Computer Science Journal of Moldova, vol.**25**, no.3(75), 2017

A linear model for multidimensional Big Data visualization

Vadim Grinshpun

Abstract

The author introduces and analyzes a model that allows organizing visualization of primary linear constructs such as interval, simplex and polygonal lines in multidimensional space.

Keywords: computer science, big data, data visualization, multi-dimensional data, exploratory data research.

1 Introduction

There are a number of well-known methods to visualize multidimensional data. There are Andrews plots, Bergeron's or Wong's model, Zinoviev model as well as Klaft, Barrett and Kleiner-Hartigan, and every one of them introduces their own unique mechanism for data visualization [1]. However, every method has its own limitations, narrowing the field of direct applicability. For instance, the Bergeron's model visualizes the wave lines and the time interval for a single frequency [2].

Graphical methods are especially helpful during the Exploratory Data Research (EDR) of the large sets of multi-dimensional data and the clustering problems, enabling the analyst to discover patterns and relationships hidden in the data set. The main advantage of the modeling data as a multi-dimensional set of points or observations is the convenience and effectiveness of analyzing a big volume of data, particularly when applied to a time-series. The problem with such model is its bulkiness and poor suitability for simple tasks of the operational data processing [3].

2 Model Definition

Theorem 1. *As a basis for visualization of the multidimensional data a linear modification of a multidimensional observation $\mathcal{H}$ into two-dimensional curved line $\mathfrak{L}_{\mathcal{H}}(t)$ is used, so $\mathcal{H}$ approximates $\mathfrak{L}_{\mathcal{H}}(t) : \mathcal{H} \leftrightarrow \mathfrak{L}_{\mathcal{H}}(t)$, with the provable condition that the values of the dimensional attributes of observations $\mathcal{H}$ and $\mathcal{X}$ correspond to graphics $\mathfrak{L}_{\mathcal{H}}(t)$ and $\mathfrak{L}_{\mathcal{X}}(t)$ that visually appear near each other, reflecting the relative closeness of $\mathcal{H}$ and $\mathcal{X}$. Conversely when these values are relatively distant, the graphical lines will appear to be far apart.*

Proof Theorem 1. For the analysis of the proposed method we will use the most general system of data presentation. Let's pick a vector H in Pn – a space with finite number of dimensions.

$$H = (h0, h1, h2, h3, \ldots hn - 1) \in Pn. \tag{1}$$

To create the visualization of the vector we have to create a basis for transformation as a set of orthogonal functions $\{\varphi(t)\} \to \infty$. Legendre orthogonal polynomials can be applied on a 0 to 1 interval, set of which can be shown as $\zeta(t) \to \infty$. In this case the vector H with coordinates $(h0, h1, h2, h3, \ldots hn - 1) \in Pn$ corresponds to the following function:

$$E_{H(t)} = \sum_{i=0}^{n-1} H_i L_i(t). \tag{2}$$

Conversion of the vector H is accomplished by conversion of its multidimensional data. In order to characterize the observable multidimensional object its coordinate values play a significant role. In the extreme cases, each coordinate should have its own measurement defined, and its value should affect the appearance of the $Eh\,(t)$ function. To exclude the influence of the individual measurement types over the $Eh\,(t)$ function, it is necessary to switch to a neutral set of values, by using one of the known methods.

It should be noted that the of inclusion of the dimensional values in vector H also can influence the look of the $Eh\,(t)$ function. To justify the order of inclusion of these characteristics in certain applications, an expertise determining the “informativeness” (the degree of influence) of each individual one can be performed, accompanied by an analysis of the optimal sequence of inclusion of these characteristics into the vector H [4].

Let’s introduce a second vector into the model:

$$X = (x0, x1, x2, x3, \dots xn - 1) \in Pn \tag{3}$$

and its corresponding function:

$$Ex(t) = \sum_{i=0}^{n-1} H_i L_i(t). \tag{4}$$

And now we can transform two points $H\&X$ from the Pn space, into the graphical view of their representative functions $Eh(t)$ and $Ex(t)$ (Fig.1).

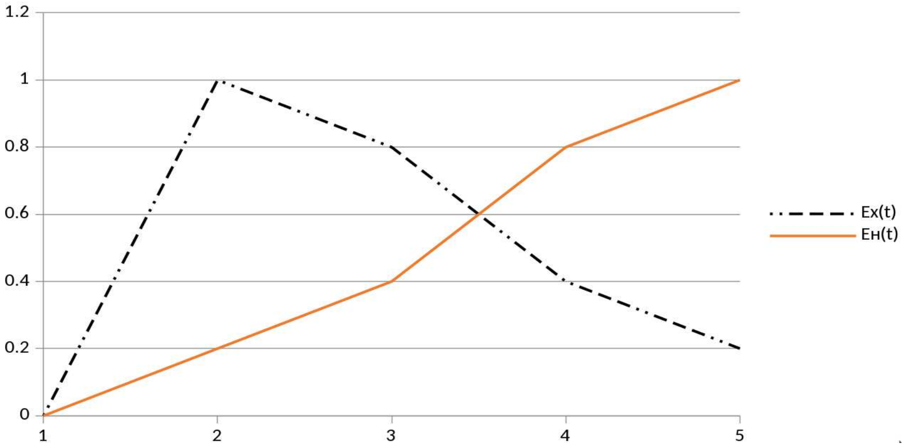


Figure 1. Visualization of $H\&X$ from the Pn space.

When we consider $H\&X$ to be vectors, with the beginning located at the beginning of the coordinate system selected for the Pn space

– then the relative proximity between all points in the Pn space becomes definitively tied to the graphical representations of their corresponding $Eh(t)$ and $Ex(t)$ functions, with axes values defined as $h0, h1, h2, h3, \ldots hn - 1$. By introducing a variable, we can create an equation:

$$\begin{aligned} C(C) = (1-c)H + cX = ((1-c)h0 + cx0, (1-c)h1 + cx1, \\ \ldots (1-c)hn1 + cxn1. \end{aligned} \tag{5}$$

From which obviously follows $C(0) = H$ and $C(1) = X$, which can be viewed as a definition of a multidimensional "straight" line connecting $H\&X$ in the Pn space, and we can use the expression like (5) to represent a multidimensional segment HX:

$$HX = (1-c)H + cX, \text{ where } \in [0,1]. \tag{6}$$

Assuming "c" represents the distance in the Pn space, the equation (6) can be shown as the proposed model:

$$E_{HX}(c) = \sum_{i=0}^{n-1} (1-c)H_i L_i(t) + cx_i L_i(t). \tag{7}$$

This function has two arguments $\{c, t\}$, which allows us to get a graphical function $Ehx(c) = Ehx\{c, t\}$ that visually represents the HX segment, as shown in Figure 2.

When defined over the $[0,1]/[0,1]$ square, it is possible to produce a smooth surface based on (7), that corresponds to an analytical expression (6) that represents a multidimensional segment HX [5] **QED**.

3 Sample Application

To test the model, we will apply it to a sample set of multidimensional objects with the following values: $H1 = \{1,0,0,0\}, H2 = \{0,1,0,0\}, H3 = \{0,0,1,0\}, H4 = \{0,0,0,1\}$, and transform them

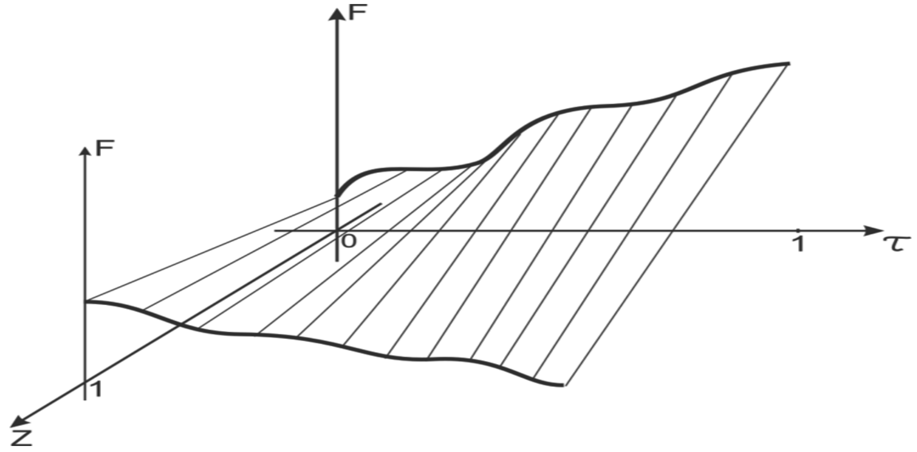


Figure 2. Visualization of smooth surface, corresponding to the HX segment from the Pn space

using polynomial matrices

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \to 1 * l0(t) + 0 * l0(t) + 0 * l2(t) + 0 * l3(t)$$
$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \to 1 * l0(t) + 0 * l0(t) + 0 * l2(t) + 0 * l3(t)$$
$$\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \to 1 * l0(t) + 0 * l0(t) + 0 * l2(t) + 0 * l3(t)$$
$$\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \to 1 * l0(t) + 0 * l0(t) + 0 * l2(t) + 0 * l3(t)$$

and get the general formula

$$E = \begin{bmatrix} f0 & f1 \\ f2 & f3 \end{bmatrix} \to f0 * l0(t) + f1 * l0(t) + f2 * l2(t) + f3 * l3(t).$$

The polynomial argument $\{t\}$ is the characterization of the composite representation and has a value, but no measure. Vector E cannot be shown in 3D, and therefore it is being substituted with a 2D line $E1(t)$ [6].

Let's see what the graphic would look like for the following 10-dimensional values.

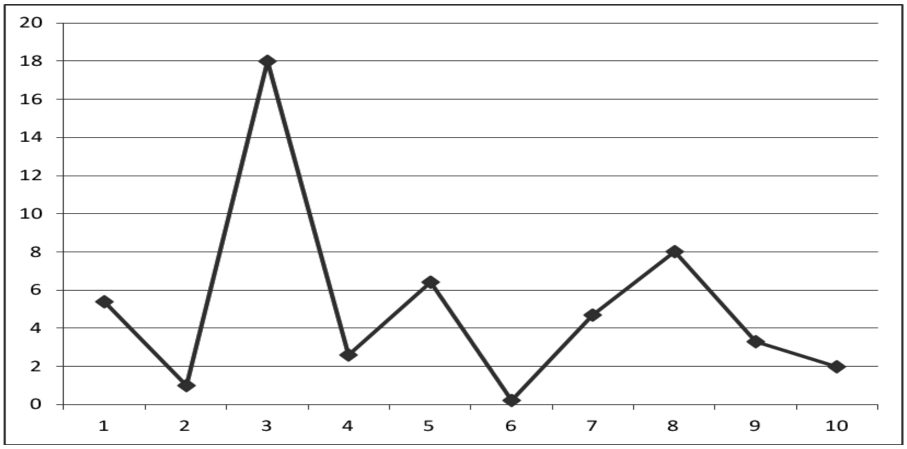


Figure 3. Visualization of $H : 54, 1, 18, 2.6, 6.4, 0.2, 4.7, 8, 3.3, 2$ in the Pn space

When these two graphics are joined, it becomes very clear that they not only look very similar, but are very close to each other in the dimensional points, indicating that the original raw observations are in close proximity in the Pn space as shown in Figure 5.

The more indistinguishable are the graphical representations of the raw observations, the closer to each other are these observations in their original space, as the multidimensional points are bijected into their corresponding graphics.

It is possible to visualize many interesting characteristics, by reproducing the Figure 5 graphics in 3D, by defining a Z-order as the distance between the points in the Pn-space or the time-interval between the observations. The Pn-space distance or the time interval can be measured in any applicable way and scaled to fit the relative distance into the graphic, which makes it possible not only evaluate

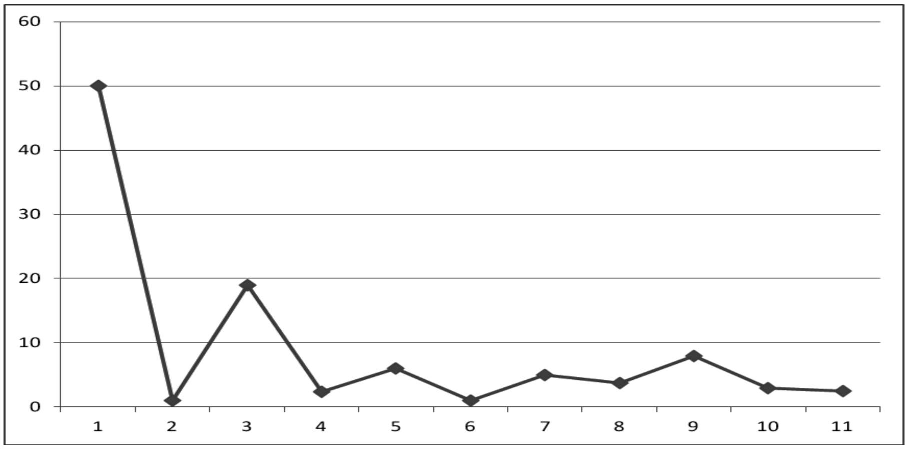


Figure 4. Visualization of $X : 50, 1, 19, 2.4, 6, 1, 5, 3.8, 8, 3, 2.5$ in the Pn space

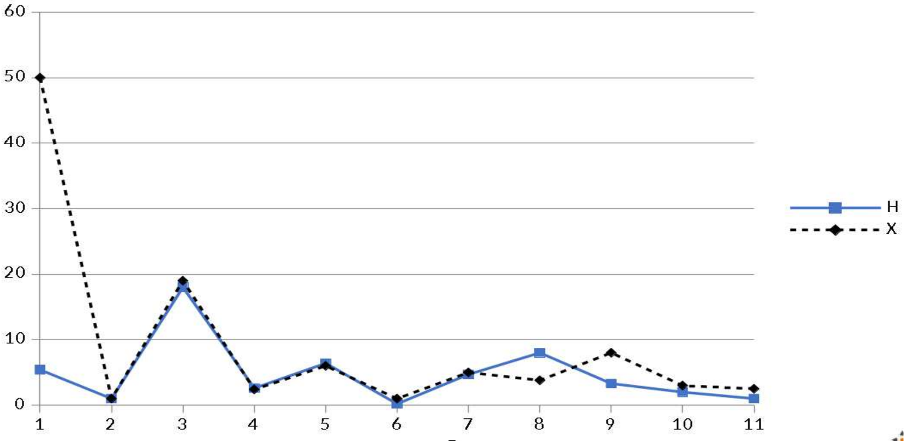


Figure 5. Proximity visualization of $H\&X$ from the Pn space.

the static characteristics of the observation data, but to view some of the changes dynamically [7].

4 Model Optimization

In linear transformation of $H \leftrightarrow F_H(t)$, using the segment between the multidimensional observations $H\&X$, we obtain a corresponding surface, that ties the projected observations. Every line representing observations with intermediate values (observations that belong to the $[H;X]$ segment in Pn space) will appear on that surface [6]. Let's consider Figure 6.

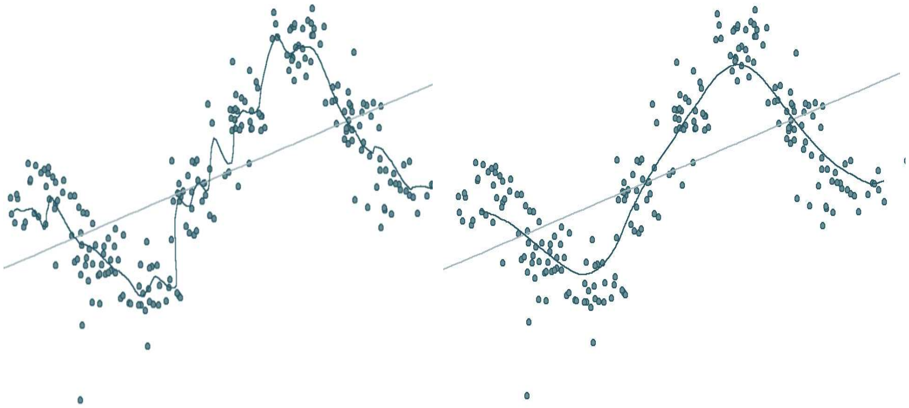


Figure 6. Proximity visualization of $H\&X$ from the Pn space.

In order to compare the observations in greater detail, particularly in the case of the heterogeneity of the units of measurement of the various data characteristics, a traditional modification mechanism should be applied:

- *Normalization* – to be applied to express the results in a coherent system of measurements;

- *Standardization* – to be applied to enable comparison of the data characteristics with variant attribute distributions and/or different units of measure [8];

Currently there are numerous programmatic and algorithmic visualization tools for multidimensional data structures. However, quite often basic visualization technique cannot be directly applied to a task at hand, since the researchers are usually interested in some very specific properties of the data that cannot be identified using standard approaches. Cases like these, call for development of the specialized types of presentation, focusing on the specific requirements of task at hand.

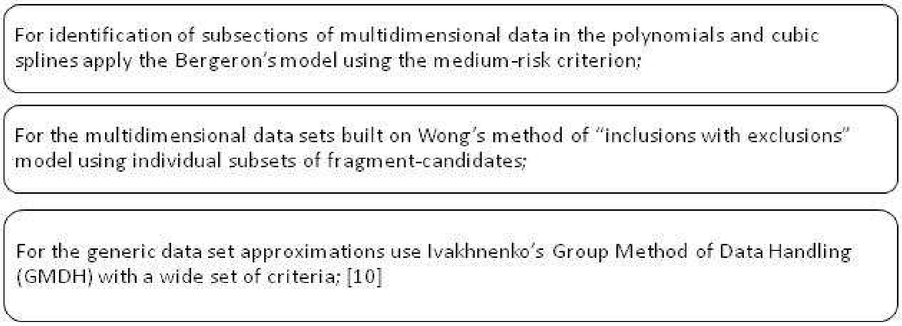
So, the developed model (6) of multidimensional data visualization demonstrates that the proposed approach holds promise in the area of analysis and representation of raw multidimensional data.

The particulars in the model selection process and the characteristics of its generation depend on the specific expectations of the outcome of the research. To formulate the task, it would be more reasonable to produce a limited and compact model. For example, an exponential function or a spline with two or three junction points. If we consider a forecasting model, there are no severe restrictions on the function's look and feel, with the one and only requirement being the authenticity and immutability of the prediction, when extrapolating the multidimensional data. However, the same principle of searching for the best model by the way of self-organization applies in all cases.

The essence of the development process of the model of optimal complexity by means of model's self-organizing is contained in its gradual organizational identification, i.e. setting of the model's optimal structure and isochronous analysis of its characteristics. In the cases like that, the specific sets of models of varied complexity are generated, and the best of them are identified based on a given rational indicator of regularization.

Figure 7 shows the selection of the criteria for the optimization of multidimensional data [9].

An important question in the research of multidimensional data visualization is the analysis of the relationships between data point's



For identification of subsections of multidimensional data in the polynomials and cubic splines apply the Bergeron's model using the medium-risk criterion;

For the multidimensional data sets built on Wong's method of "inclusions with exclusions" model using individual subsets of fragment-candidates;

For the generic data set approximations use Ivakhnenko's Group Method of Data Handling (GMDH) with a wide set of criteria; [10]

Figure 7. Selection of models for development of multidimensional data set optimizations for application in visualization techniques

individual characteristics and its effect on the overall informational content of the data set.

When considering the Bergeron and Wong models (Figure 7) it is important to note that when applied they produce the same effect on a given data-set. Because of it the following statement can be referred to as the "Bergeron-Wong" theorem.

Theorem 2. *Full linear big data model, used for the synthesis of the selected multidimensional predictive collections is dependent on the processes characterized by the multicollinearity, neural networks and robust statistical methods applied against large volumes of data for targeted selection of the most informationally significant data attributes.*

Proof Theorem 2. The specifics of the Bergeron Model lie in its applicability to the large volumes of data calculation methods with multicollinearity. The model looks as follows:

$$\begin{gathered}\Delta v = \frac{a \cdot *vm}{g} \\ a \cdot = ab + Dh \\ \Delta v = \frac{ab \cdot *vm}{g} \\ b \cdot = -(ab + Dh) \end{gathered} \tag{8}$$

This model demonstrates that the visualization problems of information spaces may be resolved by application of the self-organizing modeling described in Figure 7: through various transformations with targeted selection of the informational value of the attributes, breeding records' observed indicators, neural networks and robust statistical methods applied against large volumes of data for targeted selection of the most informationally significant data attributes [2].

Wong's model corresponds to the same representation however it works from the large data sets to the minimalistic ones [11]. The Wong's model can be represented as follows:

$$\Delta = -gm * \Delta t = \frac{-gt - \Delta r + gt}{2} \Delta t. \tag{9}$$

The formulae for Bergeron's and Wong's models shown above represent a complete linear spatial model for the large sets of data, used for the synthesis of the selected multidimensional predictive collections. Thus the self-organizing approach based on the above statements allows constructing methods and models developed for a specific set of large multidimensional observations. Also it is representable as a collection of the high-level polynomials, through the application of the model connection technique [2].

$$\begin{gathered} \Delta = a_0 + \sum_{i=1}^{m} a_i g_i + \\ \sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j g_i g_j + \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{m} a_i a_j a_k g_i g_j g_k + m \ldots \end{gathered} \tag{10}$$

The key issue in such complex structures based on the very large data sets is to cull the (9) by removing the low-information data attributes, that prove largely irrelevant and to leave a necessary and sufficient number of most meaningful attributes.

The complexity of the model being analyzed is considered optimal when the model remains adequate for the stated purpose with the fewest number of attributes used to comprise the transformed model [7]. Let's illustrate this statement with the Figure 8

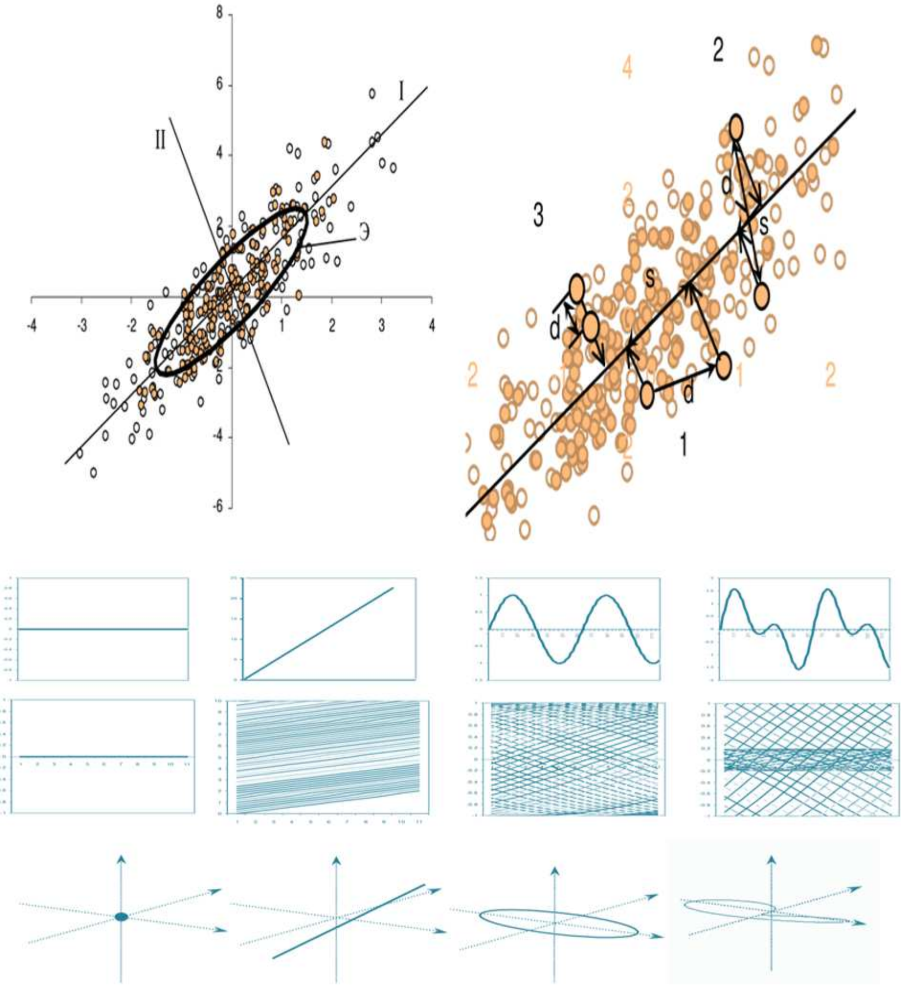


Figure 8. Graphical representation of the combined model

5 Conclusion

In conclusion, it is important to point out that the data visualization methods are necessary to answer the data interpretation questions, however to determine all the characteristics of the data, manage its size and complexity and eventually reduce the time it takes to develop the answers, these visualizations methods must be preceded and often augmented by the non-visual means of analysis and interpretation of a given set of data.

The more general outcome of this analysis is a methodology for accounting of the group's characteristics, that can be viewed as the linkage between a variety of methodological concepts and can be depicted through modern methods of artificial intelligence [12]. In continuation of this research an original program module will be developed. That model shall implement a generalized multi-faceted algorithm for visualizations of big data with its attributes described and optimized through a set of linear functions, implementing the reflecting "distances" between the attributes at every step of the subset selection.

References

[1] K.A. Sharopin, O.G. Berestneva and G.I. Shkatova, "Visualization of the results of the experimental research," *Tomsk Politechnical University News,* vol. 316, pp. 172–176, 2010. (in Russian)

[2] R. Venet, P. Leger and A. Pavie, *La méthode graphique de bergeron pour étayer l'hypothèse sur l'origine de l'hémorragie cérébrale du sujet hypertendu : le coup de bélier hydraulique á l'origine d'une cavitation engendrant un phénomène de thixotropie. le miracle de la saint janvier.* 2013. [Online]. Available: hal.archives-ouvertes.fr/hal-00830952/file/BERGERON_HAL.pdf. (in French)

[3] *Principal Manifolds for Data Visualization and Dimension Reduction,* (Lecture Notes in Computational Science and Engineering, vol. 58), Alexander N. Gorban, Balzs Kgl, Donald C. Wunsch,

Andrei Zinovyev, Eds. Springer-Verlag Berlin Heidelberg, 2008, XXIV+340 p.

[4] O.V. Marukhina, O.G. Berestneva, V.A. Volovodenko and K.A. Sharopin, "Technologies for visualization of the experimental research results," in *Records of XVI Bajkal Conference; Part 3*, (Irkutsk), 2010, pp. 165–171. (in Russian)

[5] S. Rufiange, "Visualizing Dynamic Graphs with a Hybrid of Difference Maps and Animation," *IEEE Transactions on Visualization and Computer Graphics,* vol. 19, no. 12, pp. 2556–2565, 2013.

[6] A.N. Gorban, D.A. Rossiev, E.V. Butakova, S.E. Gilev, S.E. Golovenkin, S.A. Dogadin, M.A. Dorrer, D.A. Kochenov, A.G. Kopytov, E.V. Maslennikova, G.V. Matyushin, Ye.M. Mirkes, B.V. Nazarov, K.G. Nozdrachev, A.A. Savchenko, S.V. Smirnova, V.A. Shulman and V.I. Zenkin, "Medical, psychological and physiological applications of MultiNeuron neural simulator," in *The Second International Symposium on Neuroinformatics and Neurocomputers,* (Rostov on Don), 1995, pp. 7–14.

[7] O.G. Berestneva, Y.S. Pekker, K.A. Sharopin and V.A. Volovodenko, "Discovery of hidden patterns in the medical and socio-psychological research," in *Proceedings of the 2 International Conference on Computing Application Systems*, (Moscow), 2010, pp. 287–296. (in Russian)

[8] O.V. Marukhina, O.G. Berestneva, K.A. Sharopin and I.A. Osadchaya, "Cognitive graphics in socio-psychological research," in *Records of XVII Bajkal Conference; Part 3*, (Irkutsk), 2011, pp. 176–181. (in Russian)

[9] J.Nikander, "Exploratory vs. Model-Based Mobility Analysis," *Nordic Journal of Surveying and Real Estate Research,* vol. 9, no. 1, pp. 7–29, 2012.

[10] F.R. Hampel, E. Ronchetti, P.J. Rousseeuw and W.A. Stahel, *Robust statistic: the approach based on influence functions,* 1986.

[11] D.J. Walker, J.E. Fieldsend and R.M. Everson, "Visualising many-objective populations," in *Proc. of the 14th annual conference companion on Genetic and evolutionary computation (GECCO '12)*, (New York), 2012, pp. 451–458.

[12] G. Jornod, E. Di Mario, I. Navarro and A. Martinoli, "SwarmViz: An open-source visualization tool for Particle Swarm Optimization," in *IEEE Congress on Evolutionary Computation (CEC)*, (Sendai, Japan), 2015, pp. 179–186.

Vadim Grinshpun

Received September 25, 2017

Institute of Mathematics and Computer Science
Academy of Sciences of Moldova
5 Academiei str., Chişinău, MD-2028, Moldova
E–mail: `vgrinshpun@hotmail.com`