# A Model of Symbol Size Discrimination in Scatterplots

**Jing Li**
Eindhoven University of Technology
P.O. Box 513, 5600 MB, Eindhoven
The Netherlands
j.li@tue.nl

**Jean-Bernard Martens**
Eindhoven University of Technology
P.O. Box 513, 5600 MB, Eindhoven
The Netherlands
j.b.o.s.martens@tue.nl

**Jarke J. van Wijk**
Eindhoven University of Technology
P.O. Box 513, 5600 MB, Eindhoven
The Netherlands
vanwijk@win.tue.nl

## ABSTRACT

Symbols are used in scatterplots to encode data in a way that is appropriate for perception through human visual channels. Symbol size is believed to be the second dominant channel after color. We study symbol size perception in scatterplots in the context of analytic tasks requiring size discrimination. More specifically, we performed an experiment to measure human performance in three visual analytic tasks. Circles are used as the representative symbol, with eight, linearly varying radii; 24 persons, divided across three groups, participated; and both objective and subjective measures were obtained. We propose a model to describe the results. The perception of size is assumed to be an early step in the complex cognitive process to mediate discrimination, and psychophysical laws are used to describe this perceptual mapping. Different mapping schemes are compared by regression on the experimental data. The results show that approximate homogeneity of size perception exists in our complex tasks and can be closely described by a power law transformation with an exponent of 0.4. This yields an optimal scale for symbol size discrimination.

## Author Keywords

Symbol Size, Scatterplots, Size Discrimination, Graphical Encoding, Visual Analytic Task, Quantitative Model, User Experiment.

## ACM Classification Keywords

H.1.2 Models and Principles: User/Machine Systems – Human information processing; H.5.2 Information Interfaces and Presentation: User Interfaces - theory and methods.

## General Terms

Experimentation, Human Factor.

## INTRODUCTION

Symbols are used in scatterplots to denote objects of interest. Besides their position, other visual attributes can be used to represent multivariate information. The size of symbols is believed to be the second dominant visual channel of symbols after color [1]. It has an intuitive association with many analytic features of data, such as order, quantity and difference, and thus supports information decoding and interpretation well. However, there are few quantitative results on the perception of symbol size in an applied perspective, such as visual analytic tasks. Our question here is what sizes to use to support a visual analysis as well as possible.

Existing psychophysical research provides us relevant clues to design graphical encoding schemes. Among them, Stevens's Power Law depicts the relationship between the physical magnitude of a stimulus and the corresponding experienced magnitude [2,3]. The perceived magnitude is a power function of the stimulus, where the power coefficient depends on the tested physical channel of the stimulus [3]. However, for symbol size encoding it is not straightforward to select the appropriate value of the power. This is due to a number of reasons. First, experiments in psychophysics are set up in a context quite different from practical user tasks. Second, these experiments require the user to make judgments on a specific physical scale, for instance the length of projected lines or area of squares or circles [4]. However, we are not sure yet how these scales might be involved in size discrimination tasks. Third, individual variance has often been ignored in the analysis as only average data are reported, while visualization could potentially be optimized for a specific user.

We consider size encoding in applications of information visualization. Our aim is to pick sizes such that analytic tasks, such as distinguishing sets and counting outliers, are as easy as possible. We use perception models based on psychophysical laws as starting point. Furthermore, we hypothesize that the perceptual difference between symbol sets is negatively correlated with task difficulty, based on the Guided Search theory of visual attention [8,9]. Or, put simply, we assume that the more different two sets of symbols look, the easier it is to discriminate them, thereby enabling fast and precise pattern discovery.

Our earlier user experiments [10] revealed that certain non-linear patterns uniformly exist for varying sizes across different shaped symbols (circles, pentagons, squares, triangles, and stars with 3, 4, 5 and 6 legs) in discrimination tasks, but only four different sizes were used there. In this

*CHI 2010,* April 10–15, 2010, Atlanta, Georgia, USA.

paper, we aim to model the relation between size and discriminability in more detail. We take the circle, as the most commonly used shape, use eight linearly varying sizes, test with three visual analytic tasks, and measure user performance both objectively and subjectively. We leave the study of interaction between size and other visual channels to the next stage.

A quantitative model is proposed for describing the relation between stimulus sizes and task performance, and the model parameters are estimated from the experimental data. Following the principle of Homogeneity of Perception [22], a uniform perceived size scale across users is assumed. The resulting size scale can for instance be used to produce encoding schemes for visualization designers. In practice, continuous ranges are often split up into a limited number of discrete bins, and an ordinal scale is used to represent different data classes. Our model suggests how to pick the size of the symbols used to optimize discriminability rather than optimizing the accuracy of quantity.

In the following, we construct our quantitative model based on related work, report our user studies, analyze the experimental data, compare alternative models, and finally discuss the implications of our work for interface designers.

## RELATED WORK

Many different research fields study how humans process visual information. In psychophysics, the aim is to derive quantitative models for sensation and perception via vision and other modalities. Generic laws have been developed to describe perceptual mappings. Meanwhile, the fields of information visualization and statistical graphics focus on visual features of graphical objects and analytic tasks. The goal is to provide guidelines for the design of better displays of information in terms of easy discovery of patterns, and veracious interpretation. In the following subsections, we briefly review related work in these fields and point out the gap and challenge of modeling based on real user tasks.

### Laws in Psychophysics

Psychophysics deals with the relation between physical stimuli and subjective percepts, and therefore measures the human sensation of various physical stimuli quantitatively. The goal is to determine whether a subject can detect a simple stimulus, differentiate it from another, and describe the magnitude and nature of the difference. As the name of this field suggests, it is inspired on ideas and methods in physics. In parallel with the measureable physical world, psychologists try to define and measure the subjective world. The continuum nature of perception is assumed preconditionally in most of the studies [6]. Pioneers in this field have developed general relationships, such as the Weber-Fechner Law: $P = \log X$ and Stevens's Power Law: $P = X^{\beta}$, where $P$ presents the sensed or perceived magnitude and $X$ is the physical magnitude of the stimulus. Stevens's power law was empirically verified for many perception channels and also across different modalities [4, 5]. Furthermore, the law had been claimed to be true for the overall statistical properties such as the mean size of sets of symbols or mean brightness of sets of spots [11,12,13]. However, the common criticisms in the field are as follows:

- Stevens's method averages data in observations from different observers. It had been seriously questioned if averaging can be correctly applied to the sensation variability [6,14]. We believe that a more appropriate model can be achieved by modeling the individual perception and performance of different subjects in distinct test conditions.
- The judgment of a single stimulus was used in Stevens's experiments, *i.e.*, subjects were required to make judgments without an explicit reference. However, only relations between stimuli might provide a basis for judgment [6]. We believe that a discriminability scale is more meaningful in real use cases.
- Besides these methodological issues, Stevens's experiments require a clear instruction on what to judge. Alternative descriptions are however possible for object size, such as the width or height or the area. We are not sure what criteria people use when they perform tasks that involve size discrimination. Hence, it is relevant but not complete to understand human judgments regarding the length of projected lines or the area of squares or circles tested by Stevens and his followers.

In this paper, we focus on how to improve visual analytic work in practical applications, and argue that size judgment thus cannot be isolated from the working context. We moreover propose an alternative way to analyze size perception in such a more complex context.

### Study on Visual Encoding

Visual encoding of data aims at enabling high quality data analysis, for instance by selecting optimal visual channels. Thus, the rank order of different visual channels is determined by the task performance for visual analytic tasks.

In the work of Cleveland and McGill, ten visual channels were identified and related with elementary perceptual tasks. A ranking list was given in order of decreasing accuracy: position along a common scale, position along nonaligned scales, length or angle or slope, area, volume or curvature, shading or color density [16]. Among these elementary channels, length and area possibly relate with size judgment. The rank orders were based on Stevens's power coefficients, using the reasoning that smaller coefficients produce lower accuracy. However, we question the validity of this argument, since Stevens's Power Law does not model judgment errors or noise.

Christ gave different rankings according to the efficiency of search tasks and the effectiveness of identification tasks [1]. Size was ranked after color and before shape in both cases.

Nowell adopted search tasks and identification tasks as fundamental tasks in a visual scan and semantic interpretation [17]. Size ranked ahead of shape in time of task completion for nominal data, but not in any other situations. Tasks of judging correlations were used in a study by Lewandowsky and Spence [18]. This work has supported the contention that psychophysical judgments on single stimuli do not extrapolate simply to the perception of more complex displays.

In a study of visual separability of symbols [19], Tremmel reduced the level of task difficulty in order to inhibit mistakes. Subjects were shown two sets of symbols in a scatterplot, and they were instructed to find the largest set as quickly as possible. A 2D separation space of symbols was produced by means of a multi-dimensional scaling approach, using log decision time as similarity measure.

The applicability of all of the above results was constrained by few analytic tasks and inconsistency between them. Moreover they provide no quantitative models and no information about the configuration within a single visual channel. However, we can deduce that size is indeed an important channel, and mostly ranked right after color. Meanwhile, the different methods and approaches provide us with a comprehensive understanding of the problem in graphical encoding.

### Previous Work

In a previous study [10] we have assumed the existence of a separation space: a space in which different symbols can be positioned and where distances between them are proportional to their discriminability.

In that study, we have tested 32 symbols by configuring four linearly varying sizes and eight different shapes. The symbols were tested pairwise in three visual analytic tasks. These tasks were selected based on taxonomy of low-level visual analytic tasks [20] and are addressed in detail in the next page of the current paper.

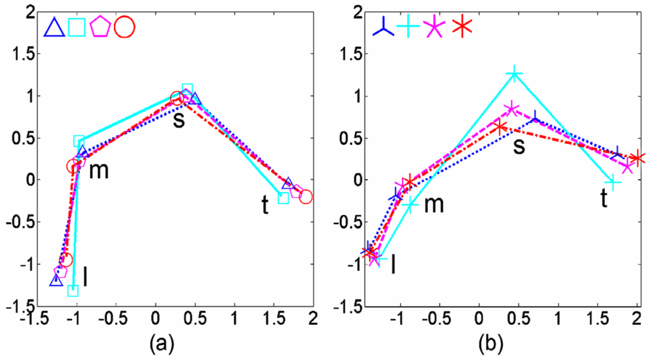


**Figure 1. Superposition of 2D projected size scale for different shapes: (a) polygons; (b) asterisks. t – tiny, s – small, m – medium, and l – large.**

A 3D separation space was established to position the symbols according to the discriminability between them. Further, different sizes of the same shape could be projected onto a 2D plane for each shape with little error. A superposition of these 2D planes is shown in Figure 1, which reveals a generic non-linear pattern for size across different shapes. First, the order of sizes is preserved so that larger size differences also result in larger separation distances. Second, the separation distances between larger symbols are relatively smaller than between smaller symbols. This indicates that equal size differences do not yield equal separation and particularly it is more difficult to distinguish larger symbols. Finally, the connected polylines bend significantly, expressing for instance that the size discrimination between tiny (t) and medium (m) is comparable to that between tiny (t) and large (l). This phenomenon repeats in our current study, and later we call it the saturation effect (see Figure 5).

In the current paper, we study size perception in more details.

## DISCRIMINATION MODEL

Symbols sizes can be expected to influence both objective task performances and subjective judgments on the difficulty of tasks. We aim to model both these overall mappings, with size perception as an intermediate step.

### Symbol Discrimination Model in the Task Context

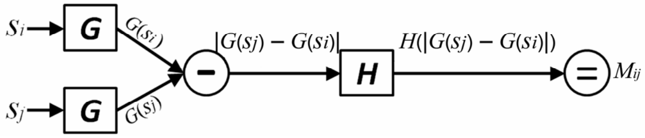


**Figure 2. Model of size discrimination in visual analytic tasks: $s_i$, $s_j$ are two different sizes; function *G* models the size perception; function *H* models the size discrimination in an analytic task.**

In Guided Search theory [9], visual attention involves top-down and bottom-up processes, and the bottom-up process is determined by how different a target is from its context. Also, according to Ware [15], whether something stands out preattentively is determined by the degree of difference of the target from the non-targets and the degree of difference of the non-targets from each other. We make the assumption here that a measure of the discriminability between targets and distractors should be based on the difference between the perceived strengths of their visual channels, instead of the difference between the physical scales of those channels. In other words, it is how they look different which determines the successive processing, rather than their difference in physical measurement. This enables us to model the visual analytic task process as two consequent steps, as shown in Figure 2. We assume a function *G* that transforms the original size stimulus onto a perceived or sensed scale, such that equal distances between stimuli on this scale denote equal separability. Stevens's power function could be a candidate description for *G*.

Further, we consider judgment in the context of analytic tasks. Typical visual analytic tasks are to compare symbol sets and to distinguish patterns from random clutter. In the simplest situation, just two different sets of symbols are used. If the perceived strengths of the two sets are very different, the task becomes easy, and can be finished quickly and precisely. More generically, we could assume that there is a monotonically increasing function $H$, which maps the difference of the perceived size strengths $|G(s_j)–G(s_i)|$ into the measured task performance $M_{ij}$, *i.e.*, the larger the size difference, the easier the task can be performed.

Function $G$ describes the relation between a physical measure and a value $P$ on the perceptual scale. Taking the radius $r$ of the circular symbol as a physical measure, it is given by $P=G(r)$. Given a certain $G$, we can construct perceptually uniform scales. The radii follow from $r_i=G^{-1}(P_i)$ ($i=1,…,N$), where $P_i$ is a linear interpolation on the perceptual scale. Using different instances of $G$, we can generate different sequences of symbols sizes. Some candidates are the power function, with different values for the exponent, and the logarithmic function:

- Physical length, Stevens's length judgment: $P=G(r)=\alpha\cdot r$, $\beta=1$;
- Physical area: $P=G(r)=\alpha\cdot r^2$, $\beta=2$;
- Stevens's area judgment, with the exponent 0.7: $P=G(r)=\alpha\cdot X^{0.7}=\alpha\cdot(r^2)^{0.7}=\alpha'\cdot r^{1.4}$, thus $\beta=1.4$;
- Fechner's logarithmic function: $P=G(r)=\log r$.

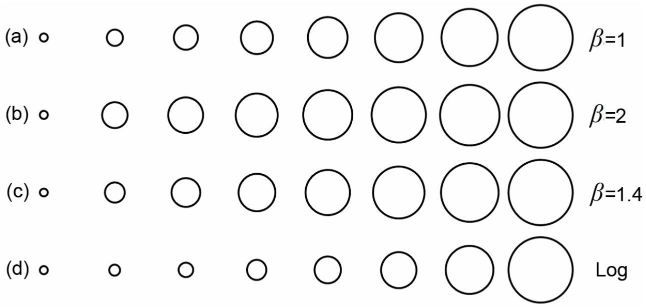


**Figure 3. Sequences of circles configured by linear interpolation on different scales: (a) radius scale; (b) area scale; (c) Stevens's area judging scale; (d) Fechner's logarithmic perception scale.**

The resulting scales are shown in Figure 3. We are not sure if one of these scales will yield the optimal discriminability in an analytic task context, or whether yet still a different scale is required. In the following we describe our experiments to obtain relevant measurements, followed by an analysis.

## EXPERIMENT 1

For an extensive study on the effect of size, we selected one standard symbol shape, the circle, and eight different sizes. Three relevant visualization tasks and two types of measures were selected and evaluated.

### Stimuli and Apparatus

Eight sizes of circles were used, with radii $i\cdot r$, $i = 1,…, 8$ and $r$=0.625mm. Participants were required to sit at one meter distance from a PC monitor. This setup gives experimental sizes within the range of 0.072°~ 0.573° in the visual angle of diameters, which covers those normally used in both print media and digital media. Below this range, symbols might not be viewed clearly, as the lower limit of visual acuity is around 0.01°; while for larger sizes, they might not be considered as symbols anymore.

We displayed 2D scatterplots on the PC monitor in a white plotting area of size 25×25 cm$^2$. In each plot, two sets of circles were used with different sizes. One size corresponded to target symbols, while the other corresponded to distractors. The total number of circles displayed in each plot was fixed to 50, occupying 10%-25% of the plotting area, in order to keep the viewing context more or less the same.

The plotting area was divided into 20×20 cells and each cell contained only one symbol to avoid unwanted overlap. Furthermore, the area of a cell was slightly larger than a specified symbol, which allowed for some random shift of the symbol shown inside the cell. This prevents symbols to assemble into lines. For each user task, 56 randomly generated plots were used. Subjects performing the same task viewed the 56 plots in different random orders.

### Analytic Tasks and Measures

To be consistent with previous work, we used the same analytic tasks as before. These were:

Task 1 ($T_1$): Visual Segmentation and Quantity Comparison

Instruction: Select the symbol that is presented most frequently in the plot.

Task 2 ($T_2$): Outlier Detection and Subitizing

Instruction: Locate and count the symbols of the type that is least presented (less than 5 times).

Task 3 ($T_3$): Distribution Characterization and Cluster Detection.

Instruction: Select the symbol that is distributed around the center of the plot with the smaller variance. (Or, select the symbol that has higher chance to appear in the center of the plot and converges more into a cluster.)

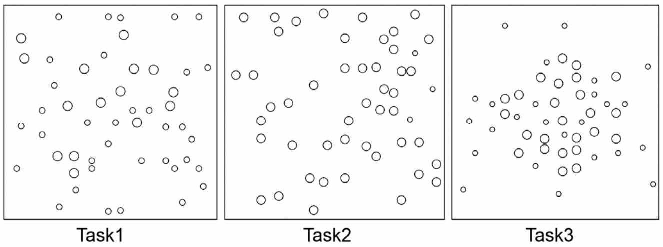


**Figure 4. Examples of the plots shown to subjects in the three different tasks.**

Examples of plots for these tasks are shown in figure 4. These three tasks occur frequently in routine analytic cases. They represent at least four fundamental tasks in a ten-task list of a recent taxonomy of low-level analytic activities [20], *i.e.*, computing derived value, characterizing distribution, finding anomalies, and clustering. Particularly, $T_1$ is the same as the quantity estimation task used by Tremmel [19]; $T_2$ is similar to the identification task in Nowell's work [17]; and $T_3$ is comparable to the correlation judgment task from Lewandowsky and Spence [18].

These tasks and their instructions were presented to the subjects in such a way that no prediction of the target is possible before the display of the plot. This design is in order to prevent the influence of a top-down attention process as indicated by Guided Search theory [9].

The three tasks were measured in both an objective session and a subjective session. The two standard measures for task performance are the time needed and the number of errors. Since we aim at the perceived scale in a separation space, we selected time as measure, as it gives a higher sensitivity in outcomes. We aim at minimizing errors, to assure the direct link between task accomplishment time and the symbol discrimination. For the different tasks we took the following measures to assure this:

$T_1$. The target type was presented 3 times more frequently than the distractor type;

$T_2$. Outliers were constrained to be within a range of 2° visual angle to keep the glances at one location and eliminate extra effort of eye movement for counting.

$T_3$. The target type had the same number of samples as the distractor type and had a four times smaller standard deviation than the distractor type.

These arrangements make the tasks fairly simple and the difficulty is mainly determined by the contrast between the two sets of symbols rather than by the complexity of the task itself. Therefore, a higher contrast makes the task easier, and can be performed quicker; while a lower contrast makes the task more difficult, and requires more time. Our pilot studies and the formal experiment showed that the above task setup indeed yielded very few errors (0−5%). As performance measure in the data processing, we use $M_{ij}=1/t_{ij}$ ($t$ denotes the measured time, $i,j=1,2,\ldots,8$ and $i\neq j$), such that a higher sensitivity corresponds to a shorter time performance.

For the subjective measurement, we used a zero-to-ten-point scale to measure the subjective opinion of the symbol discriminability in terms of task difficulty. All plots used in the objective sessions were again presented to subjects in the subjective session. Subjects were instructed to rate the plot difficulty in two steps: first categorize plots in difficult (0−3), neutral (4−6) and easy (7−10), and then give further differentiation within a category. Point 10 indicated the easiest plots among the total of 56, and since the tasks were simple and produced almost no errors, point 10 also implies the highest contrast between the two sets of symbols among the 56 different random plots. Particularly in the data processing, we use $M_{ij}=\text{rating}_{ij}/10$ ($i,j=1,2,\ldots,8$ and $i\neq j$), *i.e.*, the normalized rating.

### Subjects and Procedure

Twenty-four subjects were recruited from different departments of the Eindhoven University of Technology, all students or researchers. They all had experience in using statistical graphs, but in different fields. All of them had normal or corrected-to-normal vision and they were aged between 24 and 34 years, and balanced in gender.

Subjects were divided into three groups, with each group assigned to one task condition in two separate measurement sessions. The objective session was performed before the subjective session. A training session preceded the test session, and instructions were given together with the training session. It was emphasized for the objective session that the decisions should be made as quickly as possible, under the precondition that subjects took sufficient time to perceive the plot clearly and did not guess the answers.

In the objective sessions, the test was started by pressing the space bar. A blank screen was presented for 1.5 seconds to clear the viewport, after which a plot was displayed and an internal clock was started. As soon as the subject decided on the answer, she or he hit the space bar to stop the timing and the plot was replaced by a response interface. The response screen displayed the two symbols that were used in the plot for $T_1$ and $T_3$, or displayed “0”, “1”…, ”5” as the number of outliers for $T_2$. Next the subject could select the answer. For $T_1$ and $T_3$ the smaller circle was always at the left side and the larger circle at the right side on the response screen, to reduce cognitive error of answer inputting. Next, the test continued with the next plot by pressing the space bar again.

In the subjective sessions, the test procedure was similar as in the objective sessions. The only difference was that decisions were made without time constraint and that the same10-point rating scale was displayed on the response screen in all task conditions.

## DATA EXPLORATION

We have specified our model in Figure 2 in terms of functions $G$ and $H$. Before we fit specific functions, we first consider some underlying assumptions of this model.

### Symmetry of Symbol Discrimination

One underlying assumption is that how symbol $A$ differs from symbol $B$ is the same as how symbol $B$ differs from symbol $A$. In other words, the output of $H$ should be constant when the target set and distractor set are swapped. We verify this as follows.

For each subject the 56 plots viewed in a session lead to a 8×8 dissimilarity matrix, filled with $M_{ij}$ ($i,j=1,2,\ldots,8$ and $i\neq j$) and with an empty diagonal. The assumption made is valid if this matrix is symmetric. We use the following method to test the matrix symmetry. A matrix of a symmetric predictor is created by averaging the data of the same symbol pair tested in the observations of role swapping for each subject. Then an ANOVA test can be performed between the symmetric predictor and the experimental samples [21]. For all the twenty-four subjects, we found no evidence of asymmetry, which indicates that swapping the role of the target and distracter symbol has no significant effect on the experimental data.

### Visualize the Perceived Size Scale

Stevens's power law is a possible candidate for the function *G*. Furthermore, we assume that *H* is monotonic and describes the cognitive process, starting from the perceived difference. To get more insight, we visualize the results first in an explorative way.

One straightforward approach to deal with dissimilarity data is multi-dimensional scaling (MDS) [21]. We can use it here to observe how the eight circles with different sizes are arranged in a separation space. We have tried one dimensional to five dimensional MDS, and found that a 3D space gives the best balance between model complexity and prediction error. Figure 5 shows two 2D views of the 3D separation space, for $T_1$ tested with 8 subjects, taking $M_{ij}$ as the measure for dissimilarity. It agrees with what we obtained in previous work [10].

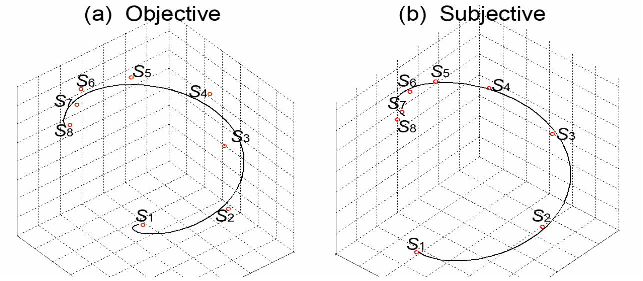


**Figure 5. Curve fitting with 3D MDS results: the red dots represent the 8 sizes; the distance between every two dots represents the size contrast; the spiral curve represents the continuous scale of the perceived sizes.**

Assuming the continuous nature of perception, if we sample as many as possible different sizes, the projected points of these samples into the perceptual separation space should form a continuous scale. Therefore, we used a 3D curve fitting to observe the status of the continuous perceived scale of size. Our observations are:

- The 8 points are located on a conical spiral curve, in order of increasing radius. This indicates that both *G* and *H* increase monotonically.
- The spacing between the points along the curve (on-path length) decreases gradually for increasing radius. This indicates that the derivative of *G* decreases monotonically.
- The curve indicates a saturation effect. If three points *l*, *m* and *n* stand for sizes that $S_l<S_m<S_n$, and $D_{ij}$ stands for the separation distance between points *i* and *j*, then $D_{ln}<D_{lm}+D_{mn}$. As shown in Figure 5, the saturation effect is strongest for high radii compared to the smallest one, *i.e.*, $D_{18}$ is almost equal to $D_{17}$. This indicates that the derivative of *H* also decreases monotonically.

If we use a power function for *G,* the power coefficient must be larger than 0 and smaller than 1 to fulfill these requirements. Also, if *H* is a power function, its power coefficient must also be larger than 0 and smaller than 1.

### Relate the measured data and the perceived size difference

One of our assumptions is that the objective measurements, *i.e.*, the measured time, are negatively associated with the size differences perceived by subjects. Independent of the specific form of *H*, this indicates that the rank order of time should correlate negatively with the rank order of perceived size difference. Using task performance $M_{ij}=1/t_{ij}$, we expect a highly positive correlation for the rank orders, and the same holds for the subjective measurements, a highly positive correlation is expected for the rank orders between ratings of task easiness and perceived size differences.

We hypothesize *G* as

$$P = G(S) = \alpha(S + s_0)^{\beta} \qquad (1)$$

where $s_0$ models a possible threshold of size perception.

Thus the Spearman's rank order correlation $\rho$ can be treated as a function of $\beta$ and $s_0$

$$\rho = \psi(\beta, s_0) = 1 - \frac{6 \cdot \sum D_{ij}^2}{N(N^2 - 1)} \qquad (2)$$

where , $D_{ij} = Rank_{|G(s_j;\beta,s_0) - G(s_i;\beta,s_0)|} - Rank_{M_{ij}}$ and $N$=56.

We studied the graphs of the 3D surfaces of $\rho=\Psi(\beta,s_0)$ for every subject assuming the meaningful range that $0<\beta<1.4$ and $0<s_0<2$ with step 0.05. We observed that the peaks of the surfaces were always at or near to $s_0=0$. Thus, we decided to drop the parameter $s_0$ to simplify the optimization problem. The values of the optimal $\rho$ obtained for each subject are presented in Figure 6 with solid green line for objective data and solid red line for subjective data.

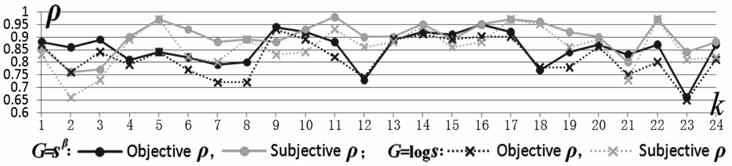


**Figure 6. The optimal rank order correlation $\rho$ for every subject (denoted by *k*) in different measurements: objective data and subjective data; and with different G: power function and logarithmic function.**

We can observe that the correlation values are quite high and approach the maximum 1 (the critical value of $\rho$ is smaller than 0.4 for $N$=56 at the 0.01 significance level) which suggests that our assumption is sound. The power coefficients which produced the optimal $\rho$ are in the range of 0.1–0.7, average to 0.4763 in $T_1$, 0.3738 in $T_2$, and 0.4325 in $T_3$ for objective measurement and average to 0.4725 in $T_1$, 0.4963 in $T_2$, and 0.3650 in $T_3$ for subjective measurement.

The rank order correlation $\rho$ produced by the logarithmic form of $G$ (without parameters) is also presented in Figure 6. The correlation values are also quite high, which suggests another possible form of $G$.

| Measure -ment | $G$ | Max $\Sigma\rho$ | Ave $\rho$ | $\beta\rightarrow$ |
|---|---|---|---|---|
| Objective | $G(s_i; \beta_{obj})$ | 20.12 | 0.84 | [0.38, 0.48] |
| Subjective | $G(s_i; \beta_{sub})$ | 21.26 | 0.89 | [0.38, 0.48] |
| All | $G(s_i; \beta_{all})$ | 41.37 | 0.86 | [0.38, 0.48] |

**Table 1. The optimized sum of Spearman rank-order correlation coefficients for all subjects assuming a shared power function.**

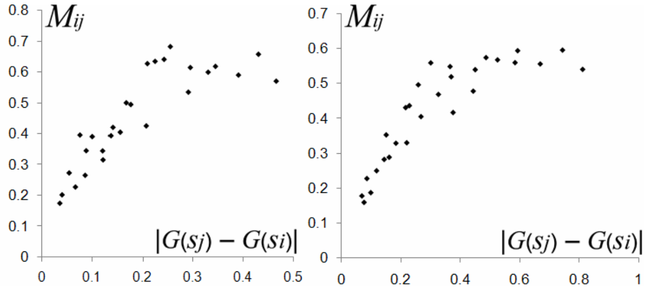


**Figure 7. Examples for subject 1 and 17: *x* – the perceived size contrast given by *G*; *y* – the reciprocal of measured time.**

Our next hypothesis is that all subjects share the same power function $G$. This hypothesis is known as the principle of Homogeneity of Perception [22]. To explore this, we add up all the Spearman's rank order correlations calculated on an individual basis. The sum is a function of the shared $\beta$ for all the subjects. The optimal $\Sigma\rho$ were obtained throughout an interval of $\beta$ in Table 1. In Figure 7 we show the relation between the perceived size difference $|G(s_j)–G(s_i)|$ and $M_{ij}$ for two subjects with $G$ determined by their individual $\beta_k$ ($k$ denotes the subject number and $k$ =1 and 17). Although there is quite some variation, the overall pattern suggests that $H$ can indeed be described by a power-like relationship.

## MODEL FITTING

This exploration of the data has verified our main assumptions and given us suggestions for the functions to use. We use nonlinear regression models to fit the data and to perform inferences [23]. $H$ is assumed to be in the following power-like [21] form:

$$H = a \cdot ((|G(s_j) - G(s_i)| + d)^b - d^b) \qquad (3)$$

where the free parameters $b$ and $d$ denote, respectively the nonlinear effect of the user responses in the discrimination tasks and the threshold of size contrast perception. Meanwhile, $G$ is given by $G(r)=r^\beta$ or log $r$. The parameter $a$ is a linear regression parameter, and the complete nonlinear regression model is:

$$M_{ij} = H + \varepsilon_{ij} \qquad (4)$$

where $\varepsilon_{ij}$ is assumed to be normally distributed with zero means and unknown standard deviation $\sigma$ [23].

### Perception Homogeneity and Estimation of *G*

We can make different assumptions for the free parameters, dependent on how we share cases. For instance, we can aim at one shared value for $\beta$, or we can fit individual values per subject or task. According to Maximum Likelihood Theory [21,23], the most appropriate model should be selected by tradeoff between the model fit in terms of the log likelihood and the penalty from the model complexity in terms of the total number of free parameters. The *AIC* index is the most widely adopted measure for this purpose [24]. We calculate *AIC* indices for different models and present them in Table 2. The model selection is based on $\Delta AIC$ [24] ($AIC_{\text{the simpler model}} - AIC_{\text{the more complex model}}$).

| Model | | *AIC* Index (-2*L*) | |
|---|---|---|---|
| Func. | Free Parameters (No. of parameters) | Obj. Data | Subj. Data |
| $G=r^\beta$ | $\beta_k, a_k, b_k, d_k, \sigma_k$ (120) | 1087 (822) | 1324 (1060) |
| | $\beta, a_k, b_k, d_k, \sigma_k$ (97) | 1048 (838) | 1313 (1106) |
| | $\beta, a_k, b_k, d_k, \sigma$ (74) | 997 (840) | 1263 (1104) |
| $G$=log$r$ | $a_k, b_k, d_k, \sigma_k$ (96) | 1110 (902) | 1353 (1146) |
| | $a_k, b_k, d_k, \sigma$ (73) | 1205 (1050) | 1433 (1278) |
| $G=r^\beta$ | $\beta_{Ti}, a_{Ti}, b_{Ti}, d_{Ti}, \sigma_{Ti}$ (15) | 1996 (1966) | 1444 (1414) |
| | $\beta, a_{Ti}, b_{Ti}, d_{Ti}, \sigma_{Ti}$ (13) | 1994 (1968) | 1442 (1416) |
| | $\beta, a_{Ti}, b_{Ti}, d_{Ti}, \sigma$ (11) | 1992 (1970) | 1440 (1418) |
| $G$=log$r$ | $a_{Ti}, b_{Ti}, d_{Ti}, \sigma_{Ti}$ (12) | 2018 (1994) | 1492 (1468) |
| | $a_{Ti}, b_{Ti}, d_{Ti}, \sigma$ (10) | 2030 (2010) | 1492 (1472) |

**Table 2. *AIC* index compared among different models. Gray cells are for the shared *H* within the specific task and measurement condition, $k$=1,2, …,24 and $T_i=T_1, T_2, T_3$**

The comparison shows that models with shared parameter $\beta$ and $\sigma$ are preferred over models with individual parameters $\beta_k$ and $\sigma_k$ for both objective data and subjective data, and the best model is ($\beta$, $a_k$, $b_k$, $d_k$, $\sigma$). For objective data, the estimation result from the best model is $\beta$=0.3880 with 95% confidence interval [0.3797, 0.3964]; and for subjective data, it is $\beta$=0.3880 with 95% confidence interval [0.3814, 0.3945]. The confidence interval of the former result contains that of the latter completely. Hence, there is significant evidence that the value of $\beta$ is even shared across objective and subjective data. This shows that the internal homogeneity of size perception is a sound assumption. This value 0.3880 of $\beta$ coincides with the estimation from previous data exploration in Table 1. As to the model with a logarithmic function for $G$, there is no evidence to support it as the preferred simpler model.

The individual variance in task performance is modeled by $a_k$, $b_k$, $d_k$ and $\sigma_k$. Modeling the individual variance aims at a more precise model approximation to the data. Due to the page limit, we are not going to present individual estimates. As a remark, the fluctuation of the estimated values across individual parameters is limited which suggests that a model with shared $H$ and common $a$, $b$ and $d$ might be acceptable. The evidence for such a shared $H$ is presented in the next section.

### Task and Measurement Influences and Estimation of *H*

In our model, $H$ models the performance in the cognitive part of a particular task. Different tasks do not necessarily share the same cognitive process, which means that the response behavior expressed by $b$ and $d$ is expected to vary across tasks. Therefore, it is reasonable to assume task associated values for the free parameters of $H$. Different models with shared $a$, $b$ and $d$ are compared in the gray rows of Table 3. The model ($\beta$, $a_{Ti}$, $b_{Ti}$, $d_{Ti}$, $\sigma$) with shared parameters $\beta$ and $\sigma$ is preferred over the model ($\beta_{Ti}$, $a_{Ti}$, $b_{Ti}$, $d_{Ti}$, $\sigma_{Ti}$) for both the objective and the subjective data. The estimation of $\beta$ by the model ($\beta$, $a_{Ti}$, $b_{Ti}$, $d_{Ti}$, $\sigma$) for objective data is $\beta$=0.3884 with 95% confidence interval [0.3754, 0.4016] and for subjective data is $\beta$=0.3878 with 95% confidence interval [0.3801, 0.3956], both of which largely overlap with the previous results. Figure 8 presents the estimates for $b_{Ti}$ and $d_{Ti}$ in case of the model ($\beta$, $a_{Ti}$, $b_{Ti}$, $d_{Ti}$, $\sigma_{Ti}$).

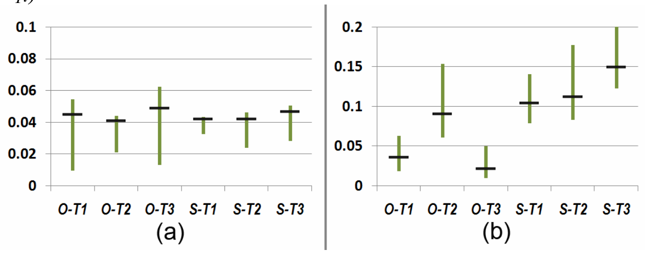


**Figure 8. Parameter values of shared *H* in different tasks and measurements with their 95% confidence intervals: (a) parameter *b*; (b) parameter *d*. *O-Ti*: objective data, *S-Ti*: subjective data.**

There is no significantly different $b_{Ti}$ for different task and measurement conditions since the confidence intervals overlap. The average estimated value of $b_{Ti}$ is 0.0442. For $d_{Ti}$, no significant difference is found for subjective data among different tasks. However, there is significant difference between $T_2$ and the other two tasks for objective data. Particularly, $d_{T2}$ for objective data produces the highest threshold for size contrast perception, and coincides with the thresholds for subjective data. This indicates that in $T_2$ a performance limit was approached in case of low-perceived contrast.

If we compare these models with task associated parameters (in gray cells) with the previous models with individualized parameters (in white cells) in Table 2, we see that the number of parameters drops sharply, but also that the AIC increases strongly. This indicates that individual variation is strong and cannot be neglected. Nevertheless, the analysis of models with task associated parameters gives a direct insight in the influence of the tasks in different measurement.

Moreover, we can observe in Table 2 that the models with $G$=log $r$ also produce a reasonable fit with the data. This elicits the question whether or not we can further simplify $G$ as the logarithmic function. If so, $|G(s_j)–G(s_i)|$ can be written as $\log(s_j/s_i)$. Hence the final output of $H$ is simply determined by the ratio of radii between two circles. Noticing that the Stevens' power coefficient of length judgment is 1, the symbol size discrimination might be simplified to judgment of the ratio of the width or height between two symbols. This hypothesis is investigated in the second experiment.

## EXPERIMENT2

Our experiments and analysis have led to a function $G$ that describes size perception based on size discriminability in the task context. We can use $G$ now to generate sequences of circles with equally perceived steps. We might even extend the perceptual mapping to other types of symbols with equal width and height. However, we do not know how such a perceived scale relates with subjective descriptions of symbols. In other words, we want to know if the internal perceived size scale looks closer to a 1D length scale or a 2D area scale subjectively and if it represents the difference or proportion subjectively. Further, we want to know if the logarithmic scale can represent the perceived scale subjectively or not.

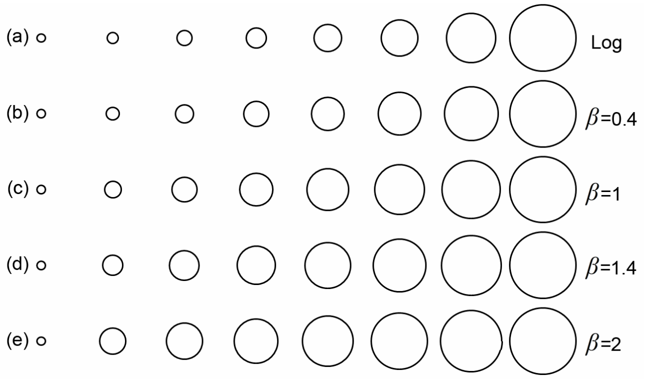
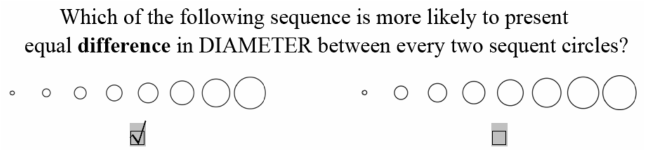


**Figure 9. Sequences of circles configured by linear interpolation on different scales (refer back to Figure 3, we added a discriminability scale obtained from the first experiment with *G* in power form and *β*=0.4).**

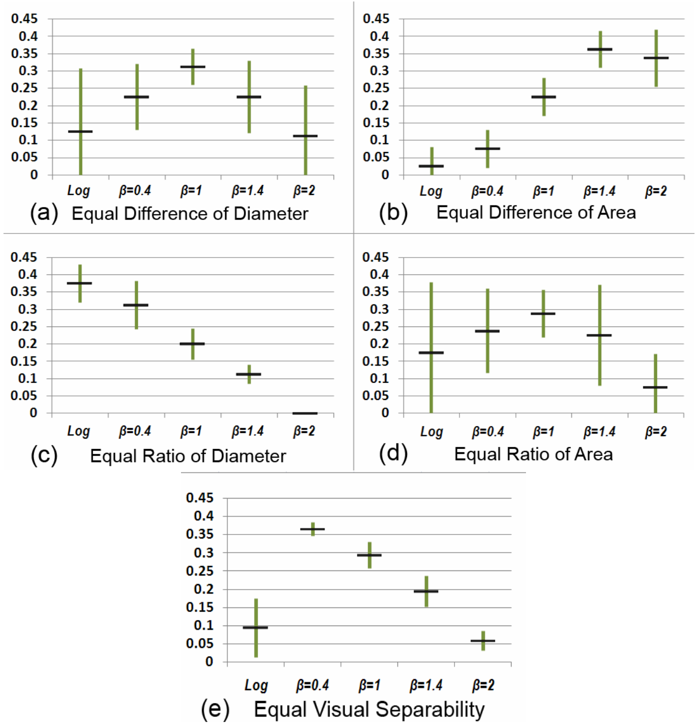
A two-alternative forced choice experiment was designed for the above purpose, in which people were required to compare two of the five sequences of circles shown in Figure 9 with respect to their optimality on five different tasks. The tasks were to distinguish the five relationships: (a) equal differences in diameter; (b) equal differences in area; (c) equal ratios of diameters; (d) equal ratio of areas; (e) equal visual separability. For each task description (an example of task (a) is shown in Figure 10), there were $C_5^2 = 10$ combinations of the five sequences which were doubled by left-right switching. Thus for each subject, there were 10×2×5=100 selections to be made, including all the tasks. We randomized the order between the tasks and the order of selections within a task and tested with 7 subjects.


Which of the following sequence is more likely to present
equal **difference** in DIAMETER between every two sequent circles?

☑ □

**Figure 10. An example of an instruction in Experiment 2.**

For each task, the frequencies of every sequence being selected by each subject were counted and normalized by dividing the total number of selections in one task (20). We estimate the most likely (normalized) frequency of a sequence being selected across all the subjects as well as its 95% confidence interval. In Figure 11, we present the results.



(a) Equal Difference of Diameter

(b) Equal Difference of Area

(c) Equal Ratio of Diameter

(d) Equal Ratio of Area

(e) Equal Visual Separability

**Figure 11. Normalized frequencies for different sequences being selected by people for the five tasks.**

We can see that the sequence with $\beta$=0.4 is significantly preferred over the alternatives in the case of equal visual separability (no overlap of confidence intervals with the second highly selected scale: linear scale). It is also relatively high in the cases of equal difference of diameter, equal ratio of diameter and equal ratio of area. The logarithmic scale is preferred for the task of distinguishing equal ratios of diameter but scores relatively low in the case of equal visual separability. Although the linear scale is preferred in the case of equal difference of diameter and equal ratio of area, the confidence intervals largely overlap with other scales. All of these suggest that in tasks requiring size discrimination, people compare symbol size in a mixed sense and we cannot simplify it to the ratio estimation of lengths.

## CONCLUSION

We found that an optimal scale with respect to equal perceptual separation of symbol size is generated by: $r_i = G^{-1}(P_i) = \alpha \cdot P_i^{1/\beta}$ ($i$=1,…,$N$), with $\beta \approx 0.4$.

Visualization designers can use this as a guideline to encode data. Also, this could be used for other cases where circles have to be discriminated optionally such as button design for user interface or coin design. As shown by the previous study (see Figure 1), this encoding scheme of symbol size should also be relevant when different shapes or colors are used. In Figure 12, we present some size-varying sequences generated by using the $\beta \approx 0.4$ scale with different lightness and shapes. They are very likely to produce equal separation as well.

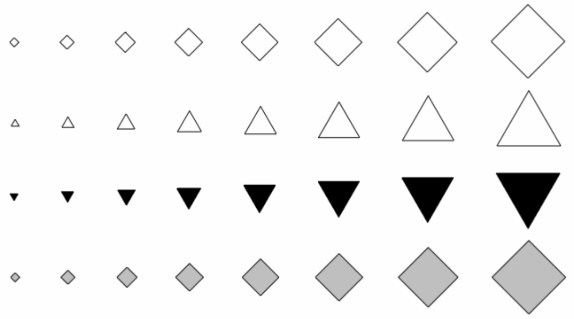


**Figure 12. Tentative sequences of different symbols configured by the discriminability scale obtained from our experiments with *G* in power form and *β*=0.4.**

In order to obtain this optimal scale, we aim at models that depict the internal homogeneity well and handle the individual variance correctly, which is missing in Stevens's method. Individual variance has been modeled and estimated by using individual cognitive functions *H*. No evidence was found to support the more complex model with an individually varied power coefficient for *G*. The estimated values of this power coefficient are also similar for objective and subjective data. Therefore, $\beta \approx 0.4$ is highly supported as the invariant power coefficient for size perception. Furthermore, we used $\beta \approx 0.4$ to produce a scale, and compared it subjectively with four other scales as candidates for the perceptually uniform scale of size. Results show that the scale with $\beta \approx 0.4$ is considered to be the optimal scale for equal visual separation, which indicates that users employ a mixed strategy, in between length and area judgment, to discriminate symbol sizes.

## FUTURE WORK

The next interesting question could be how many different sizes can be used in discrimination tasks. This question can alternatively be formulated as what is the minimum perceived size difference that can support a stable task performance. By setting up appropriate criteria on task performance, we might inversely map to the corresponding configuration via $H^{-1}$ and $G^{-1}$. This could be addressed in the future work. Although this study is limited to a single visual channel, our preliminary results of the ongoing studies show that for instance, size severely influences the perception of lightness, and lightness barely has any influence on size perception. A model of the channel interaction effect could be constructed in future work.

The methodology of modeling presented in this paper can be extended to other visual channels of symbols in scatterplots, for instance the gray scale (lightness), and furthermore to other graphical context such as maps, where symbols are heavily used.

## ACKNOWLEDGMENTS

We thank all the volunteers in different departments of TU/e. The project is supported by the VIEW program of the Netherlands Organization for Scientific Research (NOW) under research grant no. 643.100.502.

## REFERENCES

1. Christ, R.E. Review and analysis of color coding research for visual displays. *Human Factors*, 17(6):542-570, 1975.
2. Stevens, S.S. On the theory of scales of measurement. Science, 103(2684):677-680, 1946.
3. Stevens, S.S. The psychophysics of sensory function. *Amer. Scientist*, 48, 226-253, 1960.
4. Stevens, S.S. and Guirao, M. Subjective scaling of length and area and the matching of length to loudness and brightness. *Journal of Experimental Psychology*, 66, 177-186, 1963.
5. Stevens, S.S. Matching functions between loudness and ten other continua. *Perception and Psychophysics*, 1, 5-8, 1965.
6. Laming, D. *The Measurement of Sensation*. Oxford University Press, NY, USA, 1997.
7. Macmillan, N.A., Moschetto, C.F., Bialostozky, F.M. and Engel, L. Size judgment: The presence of a standard increases the exponent of the power law. *Perception and Psychophysics*, 16(2), 340-346, 1974.
8. Wolfe, J.M. and Cave, K.R. Deploying visual attention: The Guided Search model. *In AI and the Eye*, Troscianko, T. and Blake, A., eds., John Eiley & Sons, UK, 79-103, 1989.
9. Wolfe, J.M. Guided Search 2.0: A revised model of visual search. *Psychomonic Bulletin & Review*, 1(2):202-238, 1994.
10. Li, J., van Wijk, J.J. and Martens, J.B. Evaluation of symbol contrast in scatterplots. *Proc. PacificVis2009*, 97-104, 2009.
11. Miller, A.L., Pedersen, V.M. and Sheldon, R.W. Magnitude estimation of average length: A follow-up. *American Journal of Psychology*, 83, 95-102, 1970.
12. Weiss, D.J. Averaging: An empirical validity criterion for magnitude estimation. *Perception and Psychophysics*, 12, 385-388, 1972.
13. Ariely, D. Seeing sets: representation by statistical properties. *Psychological Science*, 12(2):157-162, 2001.
14. Narens, L. A theory of ratio magnitude estimation. *Journal of Mathematical Psychology*, 40, 109-129, 1996.
15. Ware, C. Information Visualization: Perception for Design, 2nd Edition, Morgan Kaufmann, SF, USA, 2004.
16. Cleveland, W.S. and McGill, R. Graphical perception: theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531-554, 1984.
17. Nowell, L.T. *Graphical Encoding for Information Visualization: using icon color, shape and size to convey nominal and quantitative data*. PhD Dissertation, Dep. of Computer Science, Faculty of the Virginia Polytechnic Institute and State University, Digital Library, 1997.
18. Lewandowsky, S. and Spence, I. Discriminating strata in scatterplots. Journal of the American Statistical Association, 84(407):682-699, 1989.
19. Tremmel, L. The visual separability of plotting symbols in scatterplots. Journal of Computational and Graphical Statistics, 4(2):101-112, 1995.
20. Amar, R., Eagan, J. and Stasko, J.Low-level components of analytic activity in Information Visualization. *Proc. IEEE InfoVis*: 112-119, 2005.
21. Martens, J.B. *Image Technology Design: A Perceptual Approach*. Kluwer Academic Publishers: Dordrecht, NL, 2003.
22. Green, P.E., Carmone Jr. F.J. and Smith, S.M. *Multidimensional Scaling, concepts and Applications*. Allyn & Bacon, Massachusetts, USA, 1989.
23. Uusipaikka, E. *Confidence Intervals in Generalized Regression Models*. STATISTICS: Textbooks and Monographs. CRC Press: Finland, 2009.
24. Burnham, K.P. and Anderson, D.R. Multimodel Inference – Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, 33(2):261-304, 2004.