el análisis

cuantitativo puede haber diferentes interpretaciones de los datos y todas o muchas de ellas pueden ser válidas.

EDA no es una alternativa al análisis confirmatorio, sino que es complementario. Ayuda en el análisis de datos y auxilia en la formulación de hipótesis, mientras que el análisis confirmatorio comprueba las hipótesis y cómo pueden ser respaldadas por la evidencia empírica.

Entre las principales *características* destacan las siguientes:

a) **Simplicidad**: han sido diseñadas estas técnicas para ser tan simples como sea posible, de manera que el investigador ve los datos rapidamente y desde diversos puntos de vista. No se requieren matemáticas avanzadas y muchas de ellas trabajan sólo con un lapiz, un papel y una regla. Ultimamente se ha desarrollado técnicas que tienen propiedades interesantes pero que requieren muchos cálculos, en muchos casos solucionados con los programas de ordenador.

b) **Resultados gráficos**: estas técnicas producen buenos resultados gráficos que permiten al investigador "ver" las tendencias de los datos y en muchos casos son parte de un proceso de comprensión de los datos.

c) **Resistencia**: EDA ha sido creado para ser resistente o robusto, teniendo la propiedad de no estar influído por errores o valores extremos.

d) **Excepticismo**: el análisis de datos requiere ser excéptico con cualquier resultado que se haya obtenido, aunque siempre existe la posibilidad de que cualquier "resultado" pueda no encontrar aspectos informativos de los datos.

e) **Apertura**: EDA pone énfasis en el uso de tantas diferentes vías de análisis de datos como sea posible. El investigador debería permanecer abierto a los modelos no previstos en los datos, hecho que no sucede en el análisis confirmatorio (ej. en la comprobación de hipótesis sólo se consideran dos alternativas y no se preven explicaciones para ellas).

En general, Tukey cree que muchos investigadores todavía piensan y trabajan con estadística clásica y lo explica como un problema psicológico. Para introducir nuevas formas de pensamiento ha creado nuevos vocablos que ayuden a escapar de los viejos conceptos estadísticos. Sin embargo estos conceptos siguen manteniendose con lo cual la confusión entre viejas y nuevas palabras es evidente (ej. los conceptos *upper hinge* y *lower hinge* se refieren al cuartil superior e inferior; o el concepto de rango intercuartílico confundido con el *h-spread* de Tukey, llamado *mid-spread* por Erickson y Nosanchuck (Erickson y Nosanchuck, 1977).

Existe un amplio número de técnicas recogidas en los libros (Tukey, 1977; Mosteller y Tukey, 1977; McNeil, 1977; Erickson y Nosanchuck, 1977; Velleman y Hoaglin, 1981), imposibles de revisar en este momento, aunque sí lo haremos con una de ellas con objeto de profundizar en su filosofía.

Una de las técnicas más simples es el **diagrama de tallo y hojas**. En este ejemplo se usan los datos del porcentaje de población que vive en áreas urbanas para 70 de los países más grandes del mundo (Tabla 1). Cada dato es dividido entre el *tallo*, el valor de las decenas que está colocado a la izquierda de una linea vertical, y las *hojas*, las unidades que se colocan a la derecha de la linea vertical (fig. 3). Este gráfico representa la forma básica de la distribución: es fácil ver que la Figura 3 es bimodal con un pico en los valores 20 y otro en los 60.

China (CHI)	21	Polonia (POL)	58	Iraq (IRK)	72
India (IND)	22	Etiopía (ETI)	14	Holanda (HOL)	88
URSS (URS)	62	Zaire (ZAI)	30	Afganistán (AFG)	15
USA (USA)	74	Sudáfrica (SAF)	50	Ghana (GHA)	36
Indonesia (INO)	21	Argentina (ARG)	82	Uganda (UGA)	7
Brasil (BRA)	68	Colombia (COL)	60	Mozambique (MOZ)	9
Japón (JAP)	78	Canadá (CAN)	76	Chile (CHL)	81
Bangladesh (BAN)	10	Marruecos (MAR)	41	Hungría (HUN)	53
Pakistán (PAK)	29	Yugoslavia (YUG)	39	Arabia Saudí (SAU)	60
Nigeria (NIG)	20	Rumania (RUM)	49	Portugal (POR)	31
Méjico (MEJ)	67	Argelia (ALG)	52	Bélgica (BEL)	95
Alemania Fed (ALF)	85	Sudán (SUD)	25	Grecia (GRE)	65
Vietnam (VNM)	19	Tanzania (TAN)	7	Cuba (CUB)	65
Italia (ITA)	69	R.P.Corea (RPC)	33	Bulgaria (BUL)	62
Reino Unido (RUN)	77	Perú (PER)	67	Ecuador (ECU)	45
Francia (FRA)	78	Kenia (KEN)	13	Suecia (SUE)	83
Filipinas (FIL)	39	Venezuela (VEN)	76	Austria (AUT)	54
Tailandia (TAI)	17	Alemania D. (ALD)	76	Alto Volta (VOL)	9
Turquía (TUR)	45	Nepal (NEP)	5	Túnez (TUN)	52
Egipto (EGI)	44	Sri Lanka (SRI)	24	Suiza (SUI)	58
Irán (IRA)	50	Checoslovaquia(CHE)	67	Bolivia (BOL)	42
Corea (COR)	57	Australia (AUS)	86	Haití (HAI)	27
España (ESP)	64	Malasia (MAL)	30	Hong Kong (HON)	92
Birmania (BIR)	29				

* Nota: Los países incluídos aquí son los sesenta mayores países más grandes del mundo junto con diez de los restantes que tienen datos fiables.

Tabla 1.- Porcentaje de población que vive en áreas urbanas de 70 de los mayores países del mundo, según el tamaño de su población

A	B		C		D	
0	0		0	75799	0	57799
1	1		1	097435	1	034579
2	2	121	2	121909547	2	011245799
3	3		3	9093061	3	0013699
4	4		4	541952	4	124559
5	5		5	078023428	5	002234788
6	6	3	6	28794077552	6	02245577789
7	7	4	7	487866620	7	024666788
8	8		8	526813	8	123568
9	9		9	52	9	25

Figura 3.- Construcción de un gráfico *"Stem-and-Leaf"* (Fuente: Tabla 1)

Un elemento de la información que se ha perdido en la Figura 3 es la asociación de los niveles de urbanización con los países, información que, para los geógrafos, es de enorme interés. El diagrama de tallo y hojas puede ser adaptado facilmente para mostrar la relación anterior (fig. 4), identificando cada país con un código de tres letras. Los países de Africa y Asia tienen menores niveles de urbanización que los de America, Europa y Australia. En la Figura 5 las datos son divididos en los dos principales subgrupos y representados separadamente, lo que quizás es más significativo para comprender los modelos de urbanización del mundo. Es además más fácil identificar los verdaderos valores extremos (Ej. Hong Kong, Japón, Irak y Arabia Saudí aparecen como *outliers* dentro del grupo africano y asiático, mientras que en la Figura 4 no está tan claro).

0	TAN NEP UGA MOZ VOL	5
1	BAN VNM TAI ETI KEN AFG	6
2	CHI IND INO PAK NIG BIR SUD SRI HAI	9
3	FIL ZAI YUG RPC MAL GHA POR	7
4	TUR EGI MAR RUM ECU BOL	6
5	IRA COR POL SAF ALG HUN AUT TUN SUI	9
6	URS BRA MEJ ITA ESP COL PER CHE GRE CUB BUL	11
7	USA JAP RUN FRA CAN BEN ALD IRQ SAU	9
8	ALF ARG AUS HOL CHL SUE	6
9	BEL HON	2
(decenas)	(unidades)	70

Nota: cada país se ha identificado por un código de tres letras.

Figura 4.- Diagrama *"Stem-and-Leaf"* usado para identificar casos individuales (Fuente: Tabla 1)

	AFRICA Y ASIA	
0	NEP UGA TAN MOZ VOL	5
1	BAN KEN ETI AFG TAI VNM	6
2	NIG CHI INO IND SRI SUD PAK BIR	8
3	ZAI MAL RPC GHA FIL	5
4	MAR EGI TUR	3
5	IRA SAF ALG TUN COR	5
6		
7	SAU IRQ JAP	3
8		
9	HON	1
		36

	EUROPA, AMERICA Y AUSTRALIA	
0		
1		
2	HAI	1
3	POR YUG	2
4	BOL ECU RUM	3
5	HUN AUT POL SUI	4
6	COL URS BUL ESP GRE CUB MEJ PER CHE BRA ITA	11
7	USA CAN VEN ALD RUN FRA	6
8	CHL ARG SUE ALF AUS HOL	6
9	BEL	
		34

Figura 5.- *"Stem-and-Leaf"* dividido en dos grupos (Fuente: Fig. 4)

La adaptabilidad del diagrama de tallo y hojas aparece tambien en la Figura 6 donde se representan enfrentados dos conjuntos de datos: muestra claramente una fuerte correlación positiva entre los niveles de alfabetismo y el empleo en la agricultura de los Estados mejicanos. También identifica aquellos estados que se adaptan a la regla (por ej. Chiapas, Guerrero, Oajaca, Nuevo León, D.F.) y, quizás de forma más interesante, los que no (Zacatecas). El original de esta figura ha sido realizado a mano y en no más de 5 minutos.

El diagrama básico de tallo y hojas tiene importantes ventajas en el análisis exploratorio de datos. Primero, es muy fácil de construir, incluso a mano, y los resultados se ven rápidamente. Segundo, no sólo no se pierde información inicial sino que puede ser adaptada al interés del investigador. En tercer lugar es una técnica flexible y puede ser adaptada para su uso con muchos tipos de datos. Algunos otros rasgos pueden ser vistos en Tukey, 1977; Erickson y Nosanchuck, 1977; y Velleman y Hoaglin, 1981. La mayor parte de las otras técnicas de EDA tienen semejantes ventajas y deberían ser usadas cuando el investigador se enfrente con nuevos conjuntos de datos o problemas.

% POBLACION ACTIVA EN LA AGRICULTURA	100	50	% ANALFABETOS
	90-99	45-49	
	80-89	40-44	CHIA, GUER, OAXA
OAXA, CHIA	70-79	35-39	GUAN, HIDA, QUER
ZACA, HIDA, GUER	60-69	30-34	MICH, PUEB
YUCA, VCRU, TLAX, TABA, SINA, SLPO, QROO, PUEB, NAYA, MICH, DURA	50-59	25-29	MORE, SLPO, VCRU, YUCA
QUER, MORE, GUAN, COLI, CAMP	40-49	20-24	CAMP, MEXI, NAYA, QROO, SINA, TABA, TLAX
TAMA, SONO, MEXI, JALI, CHIH, CALS, AGUA	30-39	15-19	COLI, JALI, ZACA
COAH, BCAL	20-29	10-14	AGUA, BCAL, CALS, COAH, CHIH, DURA, LEON, SONO, TAMA
LEON	10-19	5-9	DFED
DFED	0-9	0-4	
	0	0	

Figura 6.

ERRORES Y "OUTLIERS" (Y QUE HACER CON ELLOS)

Hemos comprobado que los errores son elementos inevitables en los datos. Incluso cuando los errores se han detectado y se han corregido mediante, por ejemplo, técnicas de EDA, todavía existe el peligro de errores no identificados. En cualquier circunstancia la postura más inteligente es asumir que el error existe, aunque haya que reconocer que hay algunos errores más importantes que otros. Consideremos los dos ejemplos siguientes (figs. 7a. y 7b.).

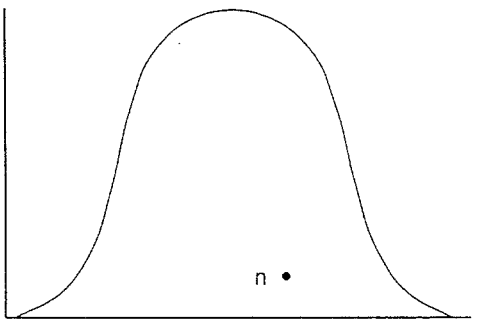

Figura 7a.

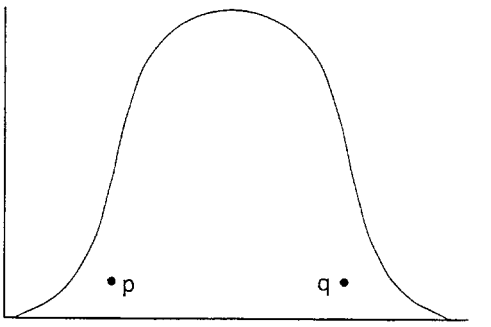

Figura 7b.

Figuras 7a y 7b.- Integración de E.D.A. en un programa de análisis estadístico. (Fuente: Velleman y Hoaglin, 1981)

Introducción
Descripción de distribuciones
Gráficos de tallo y hojas
Medidas de tendencia central
Medidas de variabilidad
Gráficos de valores alfabéticos
Reexpresión de datos
Valores extremos

Relación lineal
Linea resistente
Mínimos cuadrados
Reexpresión de datos
Análisis de residuales

Probabilidad
Inferencia para muestras grandes
Estimación de intervalos sobre la media
Comprobación de hipótesis
Estimación de diferencia de medias

Inferencia para muestras pequeñas
t Student

Inferencia para regresión lineal
Tests t para coeficientes de regresión
Correlación
Comparación de lineas resistentes

Análisis de tablas de datos
Tablas codificadas
Estadístico Xi cuadrado

Métodos aditivos para tablas de datos
Comparación de más de dos medias
Gráficos de caja múltiples
Análisis de varianza univariado
Ajuste de medianas
Análisis de varianza bivariado

Series temporales
Suavizado de datos no lineales
Modelos para series temporales.

En la Figura 7a. se supone que hay puntos que deberían estar en n, mientras que de hecho lo están en m, formando un valor extremo. En esta extrema localización el error tendrá una influencia importante en el cálculo de cualquier estadística tales como la media o la desviación típica. Sin embargo, en la Figura 7b. el error está dentro del cuerpo de los datos, ocurre en p cuando debería ocurrir en q. Aunque el error es de la misma magnitud que el de la Figura 7a. (la distancia p-q es la misma que la distancia m-n) sin embargo es probable que tenga una influencia muy limitada en el cálculo de los valores estadísticos. Por esta razón el problema del tratamiento de los datos se ha concentrado en la cuestión de los valores extremos.

La cuestión de determinar el método correcto de tratamiento de *outliers* tiene una larga historia en Estadística. Uno de los primeros científicos en enfocar este problema fue el alemán Bessel quien, en 1838, afirmó que el nunca rechazó ningún dato (incluyendo valores extremos) por razones estadísticas. La única razón para rechazar un dato es que dicho dato sea un error detectado en la recogida o en la medida. Incluso Bessel fué más lejos al afirmar que todos los datos tenían el mismo peso "para quitar la arbitrariedad en nuestros resultados" (Huber, 1972). Este representa el punto de vista más puro, permaneciendo todavía dominante entre la mayoría de los estadísticos y los analistas de datos hasta tiempos muy recientes.

Sin embargo, al final de los años 50 y principio de los 60 este punto de vista ha ido cambiando y se han hecho esfuerzos por encontrar métodos efectivos para el tratamiento de los valores extremos, con tres posibles soluciones (Kruskal et al., 1960):

- *rechazo* de valores extremos, que significa su salida completa del conjunto de datos. Fué el método tradicional de tratar el problema de los *outliers*. Los siguientes métodos son menos comunes,

- *truncado*, donde un cierto porcentaje de los valores más altos y más bajos son automáticamente eliminados de los extremos de acuerdo a alguna fórmula,

- *winsorization* donde un valor extremo conserva su signo, pero su magnitud es alterada hasta situarla entre los dos valores vecinos (anterior y posterior).

Debería ser enfatizado que el rechazo de valores extremos ha sido considerado, desde el punto de vista estadístico, como un procedimiento peligroso y sólo tendría que aplicarse bajo condiciones estrictas, por ejemplo, no con muestras de tamaño pequeño (Anscombe y Tukey, 1963). Se debe recordar que los tres procedimientos se aplican sólo cuando se calculan medidas de tendencia central y de dispersión.

En otras áreas del campo científico y estadístico no sólo se han retenido los valores extremos sino que han sido considerados como los datos más importantes, puesto que arrojan una luz muy considerable sobre el comportamiento de las observaciones.

Eventualmente se ha reconocido que un procedimiento que suponga un rechazo de valores extremos, seguido de la aplicación de técnicas estadísticas clásicas, está sujeto a importantes limitaciones (Huber, 1981):

1. Es difícil separar los dos pasos ya que las reglas para el rechazo casi siempre necesitan estimaciones fiables (robustas) de los parámetros básicos.

2. Los procedimientos de rechazo no serán perfectos. Los datos pueden contener errores de falsos rechazos y falsas retenciones. La estadística clásica normal no es aplicable a datos que han sido limpiados.

3. Los mejores procedimientos de rechazo no funcionarán tan bien como los robustos.

Por estas y otras razones se ha puesto un considerable interés en el estudio de la estadística robusta en los últimos 20 años.

ESTADISTICA ROBUSTA

Aunque el estudio de las estadísticas robustas es relativamente reciente, el concepto de *robustez* no lo es. A finales del siglo XIX y principios del XX estadísticos y científicos fueron conscientes de la robustez, de manera que fueron inventadas algunas técnicas robustas que posteriormente han sido redescubiertas. Una razón para el abandono de algunos de estas técnicas fue la disputa entre Eddington y Fisher al principio del siglo XX en relación con la mejor medida de la dispersión. Eddington abogó por el uso por la *desviación media absoluta* :

$$\delta n = \frac{1}{n} \Sigma \mid x_i - x \mid$$

mientras Fisher lo hizo por la *desviación típica:*

$$çn = \left[\frac{1}{n} \Sigma (x_i - x)^2\right]^{1/2}$$

La disputa fué ganada por Fisher quien apuntó que para observaciones normales $çn$ es aproximadamente un 12% más eficiente que δn. Sin embargo, aproximadamente cincuenta años más tarde, Huber (1977), usando medidas de eficiencia asintótica relativa (ARE), indicó que sólo dos malas observaciones en 1000 son suficientes para compensar el 12 % de ventaja de la desviación típica (Tabla 2).

Es bastante claro en la Tabla 2 que entre 2 y 500 errores sobre 1000 observaciones la desviación absoluta media es una medida de desviación mejor que la desviación

típica (existen, sin embargo, otras medidas de desviación mejores que esta última). Como muchos datos geográficos es probable que contengan entre el 0.2% y el 50% de errores, es conveniente tomar en consideración el uso de medidas alternativas a la desviación típica.

ERROR	ARE
0.000	0.876
0.001	0.948
0.002	1.016
0.005	1.198
0.010	1.439
0.050	2.035
0.100	1.903
0.500	1.017
1.000	0.876

Tabla 2.

Un último punto es que las técnicas estadísticas, y en particular las que se basan en la inferencia estadística, se fundamentan en parte en observaciones empíricas y en parte en asunciones a priori, sobre todo la forma de la distribución subyacente. Tales suposiciones son vitales si la estadística se aplica a problemas prácticos. Previamente se ha asumido que las técnicas estadísticas tradicionales se adecúan al *principio de estabilidad,* por lo que un pequeño error en el modelo estadístico debería producir sólo un pequeño error en los cálculos finales. Desgraciadamente este principio no siempre se aplica, por lo que se sabe que muchas de las técnicas usadas comúnmente son muy vulnerables a las pequeñas variaciones en los supuestos iniciales, especialmente a la distribución normal.

El término "robusto" fué usado por primera vez por Box en 1953 y es definido usualmente como "la insensibilidad a las pequeñas desviaciones de los supuestos subyacentes". La principal preocupación es con respecto a las desviaciones del modelo asumido y especialmente a la distribución normal o gaussiana. Se conocen poco los efectos con respecto a otros tipos de desviaciones. Algunos autores hacen distinción entre "robustez", las menores desviaciones con respecto a la distribución subyacente, y "resistencia", la propiedad de permanecer relativamente no afectados por valores extremos *(wild values o outliers).* Sin embargo, aunque las dos ideas son conceptualmente distintas, a efectos prácticos pueden ser consideradas como sinónimos (figs. 8a. y 8b.).

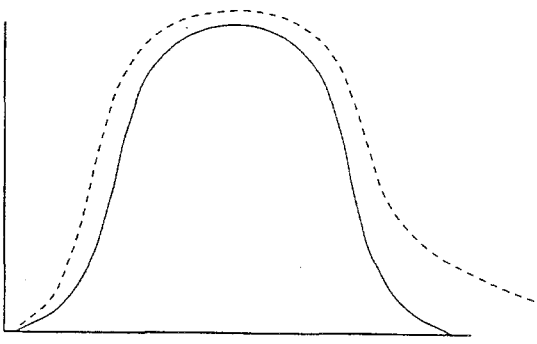

Figura 8a.

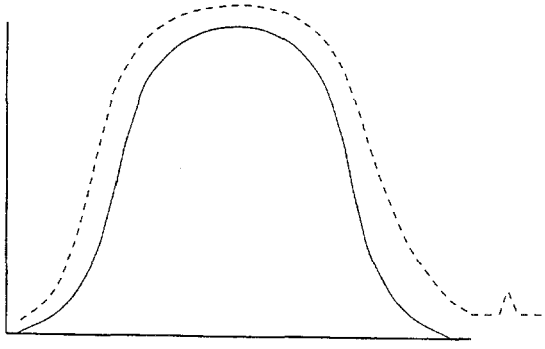

Figura 8b.

Las técnicas robustas pueden ser aplicadas a cualquier tipo de problema (Huber, 1981) y no deberían ser confundidas con las *no paramétricas* donde no se conoce nada sobre las distribuciones subyacentes. Las técnicas no paramétricas pueden ser muy sensibles a los valores extremos y no son robustas. De forma similar no se deben confundir con los tests de distribución libre *(distribution-free tests)* en los que la probabilidad de rechazar falsamente la hipótesis nula es la misma para todas las distribuciones subyacentes. Muchos de estos tests son robustos, pero ello parece ser más una feliz coincidencia que una propiedad inherente a esas técnicas.

Comparacion de medidas clásicas y robustas

Como se ha señalado la media y la desviación típica no son medidas robustas. Al contrario, la mediana y el rango intercuartílico *(midspread)* sí lo son. Consideremos el siguiente ejemplo: en el conjunto de datos (1, 2, 3, 4, 5, 6) la media es 3.5, la desviación típica 1.87, la mediana 3.5 y el rango intercuartílico 3 (5-2). Sin embargo, si los datos se modifican sustituyendo el 6 por 60, la media se convierte en 12.5 y la desviación típica en 23.31. Entonces, un cambio en un sólo valor produce una media y una desviación típica que han multiplicado su valor, por lo que no dan una idea del comportamiento general de los datos. En contraste, la mediana y el rango intercuartílico de los datos modificados permanecen sin cambiar en 3.5 y 3 y proporcionan unas buenas medidas de cinco de los 6 valores.

La media a veces proporciona una guía engañosa de los datos porque, como Erickson y Nosanchuk (1977) señalan, los investigadores a menudo preguntan por cuestiones erroneas en los datos. Por ejemplo, consideremos un país en el que en un año dado 999 personas ganan 1 peseta cada uno y otra persona obtiene 1 millón de pesetas. La renta media es aproximadamente 1.000 pts., que no está cercana a ninguna de las rentas de las personas. El problema es que se pregunta "¿cuál es la renta *media* en este país?", que es una cuestión no adecuada si se utiliza la *media*,

en vez de "¿cuál es la renta de una persona *típica*?" que es mucho más acertada porque se trata de la *mediana*.

La no resistencia a los valores extremos es incluso mucho más pronunciada con la desviación típica, porque se basa en la varianza en la que las desviaciones de la media son elevadas al cuadrado.

Dada, entonces, la relativamente mala representatividad de la media y la desviación típica es interesante considerar porqué son todavía usadas con regularidad en estudios científicos. La casi reverencial estima de las medidas clásicas se basa, en parte, en la creencia de que son las medidas "naturales" de la distribución a causa de su relación con la distribución normal. El mismo Gauss no comenzó usando la distribución normal para representar la situación normal de los datos, sino que decidió el uso de la media porque era fácil de calcular y a través del teorema de Gauss-Markov llegó a la distribución normal. Por lo tanto no existe nada sagrado en el uso de la media en análisis de datos.

Otras razones se basan en la tradición y porque todos los datos intervienen en su cálculo. Como señala Huber (1972) otras medidas habían tenido también un uso tradicional anteriormente y pone el ejemplo de que en Francia en el siglo XIX había tradición de valorar los terrenos agrícolas de acuerdo a la producción de cereales. Se tomaban valores consecutivamente durante 20 años y se quitaban los valores superior e inferior. Los 18 valores restantes servían para el cálculo de la media. Esta filosofía de medida se mantiene actualmente en algunos deportes (patinaje sobre hielo, gimnasia) y ello es así para evitar el sesgo de algún juez a la hora de otorgar las puntuaciones.

Algunas medidas robustas de tendencia central

Existe un número elevado de alternativas a la media aritmética como medida de tendencia central. Por ejemplo, en una encuesta sobre medidas robustas un grupo de investigadores de la Universidad de Princeton investigó las propiedades de 60 medidas (Andrews et al., 1972) e incluso su lista no fué exhaustiva. Tomaremos tres ejemplos con objeto de ilustrar esta lista:

a) *Trimmed means* (medias recortadas).

Parten de la idea de que la media aritmética es simple y bien conocida, pero está afectada por la existencia de valores extremos. Como se ha visto en el ejemplo de la agricultura francesa, una solución sencilla a este problema es "recortar" los valores extremos y calcular la media de los restantes. La proporcion de valores recortados se denomina α y entonces es posible tener una familia de diferentes medias α Andrews (1972) sugiere que el factor α debería ser una proporción fija del conjunto de datos (1%, 5%, etc.), dependiendo de la naturaleza de los datos que están siendo analizados. El cálculo de la media recortada se muestra en el ejemplo 1.

Ejemplo 1.

Trimmed mean (Media recortada)

Fórmula: la media aritmética de los datos restantes despues de que se hayan recortado n valores de cada uno de los extremos del conjunto de datos ordenado.

Datos: 6, 66, 67, 68, 68, 69, 70, 71, 71, 72, 74, 78.
n = 12
media aritmética = 65
mediana = 69.5
Valor de α tomado (10%) = αn (10*12)/100 = 1.2

Se redondea hasta el valor entero más cercano, es decir, 1; se quita ese valor de los dos extremos de los datos y se opera con los diez valores restantes. La media recortada es:

Comentario: En este ejemplo la media aritmética (65) no es una buena medida de la tendencia central porque su valor es más bajo que nueve de los diez números del conjunto de datos. La media recortada (69.6) es una medida más ajustada puesto que está en el centro del principal grupo de valores y es muy parecida a la mediana.

b) *Folded medians* (medianas plegadas).

Esta medida fué propuesta por Tukey a partir de una sugerencia de Hodge y Lehmann. El conjunto de n datos es ordenado y se calcula el valor medio de los pares de datos colocados simétricamente. Entonces se obtiene la media de los valores más grande y más pequeño, de los siguientes más grande y más pequeño y así sucesivamente. En el caso de un número impar de datos la media se calcula consigo mismo. En una versión iterativa de esta medida, el proceso se repite un número de veces, tomando entonces la mediana de los valores medios obtenidos (En el ejemplo 2 se muestra el cálculo).

Ejemplo 2.

Folded median (Mediana plegada)

Fórmula: En este ejemplo se "pliegan" los datos dos veces y se toma la mediana. El cálculo se realiza obteniendo la media de los datos simétricamente colocados en el conjunto de datos ordenado.

Datos: 6, 66, 67, 68, 68, 69, 70, 71, 71, 72, 74, 78

Cálculo: $\frac{6+78}{2} = 42$: $\frac{66+74}{2} = 70$; $\frac{67+72}{2} = 69.5$

$$\frac{68+71}{2} = 69.5: \quad \frac{68+71}{2} = 69.5; \quad \frac{69+70}{2} = 69.5$$

Nuevo conjunto de datos: 42, 69.5, 69.5, 69.5, 69.5, 70

$$\frac{42+70}{2} = 56: \quad \frac{69.5+69.5}{2} = 70; \quad \frac{69.5+69.5}{2} = 69.5$$

Datos resultantes: 56, 69.5, 69.5

Mediana de estos datos: 69.5

Comentario: La principal ventaja de este tipo de medidas es que se calculan con todos los datos y por ello tienen las mismas oportunidades de influir en el valor de la media, hasta que se obtiene la mediana al final de los cálculos. En este caso el valor hallado es de nuevo 69.5.

Ejemplo 3

Estimación *biweight*

Fórmula:

$$x = \frac{\sum \omega_i(u_i)\, x_i}{\sum \omega_i(u_i)}$$

donde

$$\omega_i(\upsilon_i) = (1 - u_i^2)^2, \text{ si } | \upsilon_i | \leq 1$$
$$\text{en otro caso}$$

y

$$\upsilon_i = \frac{x_i - x}{c^8}$$

y $c = 6$

$s = 1/2$ (rango intercuartílico)

Datos: 6, 66, 67, 68, 68, 69, 70, 71, 71, 72, 74, 78

media aritmética = 65

rango intercuartílico = 4

cs = 12

biweight = 69.85

Comentario: La estimación *biweight* es extremdamente lenta de calcular, por lo tanto no es una buena medida si no se dispone de un ordenador. Sin embargo, tiene un buen número de propiedades que pueden ser interesantes para análisis más avanzados. En este ejemplo, la estimación varía desde 65 (la media) en la primera iteración a 68.8, 69.49, 69.765, 69.829, 69.85 y 69.85 en sucesivas etapas. El resultado final no difiere de otros estadísticos robustos.

c) *Biweight* (Ponderación "bicuadrada").

Esta medida es más complicada y se calcula por la ponderación de cada dato por un peso "bicuadrado" ($\omega_i(u_i)$) según la fórmula:

$$\omega_i(u_i) = (1 - u_i^2)^2, \text{ si } |u_i| \leq 1$$

$$\text{en otro caso}$$

donde

$$u_i = \frac{x_i - x}{c^8}$$

y c es una constante que se sugiere tenga un valor de 6 (aunque también se ha propuesto un valor de 9) y s = 1/2 (rango intercuartílico). Para una distribución normal, s es aproximadamente igual a 2/3 de la desviación típica. Finalmente la estimación se define como:

$$x = \frac{\sum \omega_i(u_i)\, x_i}{\sum \omega_i(u_i)}$$

Desgracidamente el cálculo de x depende de los pesos $\omega_i(u_i)$ y viceversa, de manera que la solución requiere un proceso iterativo para determinar esos valores. Como los cálculos son extremadamente laboriosos, sólo es factible hacerlo cuando se disponga de un programa de ordenador.

Cuando se dispone de muchas medidas cabe preguntarse cuál es la mejor. El estudio de Princeton intentó contestar a esta pregunta de acuerdo a diversos criterios (Andrews et al., 1972). Los autores fueron remisos a definir una única buena medida puesto que la variedad de situaciones en el análisis de datos requiere también una variedad de instrumentos. Sin embargo afirmaron que la media era una de las peores en muchos casos, mientras que *"trimmed mean"* funciona bien, la mediana proporciona el método más rápido y sencillo y que *biweight* es la medida de mayor calidad (Mosteller y Tukey, 1977).

Dada la gran variedad de los métodos robustos actualmente disponibles y la complejidad en su cálculo, no siempre es fácil conocer lo que cada una de ellas hace. Sin embargo, este problema está siendo resuelto gradualmente con la presencia en los programas informáticos de algunas de estas medidas. En estas circunstancias el consejo más adecuado es obtener tanto medidas clásicas como robustas sobre los datos. Si las dos medidas producen los mismos o semejantes resultados entonces es preferible usar medidas clásicas porque son mejor conocidas y pueden ser relacionadas con otros estudios. Si no es así, entonces es que los datos tienen rasgos específicos que necesitan un estudio más detallado. La decisión final dependerá de los resultados del estudio de comparación (Bickel, 1976).

MAS VALE UN GRAFICO...

Como sucede con la Estadística (básica, EDA, inferencial, multivariante,...), el ámbito de la representación gráfica es un mundo sometido a unas reglas que tratan de diferenciar campos de aplicación, entre los que se encuentra el de las Ciencias Sociales como uno de los más nítidos.

Posiblemente "no exista un instrumento estadístico sencillo tan efectivo que un gráfico bien seleccionado" (Bradshaw y Phillips, 1992). Su conexión con las técnicas estadísticas hasta ahora analizadas esta fuera de toda discusión y de ahí el nombre más comúnmente aceptado, "gráficos estadísticos" (Schmid, 1983) en clara alusión a esta doble función.

En los gráficos estadísticos, como medio para establecer una comunicación entre el productor y el usuario, intervienen componentes psicológicos y perceptuales. Serían las dos caras de una moneda, la "graficacia" y la "percepción gráfica" (Schmid, 1983; Spence y Lewandowski, 1990; Cleveland, 1987; Cleveland y McGill, 1987). En un sentido más literario, Tufte (1983) se ha referido a la "excelencia gráfica".

Para documentar la parte más psicológica del proceso de comunicación existen diversas teorías que pretenden sentar las bases de este proceso (Spence y Lewandowski, 1990), en muchos casos corroboradas con experimentos que miden la accesibilidad a los gráficos de dístintos tipos de usuarios (Cleveland y McGill, 1985; Cleveland y McGill, 1987).

Desde otra faceta, los gráficos son el correlato lógico de muchas técnicas estadísticas, como ya se ha señalado, aprovechando casi siempre las potencialidades de los programas de ordenador. Quizás no sea la Estadística quien más ideas haya aportado al diseño de los gráficos. Al contrario, Schmid (1983) ha destacado el papel jugado en este campo por otras disciplinas como la Cartografía, la Informática aplicada, la Psicología, el Diseño gráfico, etc. Sin embargo, ello no es una garantía para una buena representación visual de los datos. Más importantes son, sin duda, los datos a representar.

En esencia, el gráfico puede ser asimilado a una fotografía que necesita ser codificada de acuerdo con los parámetros y los objetivos de la persona que desea aprehender una realidad, para que luego sea descodificada según los elementos perceptuales del usuario de la fotografía. La bondad de un gráfico depende de que la codificación-descodificación se ajuste a las leyes de la comunicación visual (fig. 9).

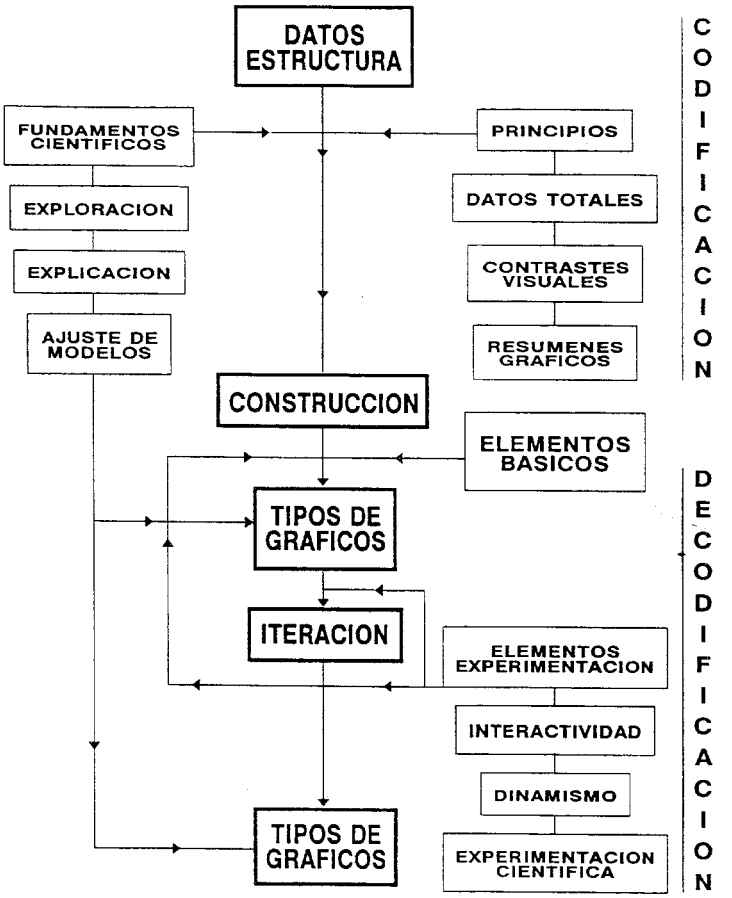

Figura 9.- Proceso de elaboración de gráficos

Por ello, antes de proceder a una representación ajustada de la información es absolutamente necesario identificar cuál es el objetivo del gráfico, porque de ello se deriva su diseño y contenido, especialmente desde que los ordenadores reducen al mínimo el tiempo material de ejecución. A modo de ejemplo, Tukey (1986) ha propuesto una serie de preguntas a hacerse antes de construir un simple diagrama de puntos.

Dos son los objetivos principales: uno es la presentación de datos, el otro el análisis. En el primer caso se trata de gráficos más sencillos, lo que no significa que estén bién concebidos (Spence y Lewandowski, 1990). Para el análisis de datos la comunicabilidad del gráfico se adapta a tres fundamentos científicos (exploración del volúmen de datos para descubrir conjuntos comunes y específicos en su distribución; la explicación de lo que es en origen poco abarcable o la aplicación a la realidad medida de modelos teóricos definidos previamente).

A partir de esta base se aplican unos principios que se basan en la claridad, la precisión y la eficacia, para tratar de presentar unos datos totales a través de los cuales se puedan extraer resúmenes y contrastes visuales de dicha información. Tufte (1983) ha señalado los siguientes:

- mostar los datos
- inducir al espectador a ver el contenido y no tanto la forma de construir el gráfico
- evitar distorsiones
- presentar el mayor número de datos en el menor espacio
- hacer que haya coherencia en los conjuntos de datos
- hacer posible su comparación visual
- resaltar los detalles, a distintos niveles, entre los datos
- estar integrado con el análisis estadístico y literal.

Es posible que llevando a cabo estas ideas se construya un "buen" gráfico. Sólo lo será si además se adapta a la función para la que se ha diseñado, a los datos, a los usuarios que lo van a leer, a los elementos físicos que lo componen (Schmid, 1983).

Los experimentos llevados a cabo y un análisis exhaustivo de los tipos de gráficos más comunes (Cleveland y McGill, 1987; Schmid, 1983; Tufte, 1983) demuestran que el incumplimiento de algunos de estos principios y condiciones puede limitar la utilidad del gráfico.

De acuerdo con lo anterior, la construcción de un gráfico es uno de los puntos clave en este proceso de comunicabilidad, cuya investigación todavía no se ha cerrado. Las posibilidades que ofrecen los programas de ordenador (estadísticos con aplicaciones gráficas o gráficos con pequeños cálculos estadísticos) permiten suponer que este es un ámbito en evolución. Sería adecuado quizás sugerir la necesidad de estudiar más profundamente las ventajas e inconvenientes de los gráficos y su relación con la informática para su aplicación en Geografía.

Un gráfico producirá un resultado adecuado si "nuestro sistema visual interpreta la percepción gráfica con exactitud y eficiencia" (Cleveland, 1987), de

acuerdo con el principio de parsimonia, que afirma que los elementos empleados en la construcción serán tan sencillos como sea posible, a la vez que su representatividad sea la más alta posible también.

Por ello es indispensable identificar una escala de elementos generalizable según su representatividad. Esta es una tarea difícil porque no vale la propia intuición del investigador si no se acompaña de una experimentación.

En uno de sus primeros estudios Cleveland y McGill (1984a) definieron una escala de 10 elementos gráficos perceptuales (fig. 10a), ordenados desde los más a los menos exactos en su percepción. Se basaron en teorías psicofísicas y trabajos experimentales. Valorando la percepción de distintos tipos de usuarios sobre gráficos diferentes según sus elementos, descubrieron que los errores tienden a ser menores cuando los elementos se sitúan en la parte alta de la escala ordenada (fig. 10b). Resumen los autores esta cuestión de la siguiente manera: "la posición y la longitud son los más exactos, mientras que el ángulo y la pendiente son semejantes entre sí, pero menos que la longitud" (Cleveland y McGill, 1984a), mientras que el área es, de todos, el menos exacto.

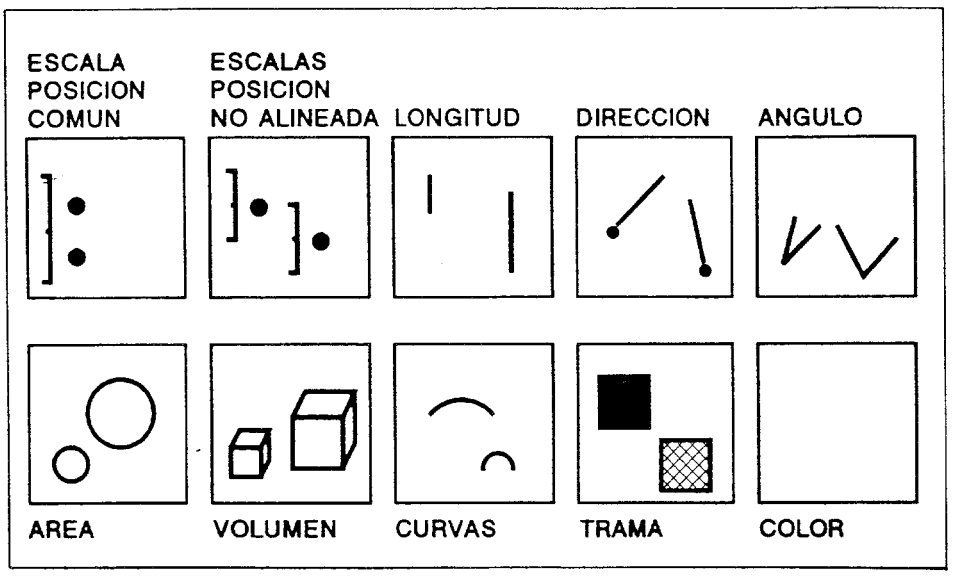

Figura 10a.- Elementos básicos en la construcción de un gráfico (Fuente: Cleveland and McGill, 1984a)

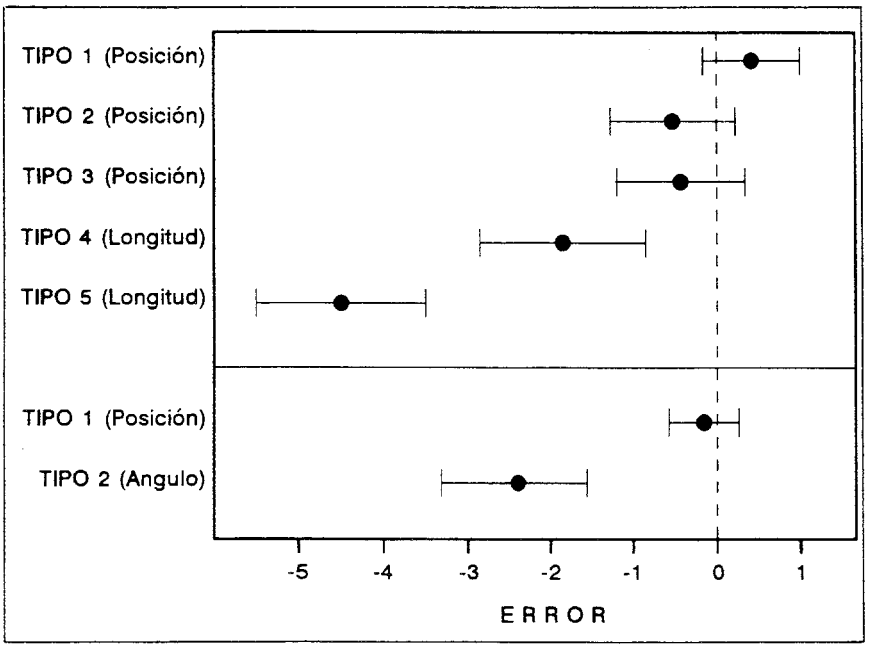

Figura 10b.- Errores medios e intervalos de confianza al 95 % para juicios sobre experimentos posición Vs. longitud y posición Vs. ángulo (Fuente: Cleveland and McGill, 1984a)

Unas breves recomendaciones ayudan a sintetizar el esfuerzo perceptual derivado de los experimentos:

- un gráfico de tarta siempre puede ser reemplazado por otro de barras (posición/longitud vs. ángulo)
- un gráfico de puntos es preferido o otro de barras (posición vs. longitud)
- un gráfico de barras divididas siempre puede ser sustituído por otro de barras agrupadas
- un gráfico de puntos agrupados siempre es preferido a otro de puntos agrupados.

En la descodificación de los gráficos, es decir el acercamiento del usuario a la información contenida en su interior, es conveniente referirse a un mecanismo interesante para el análisis de datos. Se trata de la *iteracción* que debe producirse cuando el gráfico no cumpla lo anteriormente especificado ni actúe como un vehículo de comunicación. Puede suceder, en efecto, que ni la estructura latente de los datos ni los valores extremos hayan sido destacados por lo que la apariencia y la

claridad del gráfico se ven mermadas. También es posible que los elementos formales no hayan sido bien estructurados o que el exceso de elementos reste claridad a su contenido.

Tufte (1983) ha catalogado alguno de estos efectos como *data-ink ratio* (es la relación datos/tinta, que mide la cantidad de información no redundante del gráfico), la "basura gráfica" (*chart-junk*, referida a la decoración interior innecesaria que enmascara los datos), o el efecto del "gran pato" (*big duck*, para identificar la situación que se produce cuando el gráfico se adapta en forma pictorial a una forma relacionada con el objetivo del gráfico, pero perdiendo claridad).

En todos estos casos se trata de advertencias contra la parafernalia de que se sirven algunos programas de ordenador a la hora de producir gráficos. De nuevo aparace la filosofía del análisis de datos que insiste en el papel del hombre en el proceso de investigación, especialmente en las primeras fases, antes de dejarse llevar por los oropeles informáticos.

¿Cuáles son, entonces, los elementos que deben favorecer una iteracción? La experimentación científica de los datos en el gráfico proporciona la base necesaria para integrar la representación en los objetivos científicos. En este sentido, la conjunción del análisis de datos y de gráficos juega un papel decisivo en esta fase: la literalidad de la explicación no debe por ello verse afectada por los gráficos y los datos.

La interactividad que ofrecen los ordenadores actualmente hace más fácil y rápido el proceso iterativo, apareciendo, como consecuencia, un tipo, los gráficos dinámicos, que representan un salto cualitativo. En los programas de ordenador las "utilidades" permiten insertar etiquetas en los datos, borrar datos erróneos, unir elementos dentro del gráfico, activar la información contenida en recuadros, reescalar coordenadas, rotaciones, resaltes... (Becker, Cleveland y Wilks, 1987). Sin embargo, han surgido dos problemas en relación con los gráficos dinámicos, su representación "estática" en papel y la dificultad que representa su percepción.

Tipos de gráficos

En el análisis exploratorio los gráficos adquieren plenamente la función de detectar regularidades y excepciones en los datos. Sin querer significar que esta forma de análisis sea una nueva tendencia plenamente diferenciada, los gráficos estan concebidos para integrarse plenamente en la explicación de los datos. Se pretende representar la totalidad de los datos con objeto de mantener su individualidad de forma que sea posible explorar el conjunto completo y valorar qué tipo de análisis posterior es el más adecuado. Es cierto que no siempre esto es así porque en algunos casos es preferible obtener resúmenes gráficos. El ejemplo más claro es el del suavizado de gráficos para hallar la tendencia del conjunto de datos y separarla de los valores extremos. En definitiva, los contrastes visuales permiten la discriminación de los datos totales representados.

También hay otros gráficos especificos que no se relacionan directamente con el EDA, pero que han sido diseñados para la visualización de los datos. En la Figura 11 se han agrupado los más comunes de estos gráficos, comparándolos con los tradicionales y destacando las funciones que cumplen.

TIPOS DE GRAFICOS					
TIPOS	ESTADISTICA CLASICA	ANALISIS DE DATOS	DT	CV	RG
UNIVA-RIADOS	Histograma	Gráfico de caja		*	*
	Polígono de frecuencias	"Rootograma" normal		*	
		"Rootograma" colgado		*	
BIVA-RIADOS	Diagrama de puntos	Diagrama de puntos	*	*	*
	Diagrama de puntos	Diagrama de puntos por grupos	*	*	*
		Diagrama de puntos suavizado	*	*	*
TRIVA-RIADOS	Diagrama de puntos rotado	Diagrama de puntos (3 dimensión)		*	
		Girasoles		*	
		Matriz de diagramas de puntos	*	*	
MULTI-VARIADOS		Perfiles	*	*	
		Estrellas	*	*	
		Rayos de sol	*	*	
		Draftsmans	*	*	
		Glifos	*	*	
		Curvas de Andrews	*	*	
		Caras de Chernoff	*	*	
		Arboles	*	*	
		Castillos	*	*	

DT: Representación total de los datos
CV: Contrastes visuales
RG: Resumenes gráficos

Figura 11.- Tipos de gráficos diseñados para la visualización de los datos

Entre los gráficos univariados, el EDA ha aportado, como ya se ha indicado, formas diferentes de establecer contrastes visuales entre algunos valores individuales (casos extremos) y sus medidas de referencia, que superan al histograma. Algo semejante sucede con el "rootograma" normal y colgado (diagrama de raiz cuadrada) en relación con la curva normal.

Pero es entre los gráficos bivariados y trivariados donde la variedad con respecto al diagrama de puntos tradicional es mayor. El diagrama de puntos es la representación idónea para expresar la relación entre dos variables (figs. 12 y 13), pero "su utilidad depende de la capacidad del observador para percibir e interpretar correctamente el gráfico" (Spence y Lewandowski, 1990). Las aportaciones más sustanciales de EDA en este campo son los conceptos de "linea resistente" y "suavizado". En la primera se trata de representar la individualidad de los datos para distinguir los contrastes entre individuos y ajustar el modelo adecuado. La novedad en este caso con respecto a la linea de regresión es que la linea resistente no se ve afectada por los valores extremos porque los considera formando parte del ajuste.

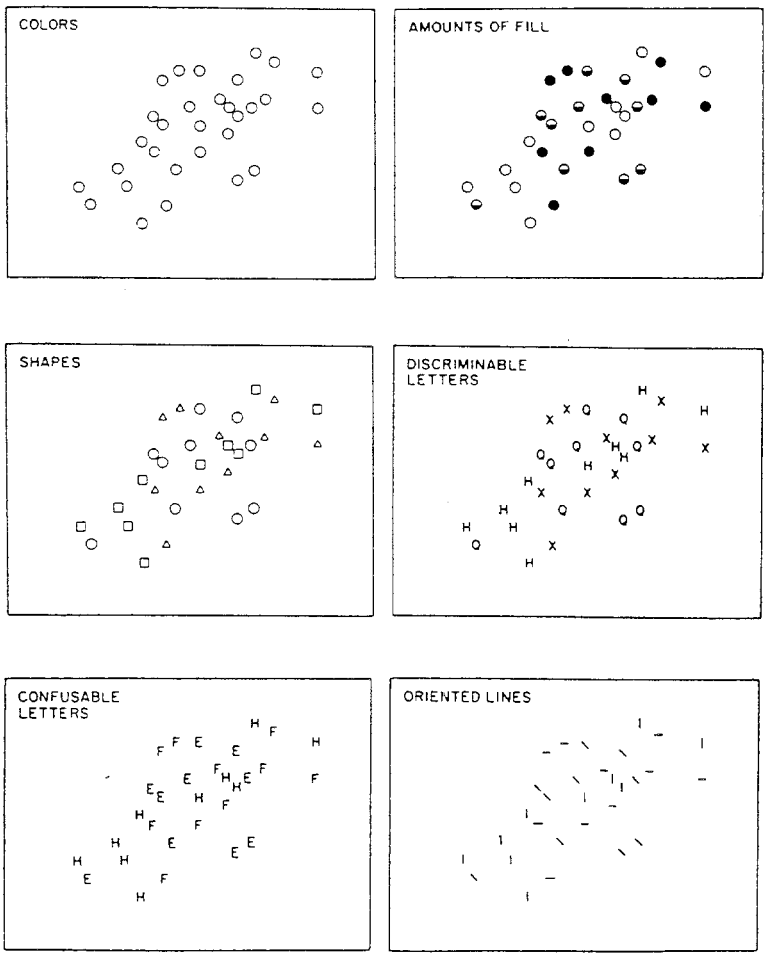

Figura 12.- Estratos de discriminación de diagramas de puntos (Fuente: Lewandowski y Spence, 1989).

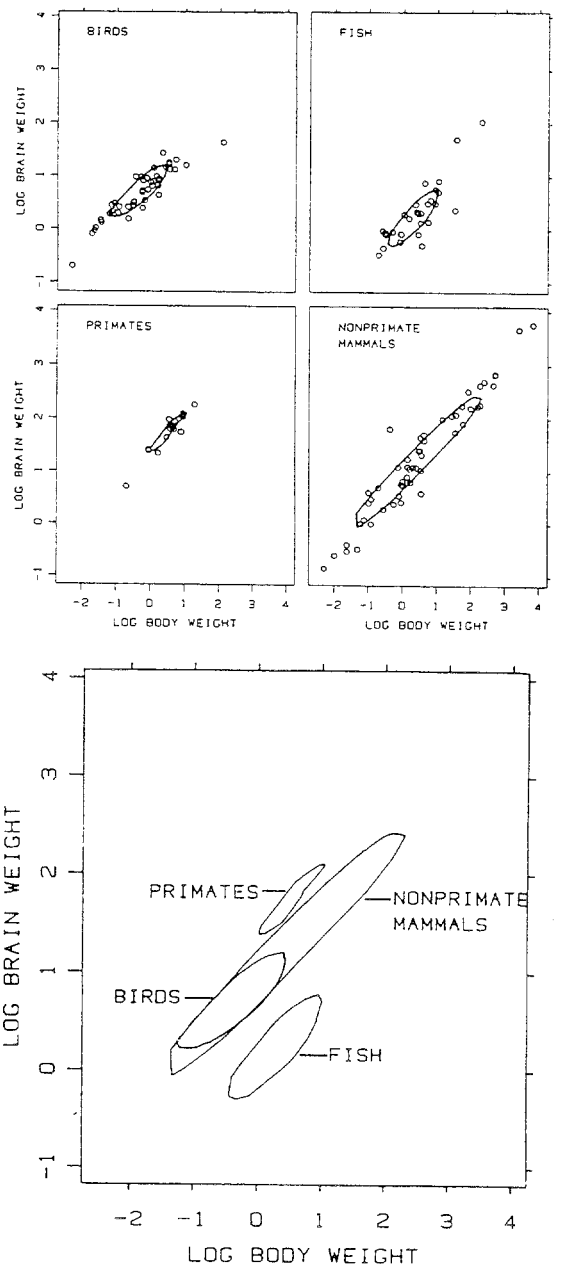

Figura 13.- Discriminación de subconjuntos en diagramas de puntos (Fuente: Lewandowski y Spence, 1989)

El problema surge cuando la cantidad de datos es elevada y pueden producirse solapamientos. Para ello el suavizado es una técnica resistente (fig. 14), para, manteniendo los principios de la representación gráfica, detectar las estructura básicas y residuales (Cleveland y McGill, 1984b; Tukey y Tukey, 1981b; Velleman y Hoaglin, 1981).

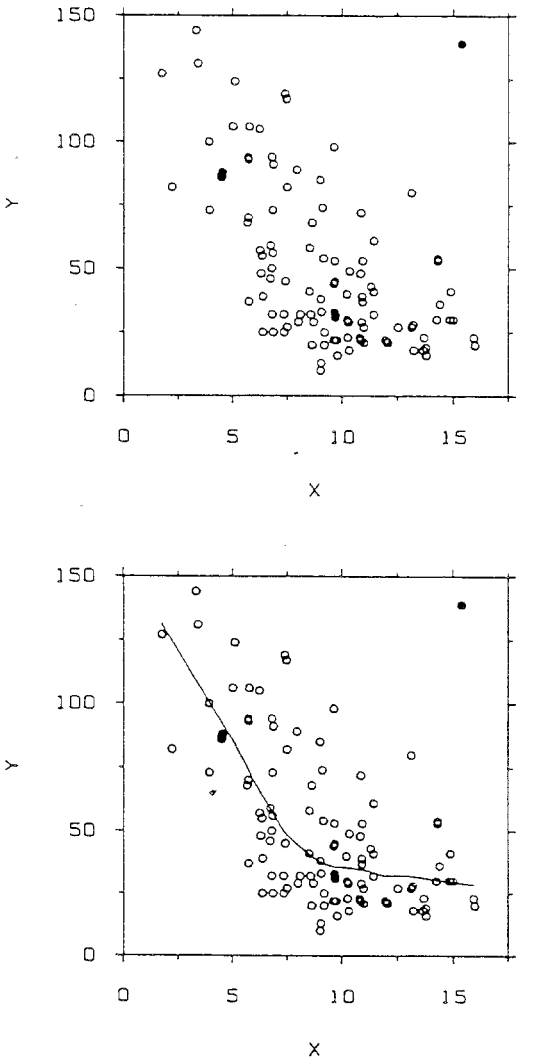

Figura 14.- Suavizado de diagrama de puntos por medio de ajuste no lineal ponderado (Fuente: Cleveland y McGill, 1985)

Desde otro punto de vista, los girasoles y los diagramas múltiples solucionan también el problema del solapamiento a través de la introducción de una tercera dimensión, que puede ser una diferente codificación para distintos conjuntos (Lewandowski y Spence, 1989), o dando valores a los datos solapados (Cleveland y McGill, 1984b) o partiendo el diagrama en varios subgráficos para identificar subconjuntos (Tukey y Tukey, 1981a).

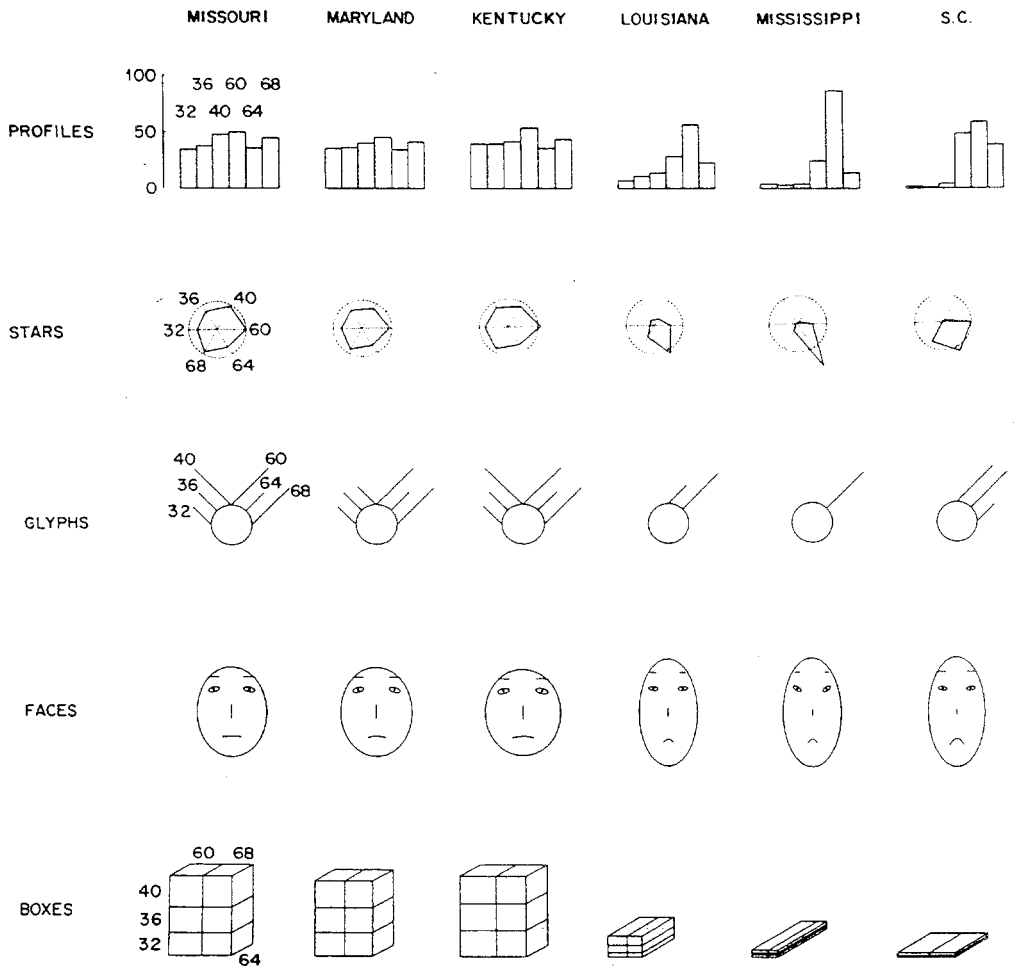

Figura 15.- Tipos de gráficos multivariados (Fuente: Kleiner y Hartigan, 1981)

Los gráficos multivariados ya no se basan en técnicas de análisis exploratorio, sino en los rasgos visuales de las composiciones múltiples. Se suele introducir la dimensión multivariada comparando gráficos individuales construídos según un patrón único. En general, ante el problema de las múltiples dimensiones, estos gráficos responden con notable imaginación, no exenta de dificultades, en muchos casos casi insalvables (Du Toit, Steyn y Stumpf, 1986).

Su objetivo fundamental es establecer contrastes visuales entre los individuos y el patrón general, pero esa función tiene una dificultad para el usuario no entrenado que suele encontrar problemas a la hora de discriminar los subconjuntos de datos.

No obstante esto, las curvas de Andrews son las más adecuadas para la representación multiple, junto con las caras de Chernoff y los árboles de Kleiner y Hartigan (figs. 16 y 17). Las primeras fundamentan la representación en una función de tipo armónico calculada para todas las variables, empleando los coeficientes obtenidos para el dibujo de las curvas. Los árboles emplean el orden de los valores de la variables en el conjunto de datos para conformar un patrón general que puede ser comparado con los casos. Las caras de Chernoff, a pesar de su positiva percepción por parte del lector, tienen problemas muy importantes para la explicación del contenido científico de los datos.

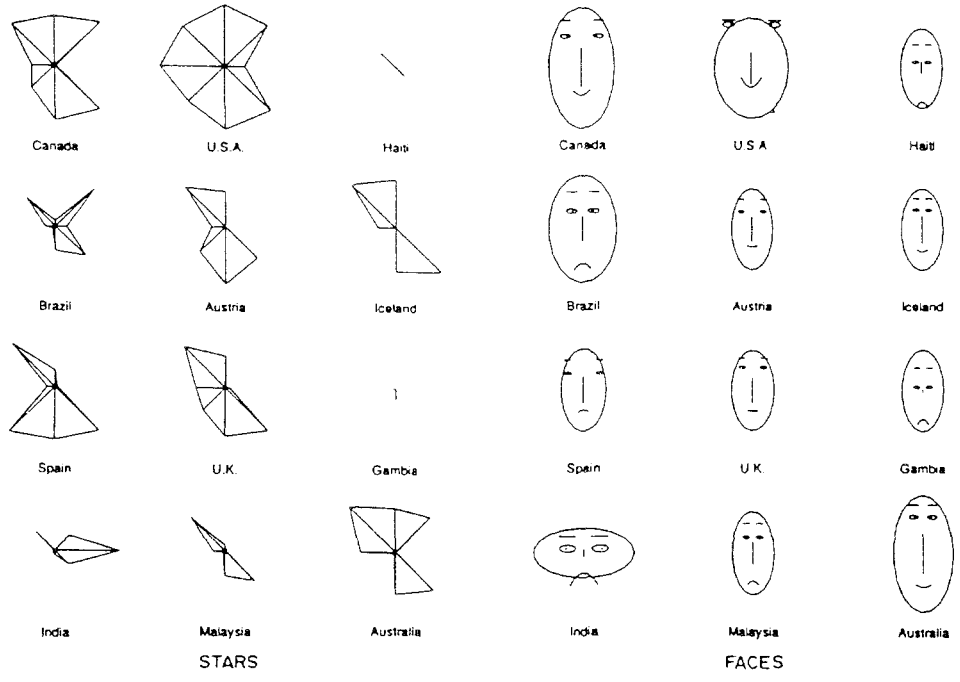

Figura 16.- Comparación de estrellas y caras de Chernoff (Fuente: Spence y Lewandowski, 1990)

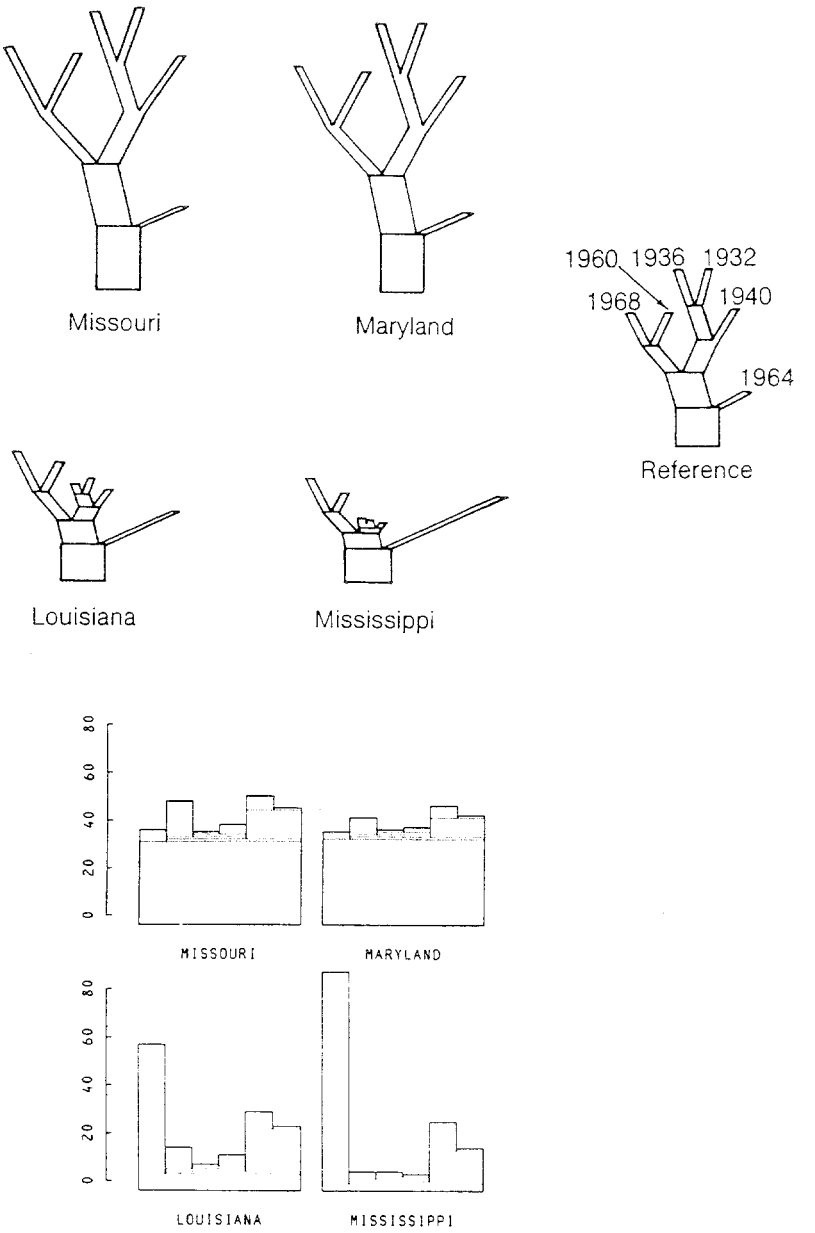

Figura 17.- Arboles y castillos (Fuente: Tukey y Tukey, 1981b)

Los otros gráficos adquieren su maxima difusión a partir de programas de ordenador, aunque su utilidad sea restringida por problemas semejantes a los de las caras de Chernoff.

BIBLIOGRAFIA

ANDREWS, D.F. et al. (1972): *Robust estimates of location: survey and advances.* New Jersey, Princeton Univ. Press.

ANSCOMBE, F.J. (1960): "Rejection of outliers". *Technometrics*, vol. 2, nº 2, p. 123-147.

ANSCOMBE, F.J. y TUKEY, J.W. (1963): "The examination and rejection of residuals". *Technometrics*, vol. 5, nº 2, p. 141- 160.

BARNETT, V. (ed.) (1981): *Interpreting multivariate data.* New York, Wiley and Sons, 374 p.

BARNETT, V. y LEWIS, T. (1978): *Outliers in statistical data.* New York, Wiley and Sons,

BECKER, R.A.; CLEVELAND, W.S. y WILKS, A. (1987): "Dynamic graphics for data analysis". *Statistical Science*, vol. 2, n. 4, p. 355-392.

BERGER, J.O. y BERRY, D.A. (1988): "Statistical analysis and the illusion of subjectivity". *American Scientist*, vol 76, p. 159-165.

BICKEL, P.J. (1976): "Another look at robustness: a review of reviews and some new developments". *Scandinavian Journal of Statistics*, vol. 3, p. 145-168.

BODMER, W.F. (1985): "Understanding Statistics". *Journal of Royal Statistical Society*, A, nº. 148, p. 69-81.

BOSQUE SENDRA, J. (1990): "Análisis estadístico exploratorio y confirmatorio en Geografía". *Actas IV Coloquio de Geografía Cuantitativa*, Palma de Mallorca, p. 405-445.

BRADSHAW, R.P. y PHILLIPS, H. (1992): "Bite less, chew more. The data analysis formula for adding value in the 1990s". *MRS 1992 Conference Papers* , p. 317-326.

BUJA, A. et al. (1986): "Discovering features of multivariate data through statistical graphics". *Proceedings of the Section on Statistical Graphics*, American Statistical Association, Washington, p. 98-103.

CARPENTER, E.H. (1987): "The evolving statistics and research process using microcomputer statistical software". *Social Science Microcomputer Review*, vol. 5, nº 4, p. 529-545.

CARPENTER, E.H. y AXELSON, R.D. (1989): "Statistical and graphical research methods: state of the art". *Social Science Computer Review*, vol. 7, nº 4, p. 503-534.

CLEVELAND, W.S. (1987): "Research in statistical graphics". *Journal of the American Statistical Association*, vol. 82, nº. 398, p. 419-423.

CLEVELAND, W.S. y McGILL, R. (1984a): "Graphical perception: theory, experimentation and application to development of graphical methods". *Journal of the American Statistical Society*, vol. 79, nº. 387, p. 531-553.

CLEVELAND, W.S. y McGILL, R. (1984b): "The many faces of a scatterplot". *Journal of the American Statistical Association*, vol 79, nº. 378, p. 807-822.

CLEVELAND, W.S. y McGILL, R. (1985): "Graphical perception and graphical methods for analyzing scientific data". *Science*, vol. 229, p. 828-833.

CLEVELAND, W.S. y McGILL, R. (1987): "Graphical perception: a visual decoding of quantitative information on graphical display of data". *Journal of the Royal Statistical Society*, 150, part 3, p. 192-228.

COX, D.R. (1978): "Some remarks on the role in statistics of graphics methods". *Applied Statistics*, vol. 27, nº. 1, p. 4-9.

CHATFIELD, C. (1985): "The initial examination of data". *Journal of the Royal Statistical Scociety*, A, nº. 148, p. 214-253.

CHATFIELD, C. y COLLINS, A.J. (1980): *Introduction to multivariate analysis*. London, Chapman Hall, p.

CHATFILED, C. y SCHIMEK, M.G. (1987): "An example of model-formulation using IDA". *The Statistician*, nº. 36, p. 357-363.

DEVILLE, J.C. y MALINVAUD, E. (1983): "Data analysis in official socio-economic statistics". *Journal of the Royal Statistical Society*, 146, part. 4, p. 335-361.

Du TOIT, S.H.C.; STEYN, A.G.W.; STUMP, R.H. (1986): *Graphical exploratory data analysis*. New York, Springer-Verlag, 314 p.

EHRENBERG, A.S.C. (1975): "Graphs or tables?". *The Statistician*, vol. 27, nº 2, p. 87-96.

ERICKSON, B.H. y NOSANCHUK, T.A. (1977): *Understanding data*. Mac Graw Hill.

FOX, J. (1990): "Describing univariate distributions", en FOX, J. y LONG, J.S. *Modern methods of data analysis*. London, Sage Publications, p. 58-125.

FRANCIS, I. (1981): *Statistical software. A comparative review*. New York, North-Holland, 542 p.

FREIXA BLANXART, M. et al. (1992): *Análisis exploratorio de datos: nuevas técnicas estadísticas*. Barcelona, PPU, 296 p.

HARTWIG, F. y DEARING, B.E. (1979): *Exploratory data analysis*. London, Sage Publications, 83 p.

HEALEY, M.J.R. (1984): "Prospects for the future: where has statistics failed?". *Journal of the Royal Statistical Society*, A, nº. 147, p. 368-374.

HEALEY, J. (1990): *Statistics. A tool for social research*. Belmont, Wadsworth, 2 ed., 431 p.

HIRSCHHEIM, R.A. et al. (1988): "A survey of microcomputer use in the Humanities and Social Sciences: a U.K. University Study". *Educaction & Computing*, nº 4, p. 77-89.

HUBER, P.J. (1972): "Robust statistics: a review". *Annals of Mathematical Statistics*, vol. 43, p. 1041-1067.

HUBER, P.J. (1977): "Robust statistical procedures". *Regional Conference Series in Applied Mathematics*. Society of Industrial and Applied Mathematics, Philadelphia, USA.

HUBER, J.P. (1981): *Robust statistics*. New York, Wiley and Sons, 302 p.

KLEINER, B. y HARTIGAN, J.A. (1981): "Representing points in many dimensions by trees and castles". *Journal of the American Statistical Society*, vol. 76, nº. 374, p. 260-269.

KRUSKAL, W. et al. (1960): "Discussion of the papers of Anscombe and Daniel". *Technometrics*, vol. 2, nº 2.

LEWANDOWSKY, S. y SPENCE, I. (1989): "Discriminating strata in scatterplots". *Journal of the American Statistical Association*, vol. 84, nº. 407, p. 682-688.

McDONALD, K.I. (1983): "Exploratory data analysis: a process and a problem", en McKAY, D.; SCHOFIELD, N; WHITELEY, P. *Data analysis and the Social Sciences*. London, Francis Pinter, p. 256-284.

McNEIL, D.R. (1977): *Interactive data analysis.*

McPHERSON, G. (1989): "The scientist' view of statistics - a neglected area". *Journal of the Royal Statistical Society*, 152, part 2, p. 221-240.

MOSTELLER, F. y TUKEY, J.W. (1977): *Data analysis and regression*. Addison Wesley, Reading, Mass.

RODRIGUEZ RODRIGUEZ, V. (1990): "Programas estadísticos a examen". *Actas IV Coloquio de Geografía Cuantitativa*, Palma de Mallorca, p. 247-260.

SCHMID, C.F. (1983): *Statistical graphics. Design principles and practices*. New York, Wiley and Sons, 212 p.

SPENCE, I, y LEWANDOWSKI, S. (1990): "Graphical perception", en FOX, J. y LONG, J.S. *Modern methods of data analysis*. London, Sage Publications, p.13-57.

TUFTE, E.R. (1983): *The visual display of quantitative information*. Cheshire, Graphics Press, 197 p.

TUFTE, E.R. (1990): *Envisioning information*. Cheshire, Graphics Press, 126 p.

TUKEY, J.W. (1977): *Exploratory data analysis*. Addison Wesley, Reading, Mass.

TUKEY, P.A. y TUKEY, J.W. (1981a): "Preparation, prechosen sequences of views", en BARNETT, V. *Interpreting multivariate data*. New York, Wiley and Sons, p. 189-213.

TUKEY, P.A. y TUKEY, J.W. (1981b): "Summarization; smoothing; supplemented views", en BARNETT, V. *Interpreting multivariate data*. New York, Wiley and Sons, p. 245-274.

TUKEY, P.A. (1986): "A data analyst's view of statistical plots". *Proceedings of the Section of Statistical Graphics*, American Statistical Association, Washington, p. 21-29.

VELLEMAN, P.F. y HOAGLIN, D.C. (1981): *Applications, basics and computing of exploratory data analysis*. Duxbury Press, Boston, Mass.

WANG, C.M. y GUGEL, H.W. (1986): "High-performance graphics for exploring multivariate data". *Proceedings of the Section on Statistical Graphics*, American Statistical Association, Washington, p. 60-65.

WOODWARD, W.A. et al. (1988): *Directory of statistical microcomputer programs*. New York, Marcel Dekker, 744 p.