Synthetic plots: some history and examples

David R. Brillinger
Department of Statistics
University of California
Berkeley, CA 94720
brill@stat.berkeley.edu
USA

April 28, 2014

Abstract Jerzy Neyman and Elizabeth Scott developed the idea of synthetic plots. These plots are a display of the data values of an experiment side by side with a display of simulated data values, with the simulation-based on a considered stochastic model. The Neyman and Scott work concerned the distribution of galaxies on the celestial sphere. A review of their wo is presented here followed by personal examples from hydrology, neuroscience, and animal motion.

Key words. Amazon river stages, bagplot, galaxies, history, monk seals. stochastic model assessment, spatial-temporal process, synthetic plots, time series.

1. Introduction

This Number of the SPJMS is meant to honor Professor Pedro Morettin and I am honored to be participating. Pedro was surely influenced by Professor Jerzey Neyman before, during and after his studies at Berkeley. Now Neyman had a serious general impact on Statistics at the University of Sao Paulo (USP) and Statistics in Brazil generally. The paper reviews both that impact and also Pedros career and then goes on to study a statistical method

that Neyman and Scott proposed for assessing a model and then moving on to a better model if necessary.

Pedro Morettin came from the city of Catanduva to live the city of Sao Paulo. He graduated in Mathematics from the University of Sao Paulo (USP) in 1969 and then moved on to g,raduate studies at Berkeley. He obtained his Berkeley doctorate in 1972 submitting a thesis titled Walsh-Fourier Analysis of Time Series. He then returned to Brazil and USP, taking a Berkeley born son with him. Pedros work with Walsh functions led several decades later to his work on wavelets. At Berkeley Pedro came to my classes, he and I interacted on o campo de futebol and became close friends as the years passed. Pedro has influenced many USP students to work on problems of time series analysis. His receiving the Mahalanobs Medal shows that applies to other countries as well. Many statistics students from Brazil have come to Berkeley to study. They include: Arujo, Bolfarino, Dantas, Folledo, Kondor, Pessoa, Rocha, Torres-Melo, Vares, and Wechsler. I played futebol with most.

As referred to above, Professor Neyman had a substantial impact on Statistics in Brazil during and following a visit in 1961. He spent a month in the city of Sao Paulo reviewing the situation of Statistics. A principal comment in his review was that the theory of statistics was not being treated there as an entity and in particular there was little focus on theoretical statistics. He proposed the creation of an Institute of Statistics to play a role like the Statistics Departments at Berkeley and Columbia. That , see Neyman (1961), details of that proposal were not realized until 1972 and classes finally started in 1975. Neyman recommended setting up a proper scientific publication. REBRAPE appeared in 1987 and the mathematical sciences journal RESENHAS started in 1993. Dantas (2002)provides historical information.

Neyman made a second visit to Brazil in 1978. He met with statisticians at the Institute of Pure and Applied Mathematics (IMPA) in Rio, particularly Kang and Barry James, and provided advice on the setting up a statistical laboratory. He gave five or six lectures there and then went on to the Second Brazilian Symposium on Probability and Statistics in Sao Paulo. That Symposium was dedicated to him.

Synthetic plots appear in Scott, Shane and Swanson (1954) and in Neyman and Scott (1956). In those papers synthetic plots were employed to develop a model for the distribution of galaxies on the celestial sphere. Clustering was a basic concern, see Abell (1975). The present paper will review, define and add examples.

The paper begins with this Introduction and follows with Sections: 2. Model appraisal methods and synthetic plots, 3. Spatial point process astronomy, 4. Time series river stages, 5. Temporal point process - neurophysiology, 6. Trajectories - marine biology, 7. Summary and discussion.

2. Model appraisal methods and synthetic plots

Model appraisal is basic to science and the scientific method. The scientific cycle involves: idea, experiment, data, model, model appraisal, new model if needed, and so on. Synthetic plots are an appraisal method capable of detecting unsuspected patterns.

Some model appraisal methods involve specific analytic formulations, for example they involve Pearsons chi-squared type statistics. Others are informal for example looking at various residual plots, or probability plots. This last can suggest model changes directly. However in Scott et al (1954) one reads that after a specific analytic check, the model must be subjected to **a number of checks** before it can be considered as more than a tentative working hypothesis.

The method of synthetics, Scott et al (1954), Neyman and Scott (1956), involves a display of actual data alongside model-based simulated data. One looks for differences between the two. One fits the model and then prepares a synthetic. Comparing the data and the synthetic together unsuspected features that are in the data, but not the model may show themselves. Features that the scientist hadn't thought of may show themselves and be used to improve on the model. In a formal testing approach a null hypothesis may have been rejected but the reason why is unclear. A comparison of the data and a synthetic plot can suggest why. Things can catch the eye, but not the formula.

3. Spatial point process astronomy

Concerning the question of why a synthetic plot one can quote Scott et al (1954), " the model must be subjected to a number of checks before it can be considered as more than a tentative working hypothesis.

Lick Observatory had long collected estimates of galaxy locations, see Abell (1975). Figure 1 shows an example that appeared in Scott et al (1954). The dimensions of the sky region studied are 6 degrees by 6 degrees. The sizes of various symbols/circles appearing provide a rough index of the brightness of the galaxies which they represent.

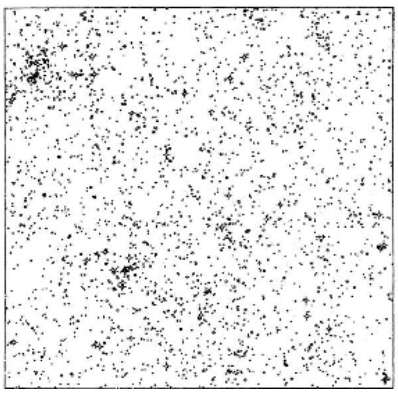

Figure 1: A plot of galaxies, see Scott et al (1954).

To construct a synthetic plot for these data a homogeneous marked Poisson model was considered for galaxy centers with marks. These last are brightness values and are assumed independent from location to location. On examination of the plot, Figure 2, the Poisson cluster center model was deemed unsatisfactory. The data plot however suggested clustering, for example note the cluster in the upper left corner.

Using "quasi-correlation functions these researchers estimated scale and amplitude parameters to describe the clustering. They then applied these parameters to the manufacture of a "synthetic" field of galaxy images based on a model that assumes that all galaxies are in clusters. They employed what is now called a Neyman-Scott cluster process, Daley and Vere-Jones (1988).

The paper Neyman and Scott (1972) lays out six basic assumptions for their model in this astronomical case. That model had been employed to create a synthetic photographic plate. The result is Figure 2. Clustering is apparent. It would not appear regularly in the Poisson case.

In the words of Scott et al (1954), "When the calculated scheme of distribution was compared with the actual distribution of galaxies recorded in Shanes photographs of the sky it became apparent that the simple mechanism could not produce a distribution resembling the one we see. They continued with, it was shown that the visual appearance of a synthetic photographic plate, obtained by means of a large scale sampling experiment conforming exactly with the assumptions of the theory, is very similar to the actual plate.

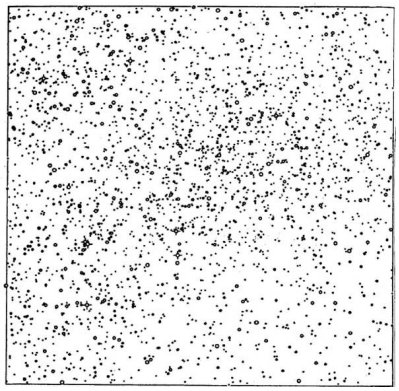

Figure 2: The simulated Neyman-Scott marked point process.

The only difference noticed between the two is concerned with the small-case clumpiness of images of galaxies."

An analytic comparison was also provided in Scott et al (1954). It compared is based the relative frequencies of counts in cells of the data with those of the synthetic. Still the synthetic let them look for an even better model.

4. Time series - hydrology

A water height measuring ruler has been situated at the end of a pier in Manaus, Brazil since 1903. It has been employed daily by the Manaus Harbour Limited to measure the daily stage there of the Rio Negro River, see Sternberg (1987). The annual mean levels for the period 1903 -1993 are graphed in the top panel of Figure 3. An outlier at year 1927 stands out. This was a period of serious fires. A horizontal reference line is also in the figure at the level of the median of the 90 observations.

Looking at the series in this top panel it seems that employing the following model for the data, Y_t, might be informative

$$(1)Y_t = \alpha + \beta Y_{t-1} + \gamma X_t + \sigma Z_t(1)$$

with $t = 1, ..., T$, $X_t = 1$ for $t = 24$ and 0 for the other times. This model was fit to the data by least squares.

Denote the estimate of $(\alpha, \beta, \gamma, \sigma)$ by (a, b, g, s). A synthetic series may be formed by evaluating $y_t, t = 2, , T$ via

$$y_t = a + by_{t-1} + gX_{t-1} + sz_t$$

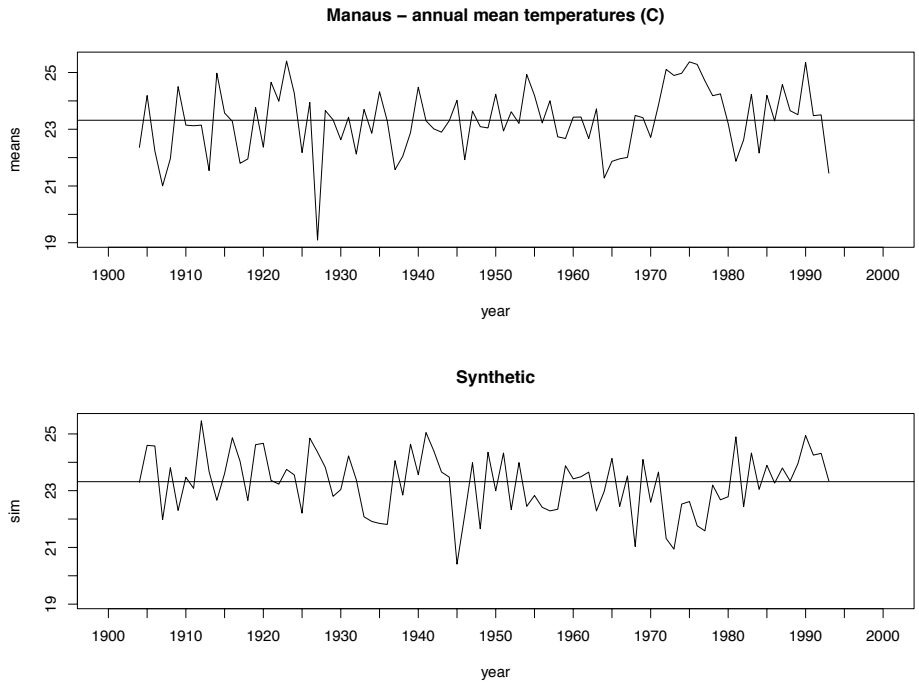

Figure 3: Top panel: Monthly Amazon river stages. Bottom panel: synthetic based on the model (1). The horizontal line is at the median level.

with the z_t independent standard normals. The synthetic series obtained is shown in the bottom panel of Figure 3.

The outlier is seen to be handled directly. Now the hump between 1970 and 1980 stands out as a difference between the two plots. One now sees some indication of autocorrelation in the top panel. That might be handled by making the error series autoregressive.

The advantage of presenting the synthetic plot for comparison is that some un-thought-of forms of departure might have been noticed.

5. Temporal point process neurophysiology

The data studied next is a nerve cell spike train. They can be viewed as a segment of a realizations of a point process. They were collected by Professor

J. P. Segundo at the University of California, Los Angeles. The data here is from the identified nerve cell L10 of *Aplysia californica*. The experiments in which these data were collected are described in detail in Bryant, Ruiz Marcos and Segundo (1973), Bryant and Segundo (1976) and Brillinger et al (1976).

These experiments and analyses have the ultimate goal of understanding how nervous systems work in strictly biological terms. *Aplysia* are often studied by because the nerve cells are large and accessible and a number are repeatedly identifiable.

Figure 4, top panel, is a rastor display of the times at which a sea hare neuron, L10, fires. It is a long time period of observation and there are many spikes. To display the data the observation time period is broken into 40 contiguous segments. These are stacked in the top panel of Figure 4.

The bottom graph is a synthetic realization obtained by a random permutation of the times between successive points. One might have generated a Poisson process by taking the times between spikes to be independent exponentials with the same mean.

One sees larger gaps along the horizontal lines of the synthetic. There would appear to be some negative correlation between the intervals of spikes in the actual data.

6. Trajectories - marine biology

By the trajectory of a particle is meant a 2-vector-valued function of time. A location is denoted by $\mathbf{r} = (x, y)$, time by t and one writes $\mathbf{r}(t)$ for the location of the particle at time t.

The model adopted is motivated by a Newtonian law of motion. It involves a scalar potential function, H, and leads to deterministic differential equations,

$$d\mathbf{r}(t) = \mathbf{v}(t)dt$$
$$d\mathbf{v}(t) = -\beta\mathbf{v}(t)dt \ - \ \beta\nabla H(\mathbf{r}(t), t)dt$$

Here $d\mathbf{v}$ denotes velocity and β denotes the coefficient of friction. If β is large the equation may be approximated by

$$d\mathbf{r}(t) = -\nabla H(\mathbf{r}, t)dt = \mu(\mathbf{r}, t)dt$$

with ∇ the gradient. The last equation shows that μ has an interpretation as velocity.

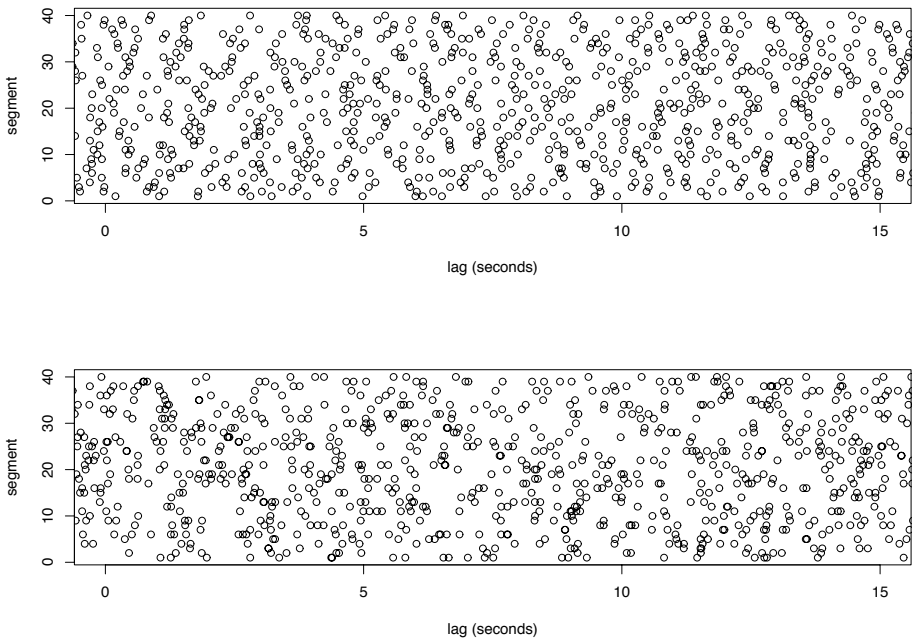

Figure 4: Top panel: rastor display of the L10 spike train data. Bottom plot: rastor display of a synthetic generated by a random permutation of the interspike times.

Examples of functions H are listed next. The symbol r may refer to the distance to a boundary or to a point.
Polynomial

$$H(x,y) = \beta_{10}x + \beta_{01}y + \beta_{20}x^2 + \beta_{11}xy + \beta_{02}y^2$$

Point of attraction

$$H(r) = .5 * \sigma^2 logr \quad \delta r$$

Point of repulsion from (0,0)

$$H(r) = C/r, C > 0$$

Attraction/repulsion

$$H(r) = \alpha(1/r^{12} \quad 1/r^6)$$

There are also nonparametric versions, for example involving splines.

A deterministic SDE can be adapted to a real world situation by working with analogous stochastic differential equations to handle uncertainty and measurement error. One form is

$$d\mathbf{r}(t) = \mu(\mathbf{r}(t), t)dt + \sigma\mathbf{r}(t), t)d\mathbf{B}(t)$$

with μ drift, σ the coefficient of diffusion and $\mathbf{B}(t)$ bivariate Brownian motion.

An Euler scheme, see Kloeden and Platen (1995), provides an approximate solution to this equation,

$$\mathbf{r}(t_{j+1}) - \mathbf{r}(t_j)/(t_{i+1} - t_i) = \mu(\mathbf{r}(t_j), t_j) + \sigma(\mathbf{r}(t_j), t_j)\mathbf{Z}_{j+1}/\sqrt{(t_{j+1} - t_j)}$$

with the entries of $\mathbf{Z}_j$ independent standard normals . This approximation may be employed with data: $(x(t_j), y(t_j), t_j)$ and an approximate likelihood function set down. Maximum likelihood estimates of the parameters may be obtained and plugged into the formula for H to obtain an estimate of the potential function.

There may be a boundary enclosing the movement. Methods for dealing with such may be found in Brillinger (2003). s In many case of the motion is on the surface of the Earth. In the presenr example the motion is that of a monk seal off the west coast of the Hawaiian island Molokai. Data were

collected by satellite observations of signals sent from a GPS transmitter attached to the seal. These seals spend their days foraging and resting and understanding their behavior and habitat use is critical to analyzing the continued decline of this endangered species.

Questions that have been asked include,

"What are the geographic and vertical marine habitats that the Hawaiian monk seals use?"

"How long is a foraging trip?"

"Are there age and sex differences in the habitats seals use when foraging?

They are discussed in Brillinger et al (2006).

The two panels of Figure 5 show the west of the island of Molokai. They also show the 200 fathom line delineating Penguin Bank. The left panel shows the well-located positions of the male juvenile animal concerned during for 15 days starting 13 April 2004, see Brillinger et al (2006).

A bagplot is defined via a bivariate median, a bag containing half the data which is defined by greatest depth, and a fence defined by inflating the bag by a factor of 3. Outliers are defined as points outside the fence. See Rousseuw, Rutts, and Tukey (1999). Bagplots are shown in both panels.

A synthetic plot is obtained in this case by generating a sample of locations making use of the fitted model. The method of fit is described above. The potential function employed is

$$H(\mathbf{r}) = \beta_{10}x + \beta_{01}y + \beta_{20}x^2 + \beta_{11}xy + \beta_{02}y^2 + C/r$$

where r is the shortest distance to the island from the seals current location. The animal is kept off the island by the C/r term. The simulated trajectory starts at the original start point. The result appears in the right hand panel. Zthe trajectory does not go outside Penguin Bank.

One sees that the bagplot has become larger and more circular. The number of outliers is reduced. It is to be remembered that the left hand plot is empirical based on the data only, while the right hand one involves estimates of the parameters of a stochastic model. This may be part of the explanation for its increased size.

8. Summary and Discussion

Thought for the day,

"The typical statistician has learned from bitter experience that negative results are just as important as positive ones, sometimes more so. Tukey (1967.)

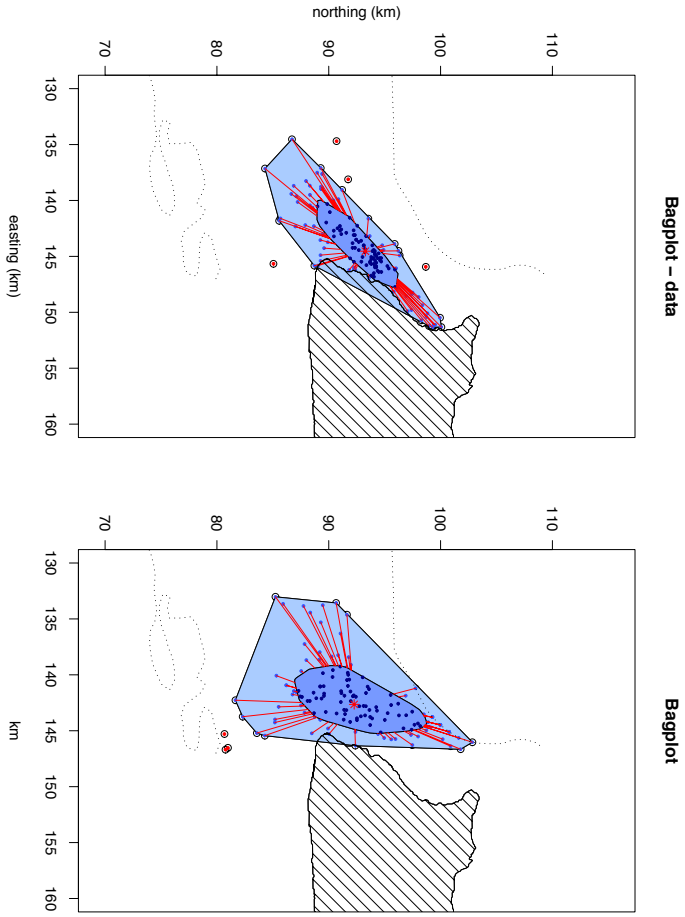

Figure 5: Left hand panel: bagplot of the data. Right hand panel: bagplot based on synthetic data. The dashed curve is the 200 fathom line. The trajectory stays inside Penguin Bank.

Because of Pedros associations with Berkeley during the Neyman years it seemed appropriate to consider a Neyman topic for this paper. Neyman and Scott made substantial use of a method they called synthetic plots in their development of a model for the distribution of galaxies on the celestial sphere. They plotted the data alongside a so-called synthetic plot, a simulation of data from a considered model. This paper has begun with details of the galaxy study and then presented synthetic plots from data from the authors applied research.

Synthetic plotting is a visualization method. Such can be very useful for model assessment and improvement. Unanticipated patterns might be noticed by the eye.

Four examples involving respectively: a spatial point process, a time series, a temporal point process and a trajectory in turn. Some model inadequacies were noted. They might be used to improve the considered model.

Acknowledgements.

Caro Pedro tenho saudade do Brasil e conversas com voce. Sou muito grato pela amizade e estimulacao das pesquisas de 1970 ate hoje. Ate a proxima.

I gratefully acknowledge the help from A. Ager, C. L. Littman, R. Lovett, A. Pinheiro, H. K. Preisler, J. P. Segundo, R. Spence, B. S. Stewart, and C. Toloi in preparing this paper.

Parts of the materials of this talk were presented at the Banff International Research Station Workshop on Forests, Fires, Stochastic Modeling in 2006.

This research was supported by the NSF Grant DMS-1007553.

References

[1] Abell, G. O. (1975), Clusters of galaxies. Galaxies and the Universe (eds. A. Sandiago, M. Sandavge and J. Kristian) IX.

[2] Brillinger, D. R. (2003) Simulating constrained animal motion using stochastic differential equations. Lecture Notes in Statistics 41, 35-48

[3] Brillinger DR, Bryant Jr, H. L., Segundo JP (1976) Identification of synaptic interactions. Biol. Cybern. 22,213-228.

[4] Brillinger, D. R., Stewart, B., and Littnan, C., (2006). A meandering hylje. Pp. 79-92 in Festschrift for Tarmo Pukkila on His 60th Birthday, Eds. E. P. Liski, J. Isotalo, S. Puntanen, and G.P.H. Styan.

[5] Bryant Jr., H. L., Ruiz Marcos, A., Segundo, J. P. (1973). Correlations of neural spike train discharges produced by monosynaptic connections and by common inputs. J. Neurophysiol. 36, 205-225
[6] Bryant Jr., H.L. and Segundo, J. P. (1976), Spike initiation by transmembrane current: a white-noise analysis, J. Physiol. 260, 279-314.
[7] Daley, D. J. and Vere-Jones, D. (1988). An Introduction to the Theory of Point Processes. Springer, New York.
[8] Dantas, C. A. B. (2002). O desevolvimento da estatistia na universidade de Sao Paulo Boletim de ABE XVIII No. 52, 49-69.
[9] Kloeden, P. E. and Platen, E. (1995). Numerical Solution of Stochastic Differential Equations. Springer, Berlin.
[10] Neyman, J. (1961). Organizational outline of the proposed Institute of Statistics at the University of Sao Paulo. Sao Paulo, Brazil.
[11] Neyman, J. and Scott, E. L. S. (1956). The distribution of galaxies. Scientific American 195, 187-200.
[12] Neyman, J. and Scott, E. L. S. (1972). Processes of clustering and applications. Pp. 546-681 in Stochastic Point Processes (ed. P. A. W. Lewis). Wiley, New York.
[13] Neyman,J. Scott, E. L. S. and Shane, C. D. (1953) On the spatial distribution of galaxies Astrophysical J. Supplement 117, 92-133
[14] Rousseuw, P. J., Ruts I., Tukey J.W. (1999). The bagplot: a bivariate boxplot.American Statistician53(4).
[15] Scott, E. L., Shane, C. D. and Swanson, M.D. (1954) Comparison of the synthetic and actual distribution of galaxies on a photographic plate. Astrophysical J. Suppliment 119, 91-112.
[16] Sternberg, H. O-R. (1987). Aggregation of floods in the Amazon River as a consequence of deforestration. Geografisika Annaler 69A, 201-209.
[17] Tukey, J.W. 1967 A statistician's comment. P.p. 41-47 in Electronic Handling of Information: Testing and Evaluation. Eds. A. Kent, O. E. Taulbee, J. Belzer and G. D. Goldstein. Washington DC, Thompson Book Co.