ZESZYTY NAUKOWE UNIWERSYTETU SZCZECIŃSKIEGO

NR 789 STUDIA INFORMATICA NR 32 2013

Olga Pilipczuk*
Szczecin University

Dmitri Eidenzon**
NovoSpark Corporation, Kanada

MULTIDIMENSIONAL DATA VISUALIZATION WITH THE NOVOSPARK® VISUALIZER SOFTWARE

Summary

In this paper the authors present a method for visualization and qualitative analysis of multivariate data implemented in the NovoSpark® Visualizer software system. An application example, based on solar activity data, is discussed as well. Selected traditional methods are compared with NovoSpark method. The results of experiment prove that traditional methods of multidimensional data visualization (such as linear plots and parallel coordinates) lack the ability to simultaneously display all dimension values, static or dynamic, in a clear single image.

Keywords: multidimensional data, visualization, NovoSpark@ Visualizer

Introduction

In the process of decision-making often appears a need to analyze multidimensional data. This usually occurs when very complex reality is described and it can be analyzed in many aspects and dimensions or when many variables are used. During the analysis and visualization of such a reality, every aspect, space or variable is seen as a separate dimension of a multidimensional space. Because multidimensional data analysis is a very complicated process, a number of methods have been created to analyze it in a quantitative manner, but

* olga.pilipczuk@wneiz.pl
** dmitri.eidenzon@novospark.com

all of these methods are scattered. The variety of these methods is induced by the diversity of the phenomena. However, in many situations it is necessary to have the possibility of visual, qualitative assessment of the analyzed data structure. Most natural source of information that can be used for the qualitative observation is human eyesight. For this to work it is necessary to be able to do observation visualization in multidimensional spaces.

Often, it is indispensable to obtain a visualization of the entire dataset as one integral image. Such a possibility is implemented in the NovoSpark Visualizer system that by acting on the basis of a mathematical model solves the problem of multidimensional data integration.

1. Traditional methods of multidimensional data visualization

Multidimensional data analysis methods are widely used in the research of complex phenomena, the description of which needs more than one variable.

The simplest methods of multidimensional data visualization are the methods of projection on two coordinate axes. They allow you to find the relationships between features (coordinates). This group of methods includes scattered plots, in which data is presented as a set of points, the position of which depends on the value of the features reflected on the coordinate axes. This method also allows finding redundant features and clustering. Another method out of this group is two-dimensional histograms presented in three dimensions.

Radar charts is another group of methods that allows presenting many features of an object or phenomenon in one image. This chart resembles a star. Each feature of the chart is shown as its radius. This type of diagrams can be effectively used to compare sets of objects with the same characteristics. A limitation of this chart is the lack of transparency in the analysis in case of a large number of features.

In the method of parallel coordinates the features are mapped to each other by using parallel coordinate axes, with the values of these characteristics reflected on them (Inselberg 1985, Inselberg 2009). Restrictions for using this method are the same as in the previous one, i.e, in the case of a large number of visual characteristics chart loses its clarity and the analysis becomes impossible (Few 2006).

In conclusion, we can say that the methods of multidimensional data visualization listed above have one common flaw: suitability to represent a large number of features. When performing samples mapping of a complex phenomena the quality of information decreases. Thus a number of methods aiming to reduce the dimensionality were created. The goal of dimensionality reduction is to transform the multidimensional data into a smaller extent. Most data is reduced to two or three dimensions.

A very common technique for reducing data dimensionality is principal component analysis (PCA). It finds the data subspaces concentrated in the main space of a coordinate system (Jolliffe 2002). Size reduction is accomplished via removal of the small variance of the characteristics. The disadvantage of this method is that it is unsuitable for the analysis of a non-linear data structure. Limitation of this method is that it is based only on the quantitative data and cannot be applied in the case when the only available data is information about objects similarity. The multidimensional scaling method fills this gap by creating an objects dissimilarity matrix.

Another method of transforming multidimensional data into two-dimensional space is Kohenen Network. This network presents the data in the form of self-organizing map (SOM) units (neurons) and has the shape of a two-dimensional grid. It is very useful for detection and visualization of clusters in a data set (Kohonen 2001). The advantage of this method is its parameterization (Graepel 1998). It does not require any assumptions on the initial distribution of the variables analyzed.

Voronoi diagram consists of cells (fields), which have only one entry point inside (Reitsma, Trubin 2007). The edge of the diagram is equidistant from two adjacent input points. It means that the dividing line between the two points located midway between them. However, this diagram has a number of drawbacks. First, using this method, we have difficulties in considering dynamic changes. The second drawback is the lack of tolerance for sensory faults.

However, dimensionality reduction is not always the best solution. Sometimes there is a need of data visualization without their transformation to the two-dimensional space. A good way out of this situation is the use of color images or charts, which utilize natural human ability to recognize shades of colors. As an example of such a method Fortson rectangles can be used where the characteristic value is determined by the corresponding tone of gray. Also, some new methods that specify the characteristics of the color using the entire spectrum of colors recently appeared.

A very effective method of multi-dimensional data visualization is Chernoff Faces. “Face” is a picture that reproduces each observation separately (Chernoff

1973). Individual part of the face attributed to the values of the tested variables. The method enables visualization of about 15 separate features on a single image. This number can be increased by dividing each face into two halves: the left and right. The analysis of such images, however, poses many problems, since it is hard to compare too many features of the faces.

In summary, we can replace the main disadvantages and difficulties of traditional visualization methods (Dzemyda at al. 2013; Gemignani 2010; Kandogan 2001; Sun at al. 2008):

– determination of dimensionality,
– different scales of parameters/variables,
– detection of data trends,
– ambiguity in mapping of dynamic changes in parameter/variable values.

We can also conclude that the visualization of the integral system state has to use all the state parameter values to create an image without loss of information and produce a single static image for state dynamics.

Visualization satisfying the requirements presented above is proposed in the NovoSpark ® Visualizer system (www.novospark.com).

2. The conception and analytical possibilities of the NovoSpark Visualizer system

In 1991 Volovodenco, Eidenzon and Mylcev have proposed a new approach that allowed visualization of both static and dynamic data on one integral image. The proposed methodology has been adapted to use modern computer technologies that resulted in implementation of the NovoSpark® Visualizer tool (Eidenzon, Volovodenco 2009).

The NovoSpark® Visualizer is used for the analysis and visualization of multidimensional data. This tool provides possibility to create two types of data images: “integral” images containing information about all parameters and “traditional” images. The available integral images are the “NovoSpark Curves”, “Andrews Curves” (Andrews 1972) and “Parallel Coordinates”. The available traditional images are “Linear Plots”, “Multiple Linear Plots”, “Scatter Plot Matrix”, “Polar Coordinates” and “Histograms”. The system creates the possibility of carrying out a series of traditional analyses: factor analysis, cluster analysis, regression analysis, Kohonen self-organizing maps, etc. The set of options to manipulate the image allows performing various views and image transformations, creating a multidimensional interval cloud and marking of abnormal observations, applying a color palette, selecting data subsets for display, and so on.

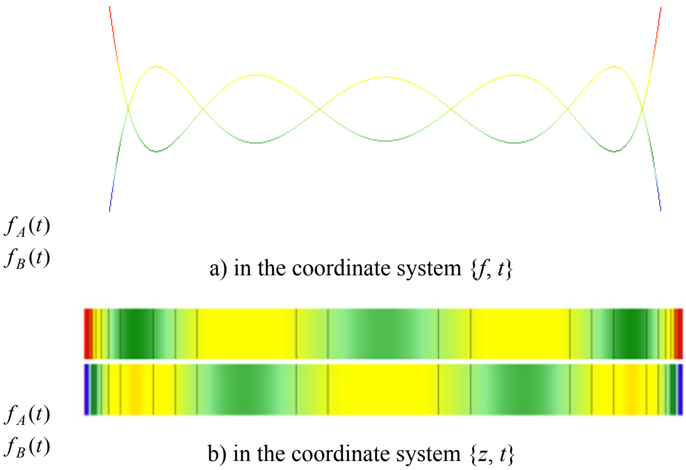
The method of visualizing multidimensional objects and processes is based on two isometric spaces: objects from one space are called the originals*,* while objects from the other space play the role of images.

A selected point-observation $\boldsymbol{A}$ in N-dimensional affine point-vector space R_N of the originals $A = (a_0, a_1, \ldots, a_{N-1})$ and form linear combination $f_A(t)$ of functions $\{P_i(t)\}_0^\infty$ by using the following equation (Figure 1a):

$$f_A(t) = \sum_{i=0}^{N-1} a_i P_i(t) \tag{1}$$

where $P_i(t)$ are Legendre polynomials, i.e. orthogonal polynomials with weight 1 defined on the segment $t = [0, 1]$.

The image of a point in a multidimensional space is presented as a function-curve and can be "painted" in accordance with the function values (Figure 1b). Applying a color palette emphasizes similarities and/or differences in images and allows viewing these images in the coordinate system $\{z, t\}$. Such an image of $f_A(t)$ function is called a "spectrum" of the multidimensional point-observation.



a) in the coordinate system $\{f, t\}$

b) in the coordinate system $\{z, t\}$

Figure 1. Images of multidimensional observations A and B

Source: prepared by the authors.

A multidimensional process is considered as a set of sequential transitions from one multidimensional observation (state) to another; or as a set of multidimensional segments sequentially connecting system states – points in the multidimensional space of the originals.

A multidimensional segment *AB* with fixed vertices, where one can define the radius vector for any point *X* belonging to the segment *AB*: $p_X = vector\ (x_0, x_1, ..., x_{N-1})$ satisfies the following equation (Figure 2):

$$p_X = p_A + z(p_B - p_A) = p_X = vector(x_0(z), x_1(z), ..., x_{N-1}(z)) \tag{2}$$

where $z \in [0, 1]$. The point *X* at a position z could be described using the following equation:

$$X = X(z) \leftrightarrow f_X(t) = \sum_{i=0}^{N-1} x_i(z) P_i(t) = f_X(t, z) \tag{3}$$

where $x_i(z) = a_i + z(b_i - a_i)$.

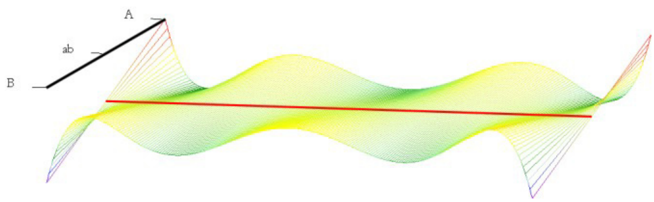


Figure 2. The image of a multidimensional segment *AB* with the point *ab* = (0, 0, 0, 0, 0, 0, 0) in the coordinate system {*f*, *t*, *z*}

Source: prepared by the authors.

The image of a multidimensional interval is a two-dimensional region between the “minimum” and “maximum” images. It is rendered in the coordinate system {*f*, *t*} and is called a “cloud” of a multidimensional interval. Boundaries of this cloud are obtained from a linear combination of separate images for each parameter from the coordinate space of the originals**.**

A more detailed description of the method has been presented in previous works (Eidenzon, Volovodenco 2009; Pilipczuk, Eidenzon 2013; Eidenzon et al. 2013).

3. An example of applying the NovoSpark® Visualizer system in the research of solar activity

In order to demonstrate the system functionality analysis of selected solar activity indicators has been conducted on the basis of data presented in Table 1 (Pilipczuk 2013).

Table 1

Monthly observations of sunspots appearance in the years 1995–2011

Year	Month											
	1	2	3	4	5	6	7	8	9	10	11	12
1995	24.2	29.9	31.1	14.0	14.5	15.6	14.5	14.3	11.8	21.1	9.0	10.0
1996	11.5	4.4	9.2	4.8	5.5	11.8	8.2	14.4	1.6	0.9	17.9	13.3
1997	5.7	7.6	8.7	15.5	18.5	12.7	10.4	24.4	51.3	22.8	39.0	41.2
1998	31.9	40.3	54.8	53.4	56.3	70.7	66.6	92.2	92.9	55.5	74.0	81.9
1999	62.0	66.3	68.8	63.7	106.4	137.7	113.5	93.7	71.5	116.7	133.2	84.6
2000	90.1	112.9	138.5	125.5	121.6	125.5	170.1	130.5	109.7	99.4	106.8	104.4
2001	95.6	80.6	113.5	107.7	96.6	134.0	81.8	106.4	150.7	125.5	106.5	132.2
2002	114.1	107.4	98.4	120.7	120.8	88.3	99.6	116.4	109.6	97.5	95.0	81.6
2003	79.5	46.2	61.5	60.0	55.2	77.4	85.0	72.7	48.8	65.6	67.2	47.0
2004	37.2	46.0	48.9	39.3	41.5	43.2	51.0	40.9	27.7	48.4	43.7	17.9
2005	31.3	29.2	24.5	24.4	42.6	39.6	39.9	36.4	22.1	8.5	18.0	41.2
2006	15.4	5.0	10.8	30.2	22.2	13.9	12.2	12.9	14.5	10.4	21.5	13.6
2007	16.9	10.6	4.8	3.7	11.7	12.0	10.0	6.2	2.4	0.9	1.7	10.1
2008	3.4	2.1	9.3	2.9	2.9	3.1	0.5	0.5	1.1	2.9	4.1	0.8
2009	1.5	1.4	0.7	1.2	2.9	2.6	3.5	0.0	4.2	4.6	4.2	10.6
2010	13.1	18.6	15.4	7.9	8.8	13.5	16.1	19.6	25.2	23.5	21.6	14.5
2011	19.0	29.4	56.2	5.4	41.6	37.0	43.9	50.6	78.0	88.0	96.7	73.0

Source: Australian Government Bureau of Meteorology www.bom.gov.au.

The system not only allows you creating a three-dimensional image (Figure 3), but also seeing the dynamics of sunspot observations in the years from 1995 (0) to 2011 (16). The violet and blue color indicates the points there the parameters values are below the minimum, the orange and red colors indicate the values which are above the maximum.

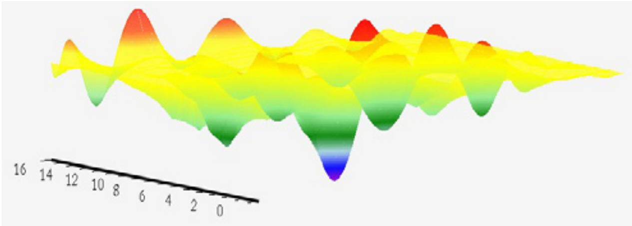
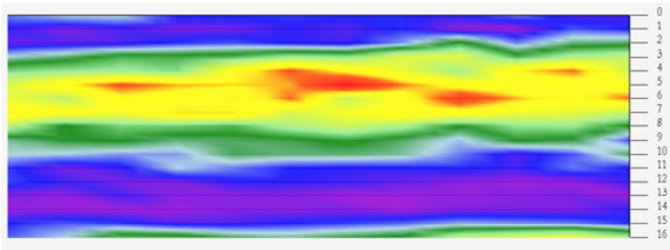


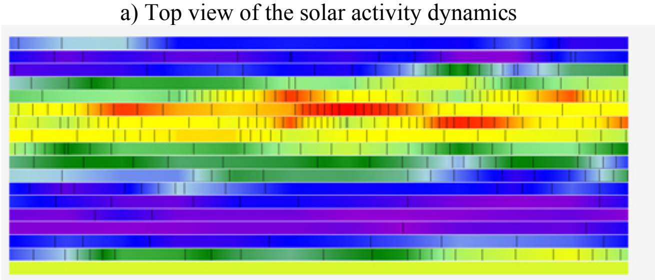
Figure 3. Three-dimensional chart of sunspot observations in 1995–2011

Source: prepared by the authors.

Figure 4 shows the dynamics of sunspots. The numerical values have been transformed into spectrum of colors. System allows building images on the basis of parallel coordinates.



a) Top view of the solar activity dynamics

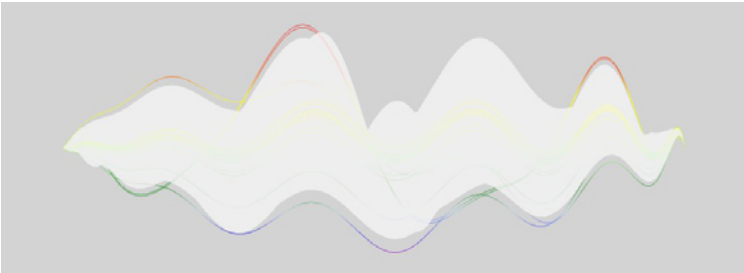


b) Spectrum view of the solar activity dynamics

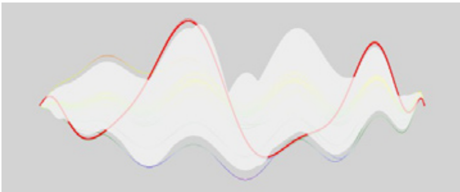
Figure 4. The presentation of the sunspot dynamics using parallel coordinates

Source: prepared by the authors.

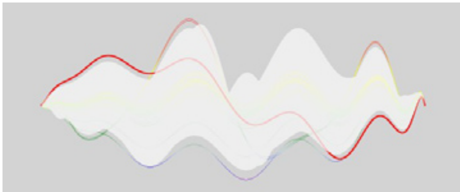
Another very useful ability of the system is identification of anomalies in the data sets. Figure 5 presents the chart for detection of abnormal observations using the NovoSpark curves. The curves showing the abnormal observations are distinguished by the shape, with some lying outside of the “cloud”. The “cloud” is marked in gray on the chart. According to the image we can identify abnormal observations in the following years: 1999, 2000, 2001 and 2002.



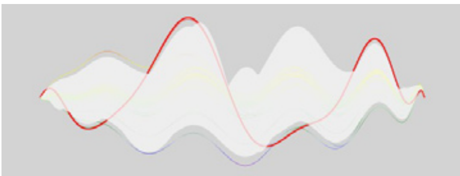
a) Abnormal observations in solar activity dynamics



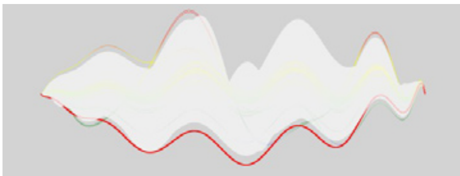
b) 1999



c) 2000



d) 2001



e) 2002

Figure 5. Identification of anomalies in data of the solar activity dynamics

Source: prepared by the authors.

Let us compare the selected traditional and NovoSpark methods. We choose the traditional methods which provide the ability to show the dynamics in data sets. Figure 6 shows linear data visualization. This figure presents clearly each parameter, but it is hard or even impossible to analyse the whole “integral” situation and show the dynamics according to months. We also can’t determine the anomalies and trends, just the maximum and minimum values using traditional parallel coordinates (Figure 7).

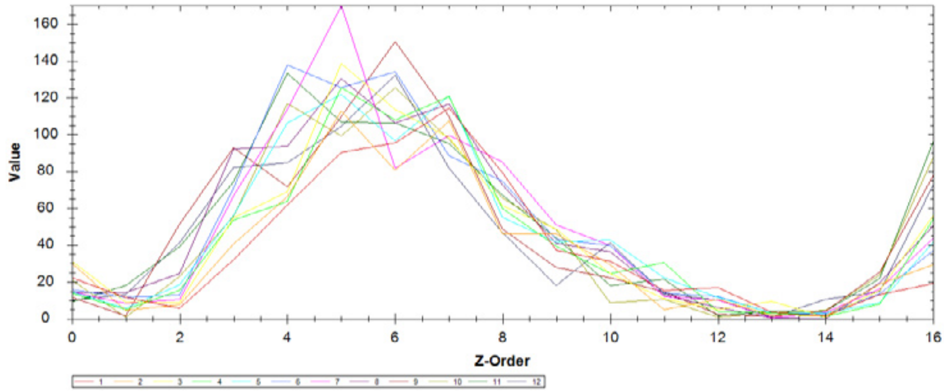


Figure 6. Linear plots for solar activity dynamics

Source: prepared by the authors.

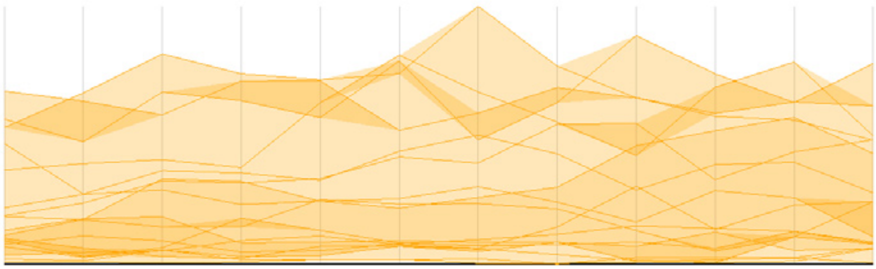


Figure 7. Traditional parallel coordinates for solar activity dynamics

Source: prepared by the authors.

The conducted experiments show the most abnormal moments in the solar activity dynamics. The 3-D and 2-D visualization using NovoSpark curves let us analyze the integral state of the whole situation.

Conclusions

Visualization is a powerful tool for data analysis, but for the moment, unfortunately, there is a number of limitations. Typically, the data used in the analysis is reflected on two-dimensional charts, the effectiveness of which is sometimes questionable. Even 3-D images can fail. New idea of increasing the effectiveness of computer visualization is the creation of integral images and conversion of numerical data to color data. Such a solution is proposed in the NovoSpark Visualizer system and works well in practice, which is proven in numerous studies (www.novospark.com).

References

Andrews D., Plots of High Dimensional Data, *Biometrics* 1972, No. 28, pp. 125–136.

Chernoff H., The Use of Faces to Represent Points in K-Dimensional Space Graphically, *Journal of the American Statistical Association (American Statistical Association)* 1973, No. 68 (342), pp. 361–368.

Dzemyda G., Kurasova O., Žilinskas J., Multidimensional data visualization, *Methods and applications series: Springer optimization and its applications* 2013, 75 XII (250), p. 122.

Eidenzon D., Volovodenko V. (2009), *Method for visualization of multidimensional data*, Patent Application 20090252436, USA.

Eidenzon D., Volovodenko V., Shamroni D., *Method and system for multidimensional data visualization*, Lambert Academic Publishing, Saarbrücken 2013.

Few S., *Multivariate analysis using parallel coordinates*, 2006, www.perceptualedge.com/articles/b-eye/parallel_coordinates.pdf.

Gemignani, Z. *Better know a visualization: Parallel coordinates*, 2010, www.juiceanalytics.com/writing/parallel-coordinates [accessed on 25.11.2013].

Graepel T., Burger M., Obermayer K., Self-organizing maps: Generalizations and new optimization techniques, *Neurocomputing* 1998, 21, pp. 173–190.

Inselberg A., *Parallel Coordinates: VISUAL Multidimensional Geometry and its Applications*, Springer, 2009, p. 580.

Inselberg A., The Plane with Parallel Coordinates, *Visual Computer* 1985, No. 1 (4), pp. 69–91.

Jolliffe I., *Principal component analysis*, Springer, New York 2002.

Kandogan E., *Visualizing multi-dimensional clusters, trends, and outliers using star coordinates. Proceedings of International Conference on Knowledge Discovery and Data Mining*, ACM Press, New York 2001, pp. 107–116. http://portal.acm.org/citation. cfm?id=502512.502530.

Kohonen T., *Self-Organizing Maps*, Springer-Verlag, 2001.

Pilipczuk O., Eidenzon D., The application of cognitive computer graphics to economic data exploration, *Journal of Automation, Mobile Robotics & Inteligent Systems* 2013, 7 (3), pp. 3–9.

Pilipczuk O., *Grafika kognitywna w podejmowaniu decyzji*, Wydawnictwo Naukowe Uniwersytetu Szczecińskiego, Szczecin, 2013.

Reitsma R., Trubin S., Information space partitioning using adaptive voronoi diagrams. *Information Visualization* 2007, No. 6 (2), pp. 123–138.

Sun Y., Tang J., Tang D., Xiao W., *Advanced star coordinates*. In: *Web-Age Information Magagement*, 2008. WAIM 08. The Ninth International conference, pp. 165–170.

Volovodenko V, Eidenzon D., Mylcev K., Method and system for the visualization of multidimensional objects and processes, *VINITII*. 1991, Vol. 1471-B91, p. 13.

www.bom.gov.au.

www.novospark.com.

ZASTOSOWANIE SYSTEMU NOVOSPARK®VISUALIZER DO WIZUALIZACJI DANYCH WIELOWYMIAROWYCH

Streszczenie

Tradycyjne metody wielowymiarowej wizualizacji danych (takie jak macierze wykresów rozproszonych, współrzędne biegunowe, twarze Chernoffa itp.) nie pozwalają na jednoczesne wyświetlanie wszystkich wartości wymiarów statycznych lub dynamicznych w ramach jednego przejrzystego obrazu. W artykule autorzy przedstawiają metodę wizualizacji i jakościowej analizy danych wielowymiarowych w systemie NovoSpark®Visualizer. Podano przykład zastosowania systemu do analizy danych wybranych wskaźników aktywności słonecznej.

Słowa kluczowe: dane wielowymiarowe, wizualizacja, NovoSpark®Vizualizer