Lo scatterplot in tre dimensioni come tecnica di analisi di dati (*)

Agostino Tarsitano
Università degli studi della Calabria
Dipartimento di economia e Statistica
87030 Arcavacata di Rende (Cs)
agotar@unical.it

Summary
The classical scatterplot represents data by placing a mark for each observations in the (x,y) plane. While a scatterplot by itself is an efficient graphic tool, its visual message can be substantially increased by showing a projection of a three-dimensional scatterplot of the data. If a series of projections is plotted in sequence as the viewer would see them while moving through three-dimensional space, the parallax resulting from motion enables the user to get a convincing and accurate perception of the point cloud as a three-dimensional object. The dynamic scatterplot is a powerful new technique which makes it possible to detect and to describe structure in data that is very difficult or impossible to see with any other technique of two-dimensional display.
.

Keywords:
Scatterplot dinamico, MACSPIN, Grafica multivariata

(*) Lavoro comparso negli Atti della XXXIV riunione scientifica della SIS, Siena 27-30 aprile 1988, 247-252

1. La rappresentazione grafica di dati multivariati

L'intento di questa nota è di mostrare l'utilità e l'importanza che ha, per l'analisi dei dati, l'applicazione dell'animazione computerizzata ad uno dei grafici più classici: lo scatterplot. Dopo una breve panoramica sui metodi di rappresentazione di dati multidimensionali seguirà, nel paragrafo due, la discussione dello scatterplot dinamico. Infine, nel paragrafo tre, verranno presentati due gruppi di applicazioni in cui si vedrà come questo cinegramma permetta un agevole studio di dati a tre dimensioni.

Sia $\boldsymbol{x}=(x_1,x_2,...,x_n)$ una osservazione n-variata e si supponga che le variabili siano espresse in unità standard. Secondo Jacob (1981) la rappresentazione nel piano della $\boldsymbol{x}$ si può ricondurre a due idee essenziali: scelta di un segno grafico che simboleggi l' unità su cui si osservano le x_i ed assegnazione delle x_i ai parametri componenti il grafico. Vediamo le più diffuse applicazioni di questi due principi:

Metodo dei profili

Il metodo dei profili rappresentata la **x** con un diagramma ad aste dove queste hanno basi su dei numeri d'ordine regolarmente spaziati ed altezze proporzionali alle x_i. I vertici delle aste vengono congiunti con una poligonale che delinea il "profilo" del dato **x**.

Metodo dei poligoni

Si tratta di una tecnica simile a quella dei profili solo che ora i vertici delle aste sono espresse nelle coordinate polari $r_i=x_i$ e $q_i=2i/n$.

Metodo dei glifi

La $\boldsymbol{x}$ viene rappresentata con dei raggi di lunghezza proporzionale alla coordinata x_i che si dirigono verso l'interno di (oppure si dipartono da) un cerchio di raggio costante.

Metodo dei cubi

Si assegnano le prime tre variabili alle tre dimensioni di un cubo. Se poi n fosse maggiore di tre le ulteriori dimensioni verrebbero rappresentate segmentando successivamente quelle già impegnate.

Metodo delle funzioni

Ogni osservazione viene associata ad una funzione del tipo seguente

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} x_i g_i(t)$$

le $g_i(t)$ vengono scelte in modo da essere ortogonali e normalizzate all'interno di un intervallo prefissato. Il caso più comune è quello delle curve di Andrews

$$f(\boldsymbol{x})=x_1/\sqrt{2} + x_2 sin(t)+x_3 cos(t)+x_4 sin(2t) \ldots \qquad per\ -p \le t \le p$$

Metodo delle facce

L'unità è rappresentata dalla caricatura di un volto e le x_i sono poste in corrispondenza con un tratto somatico: lunghezza del naso, piega della bocca, taglio degli occhi, etc.

Quelli elencati sono tentativi ingegnosi ed a volte suggestivi di realizzare la rappresentazione di dati multivariati. Hanno però in comune due carenze piuttosto serie. Per prima cosa il loro effetto è molto legato al modo in cui le variabili sono assegnate ai parametri del grafico, e secondariamente, i grafici stessi diventano illegibili allorché il numero di dimensioni e/o il numero di unità sia anche solo moderatamente grande.

Per superare il primo difetto Kleiner ed Hartigan (1981), hanno proposto dei nuovi grafici (*castles* e *trees)* in cui l'ordine di entrata delle x_i è determinato da un raggruppamento gerarchico (i due autori hanno adoperato il metodo del legame completo con distanza euclidea). Si tratta di un contributo importante che limita l'arbitrio nella combinazione variabile/parametro grafico e porta, almeno nel caso dei "trees", ad una maggiore efficacia esplicativa (Freni-Titulaer e Louv, 1984) Tuttavia, l'applicazione di queste e delle altre tecniche a matrici di dati molto grandi, richiede un pre-trattamento che riduca il numero di informazioni di cui si dovrà contemporaneamente tenere conto.

Si supponga in tal senso che a mezzo di tecniche di analisi multivariata: coordinate principali, analisi fattoriale, ricerca della proiezione ottima, analisi dello spazio minimo, etc. si riesca a limitare a tre il numero di variabili rilevanti. Si supponga pure che il numero di unità sia tale da scoraggiare l'uso dei metodi prima elencati e che inoltre si vogliano evitare tecniche poco pratiche come i plastici oppure gli ologrammi. In questo caso la visualizzazione dei dati potrà avvenire aggiungendo la terza dimensione ai tradizionali scatterplot. Ecco alcune alternative (Cfr. Schmid 1979, pp. 285-291).

Classificazione dei punti

Due variabili precisano la posizione del punto nel piano. Su questa posizione viene collocato un segno di dimensione o forma o colore congruo alla risoluzione globale del grafico, regolato dai valori della terza variabile.

Curve di livello

Come col metodo precedente due variabili servono a stabilire la collocazione nel piano del punto, mentre la terza variabile si rappresenta congiungendo i punti in cui essa presenta lo stesso valore.

Stereogrammi

La terza dimensione viene ottenuta con un solido che si innalza da un poligono del piano centrato sul punto definito da due delle variabili. Stabilito l'angolo di osservazione, dal grafico vengono rimosse tutte quelle parti che da tale angolo non sarebbero visibili.

Scatterplot isometrico

Le triple di dati vengono immaginate all'interno di un parallelepipedo regolare. Sul piano se ne riporta una proiezione isometrica con i lati trasparenti in modo da far intravvedere, stereoscopicamente, i punti.

Anche questi grafici offrono soluzioni interessanti ma parziali al problema della rappresentazione tridimensionale. I loro limiti stanno sia nella "discretizzazione" imposta alla terza variabile che nella "staticità" del grafico. Oltre a modificare l'assegnazione delle variabili agli assi nient'altro viene lasciato allo spirito di ricerca dell'utente.

2. Lo scatterplot dinamico.

I metodi discussi nel precedente paragrafo riflettono le capacità di elaborazione (numerica e grafica) che c'erano nei primi anni settanta. Da allora i progressi nell'hardware sono stati consistenti, ma ben pochi sviluppi sono seguiti nel software (interessante in questo senso l'articolo di Beninger e Robyn, 1978).

Una eccezione è stata l'impiego dell'animazione computerizzata delle cui possibilità lo scatterplot dinamico descritto in questo paragrafo è un esempio significativo. Questo cinegramma compare per la prima volta nel sistema PRIM-9 (acronimo di *Projection, Rotation, Isolation, Masking*) predisposto da Fisherkeller et al. (1975).

L'idea su cui esso si basa è molto semplice: la nube di punti dello scatterplot tradizionale viene assoggettata ad una serie di piccole rotazioni e la parallasse (il cambio apparente di posizione di un oggetto, determinato da un cambio nella posizione o direzione del punto da cui è osservato) che ne risulta, consente all'utente di cogliere lo sviluppo della nube nelle tre dimensioni. Ruotando la nube in varie direzioni si individua facilmente la prospettiva che meglio di ogni altra mette in evidenza le eventuali strutture presenti nei dati (ricerca "manuale" della proiezione ottima da contrapporre a quella "automatica" di Friedman e Tukey (1974)).

Il PRIM-9 fu realizzato con un IBM 360/91, uno dei più potenti dell'epoca, dedicato totalmente ad un terminale grafico, allora molto avanzato, l'IDIOM. Dati i costi elevatissimi (circa 175,000 dollari dell'epoca) sia della CPU che dell'hardware, il sistema ebbe scarso uso e diffusione limitata.

Il successore del PRIM-9 fu il PRIM-H, un sistema sviluppato ad Harvard sul finire degli anni settanta. Il nuovo sistema riuscì da un lato ad offrire le stesse capacità del primo ad un decimo del costo (il PRIM-H è legato ad un VAX 11/780 ed un terminale grafico ESPS) e d'altro lato, essendo incluso in un pacchetto statistico integrato, aggiungeva altre opportunità interessanti dal punto di vista del calcolo statistico. Il PRIM-H è un netto miglioramento rispetto al PRIM-9, ed ha infatti avuto una diffusione molto maggiore, Tuttavia, anch'esso dipende da un hardware costoso e non facilmente reperibile che ne impedisce un uso generalizzato nell'analisi dei dati.

Gli ultimi prodotti che si inseriscono nella linea dei PRIM sono L'ORION/I e lo SCATMAT. Prodotti che, oltre a migliorare ed ampliare le prestazioni dei loro predecessori (introduzione del colore e visualizzazione multipla), sfruttano le innovazioni (particolarmente i terminali grafici tipo raster) e la caduta dei costi nell'hardware, per arrivare a sistemi acquisibili con poca spesa. L'ORION/I costava nel 1981 circa 60,000 dollari. e si prevedeva che in pochi anni i prezzi si sarebbero più che dimezzati (McDonald, 1983).

Il forte progresso tecnico concretizzatosi negli ultimi anni con la produzione di microcomputer potenti e flessibili si è riflesso anche sui metodi di visualizzazione di dati multivariati. Ad esempio, (Donoho et al. (1986)) sono riusciti a realizzare un programma tipo PRIM eseguibile su personal computer Macintosh: il MACSPIN.

Si tratta di software che permette all'utente di dialogare in modo vivo e diretto con dati a più dimensioni. Qualità e tipo di prestazioni sono comparabili con quelle dei cinegrammi dei sistemi PRIM: visualizzazioni di nubi contenenti fino a 600 punti, rotazioni abbastanza continue, animazione rispetto ad una quarta variabile, interazione con il grafico; il tutto al costo di 140 dollari (più i 2000 dollari per il Macintosh).

3. Applicazioni

Lo scatterplot dinamico si basa su movimenti in tempo reale ed è difficile illustrarlo con figure statiche (le sue applicazioni vengono di solito presentate con dei filmati). In generale, la scarsa efficacia dei cinegrammi su supporto cartaceo si riflette soprattutto nella soggettività nella scelta della proiezione dei dati. Di questo bisognerà tenere conto nella lettura dei grafici qui presentati. In particolare ho scelto due gruppi di applicazioni: il primo sulla generazione di numeri pseudocasuali ed il secondo tratto da alcuni studi sulla distribuzione dei redditi

a) Applicazione ai generatori di numeri casuali.

Una tecnica molto diffusa per la generazione di numeri casuali è data dalla formula: $X_{i+1} = aX_i + c \ (mod\ m)$ con $a, m > 1, \ c \geq 0, \ a, c < m$. Se $c>0$ il generatore è di tipo congruenziale misto ed è di tipo congruenziale moltiplicativo se $c=0$. Numeri pseudocasuali sull'intervallo unitario possono poi ottenersi ponendo $U_i = X_i/m$. La relazione è ciclica, può cioè fornire al massimo m ($m-1$ se $c=0$) numeri diversi dopodiché la sequenza si ripete nello stesso ordine. Se il ciclo riparte dopo aver prodotto tutti i resti in modulo m, il generatore si dice a periodo completo. E' su questo tipo che si è concentrata maggiormente l'attenzione degli studiosi.

Quando n-tuple successive $(U_i, U_{i+1}, .., U_{i+n-1}), (U_{i+1}, U_{i+2}, ...,U_{i+n})$ vengono rappresentate nello spazio ad n dimensioni i punti si collocano su di un numero necessariamente limitato di iperpiani paralleli. La densità dei punti all'interno degli iperpiani e la distanza tra gli iperpiani stessi dipende dalla scelta del moltiplicatore a e del modulo m (si veda l'ottimo sunto dato da Ripley,1983).

Per illustrare il ruolo dello scattergram dinamico in questo ambito, si sono considerati 500 terne dai seguenti generatori

I) $X_{i+1} = 75X_i \ \ (mod\ 2^{16}+1)$

II: $X_{i+1} = 257X_i + 41 \ \ (mod\ 2^{16})$

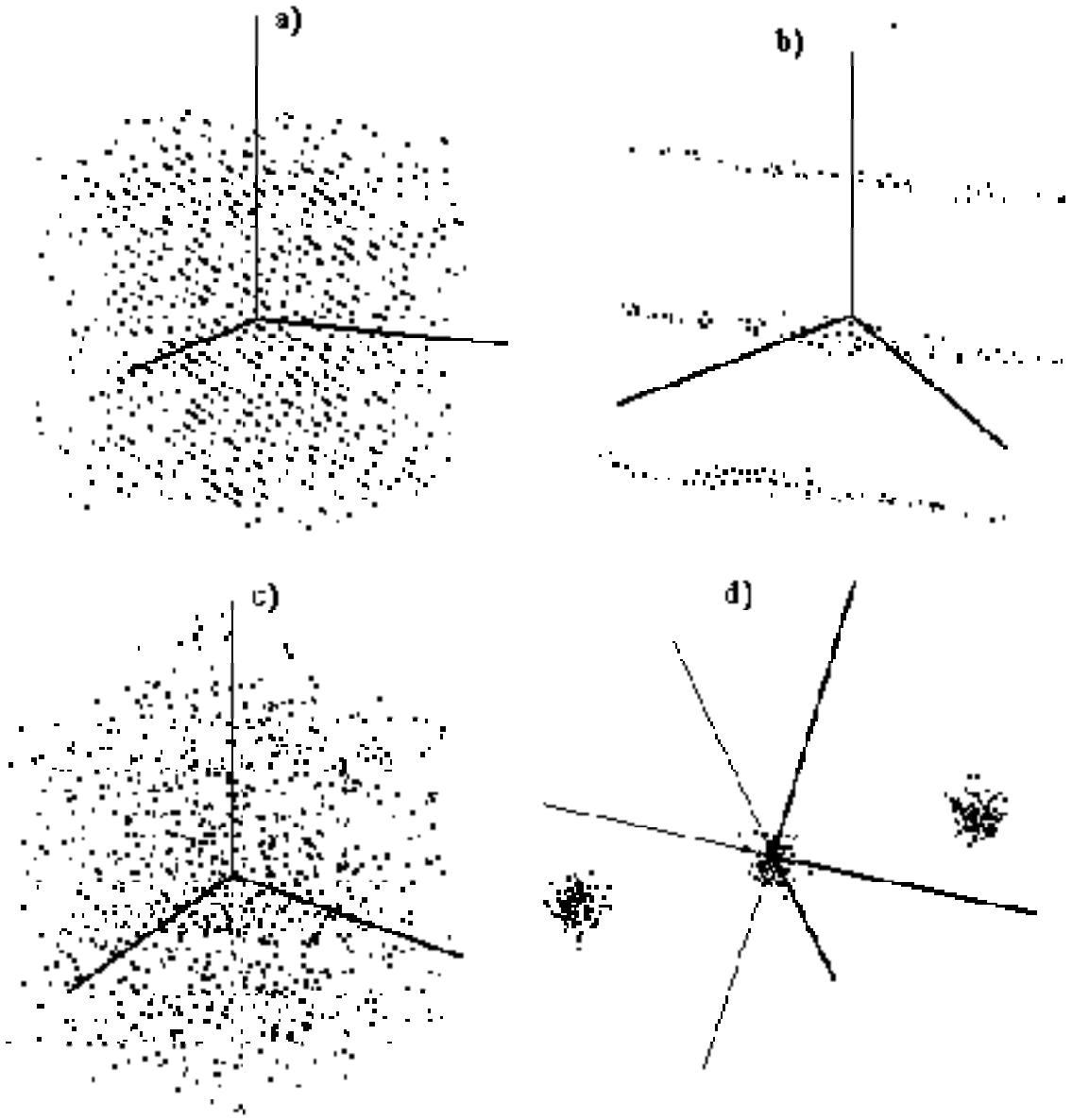
III: AS183

I primi due vengono discussi in Ripley (1983) e considerati specifici per microcomputer (il generatore I è implementato sul Sinclair ZX81). Il terzo schema è il ben noto algoritmo redatto da Wichmann ed Hill (1982)/

I grafici della fig.1 permettono di formulare un immediato giudizio sulla qualità dei tre generatori. Nel generatore I (parte a della fig.1) i piani in cui si raccolgono i punti sembrano abbastanza fitti anche se si intravedono degli allineamenti preoccupanti. Il generatore II (parte b) è palesemente scadente mentre il III conferma la sua validità.

L'ultimo generatore, insieme all'algoritmo "polare" di Marsaglia, è stato usato per la generazione di tre serie di 100 osservazioni da altrettante distribuzioni normali trivariate aventi la matrice identità come matrice di varianze-covarianze e con medie (10,10,10), (20,20,20) e (30,30,30) rispettivamente. L'esito di questa simulazione è raffigurato nella parte d della fig.1

Fig.1: Scatterplot dei generatori a) Moltiplicativo, b) Misto, c) AS183; d) Osservazioni gaussiane.

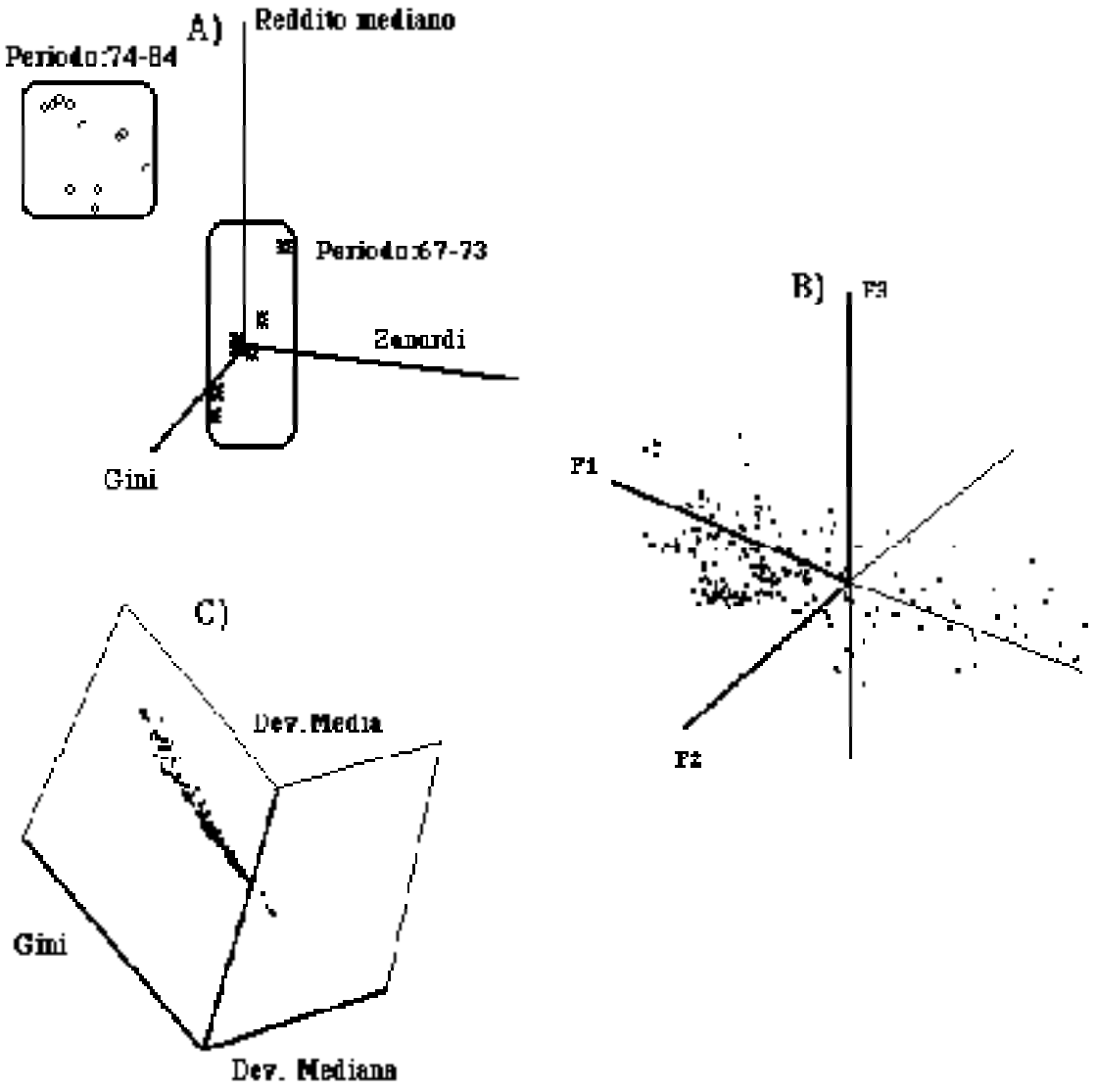


Applicazione allo studio della distribuzione dei redditi.

L'idea all'origine della parte A della fig.2 è che la distribuzione del reddito possa essere ricondotta (Gagliani e Tarsitano 1987) a tre aspetti fondamentali: livello (misurato dal reddito mediano, concentrazione (misurata con l'indice di Gini) ed asimmetria (misurata dall'indice di Zanardi).

Nel grafico sono state riportate queste misure per le distribuzioni del reddito degli USA dal '67 all'84. La presenza di due cluster ben distinti è evidente. Per ottenere il grafico B si è fatta una analisi delle componenti principali sui decili di 244 distribuzioni del reddito facenti parti della banca dati che si sta approntando presso il Dipartimento di

Economia Politica dell'Università di Calabria. Le prime tre componenti assorbono il 96% della variabilità (61% quella legata alla concentrazione, il 31% quella legata alla asimmetria e 4% la terza). Da notare Il forte addensamento nel quadrante Sud-Ovest del piano F1/F2 . Nella parte C del grafico si sono riportati i valori, relativi alle stesse distribuzioni usate per il grafico B, di tre modi alternativi di misurare la concentrazione: l'indice del Gini, la deviazione media e la deviazione mediana. L'impressione grafica è quella di una stele, segno che le misure, pur essendo concettualmente diverse, tendono però a misurare fenomeni estremamente correlati.



Riferimenti bibliografici

Beninger J.R., D.L. Robyn (1978). Quantitative graphics in statistics: a brief history. *The American Statistician*, 32,1-11.

Donoho A.W., Donoho D.L., Gasko M. (1986). MACSPIN. Graphical Data Analysis Software Austin, D2 Software.

Everitt B.S. (1978)., Graphical techniques for multivariate data, London, Heinemann Educational Books.

Freni-Titulaer L.W.J., W.C. Louv (1984). Comparisons of some graphical methods for exploratory multivariate data analysis. *The American Statistician*, 38, 184-188.

Fisherkeller M.A., J.H. Friedman, J.W. Tukey (1975). PRIM-9, an interactive multidimensional data display and analysis system. Proceedings of the 4th international congress of Stereology.

Friedman J.H., J.W. Tukey (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, *C*, 23, 881-890.

Gagliani G., A. Tarsitano (1987). Distribuzione personale del reddito, modifiche nella struttura dell'occupazione e sviluppo economico". pp. 246-251 in Strutture economiche e dinamiche dell'occupazione, A cura di C. Cazzola e A. Perrucci, La Nuova Italia Scientifica, Roma,

Jacob R.J.K. (1981). Comment" sull'articolo di Kleiner ed Hartigan, *Journal of the American Statistical Association*, 76, 269-270

Kleiner B., J.A. Hartigan (1981). Representing points in many dimensions by trees and castles. *Journal of the American Statistical Association*, 76, 260-269.

McDonald J.A. (1983). ORION I: interactive computer graphics in statistics. *Naval Research Reviews*, 2, 29-32.

Ripley B.D. (1983). Computer generation of random variables: a tutorial. *International Statistical Review*, 51, 301-319.

Schmid C.F. (1979). Handbook of graphic presentation. John Wiley & Sons, New York,

Wichmann B.A., Hill I.D.(1982). An efficient and portable, pseudo-random number generator. *Applied Statistics*, 31, 188-190.