MATHEMATICAL MODELS FOR DATA PROCESSING IN INFORMATION SYSTEMS

Ivana Linkeová
Czech Technical University in Prague, Faculty of Mechanical Engineering

Abstract: *Frequent requirement in processing the data from information systems is to find the appropriate analytical expression for the relation among variables or quantities and the possibility of accurate graphical representation of this expression. This paper is focused on the standard and non-standard tools in Microsoft Excel by means of which it is possible to achieve these requirements.*

Keywords: *Information system, data processing, trend line, regression curve, interpolation curve.*

1. Introduction

The implemented sophisticated tools for data analysis enable the user to carry out financial, engineering and statistical analysis of the data from information systems, therefore this application is widely used in many areas of economic practice. If the user is not a mathematician, then Microsoft Excel is often the only mathematically oriented program which the user has to his disposal.

A frequent requirement in processing the data from information systems is to find the analytical and graphical representation of the processed data. There are two basic approaches to interpret the processed data: approximation and interpolation. In the case of approximation, the resulted characteristic in graphical representation does not pass precisely through the data, only the data trend is satisfied. This way of interpreting the processed data can be used when studying the long-trend of the data. In the case of interpolation, the obtained characteristic passes precisely through the processed data and its shape between known values is given by the used mathematical model. This description is more sensitive and it is suitable when the detailed short-term trend is pursued and all deviations in this trend are important.

By means of implemented tools and functions in Microsoft Excel, it is, in simpler cases, possible to find an analytical expression of dependence among the processed variables using methods of regression analysis. Here, the resulted characteristic is the approximation model of the data. However, if the interdependence of the data is not evident, or if the input values are to be interpolated, these cases cannot be straightforwardly solved by standard tools in Microsoft Excel.

This paper is organized as follows: Section 2 concerns with the possibilities and restrictions of standard graphical and analytical tools which are implemented in Microsoft Excel. In section 3, the principles of suitable mathematical models for interpolation are presented, and the summary is given in Conclusion.

2. Standard tools of Microsoft Excel

The standard implemented tools and functions of Microsoft Excel offer the user the following options for the graphical representation and analytical expression of relationship among the input data: plot the input data on a chart and find analytical relationship using methods of regression analysis.

1) Graphical representation

To create a chart, the user has to enter the input data for the chart to the worksheet first. The user is limited by the choice of the chart type. Only the XY (Scatter) chart enables to plot the characteristic – curve as a graph of dependent variable on the independent variable. The plotted curves can be expressed explicitly or parametrically. All remaining chart types in Microsoft Excel only divide the *x*-axis into the appropriate number of equidistant segments according to the number of data categories which have to be displayed. In Fig. 1, there is comparison of graphical representation of the function $y = 1/x$ plotted in XY (Scatter) and Line charts. It is obvious that the shape of characteristic in Line chart is distorted.

Input data	
x	y
0.17	6.00
0.25	4.00
0.50	2.00
1.00	1.00
2.00	0.50
4.00	0.25
6.00	0.17

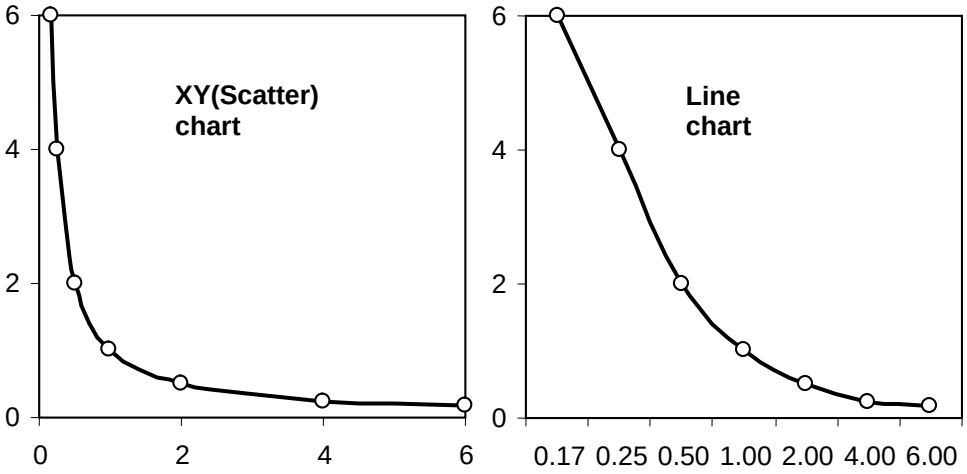

Fig. 1: Comparison of the characteristics in XY (Scatter) and Line charts

XY (Scatter) chart allows plotting numerical values of the independent variable on the *x*-axis and calculated function values of the dependent variable on the *y*-axis. The graphical representation of the dependence between input data can be plotted as a polyline (points are connected by line segments), or a smoothed line, see Fig. 2.

Input data	
x	y
0	0
0.4	4
1.5	3.8
2.5	3.6
3.2	3.4
4	3
4.7	2.5
5.7	2.3
6	0

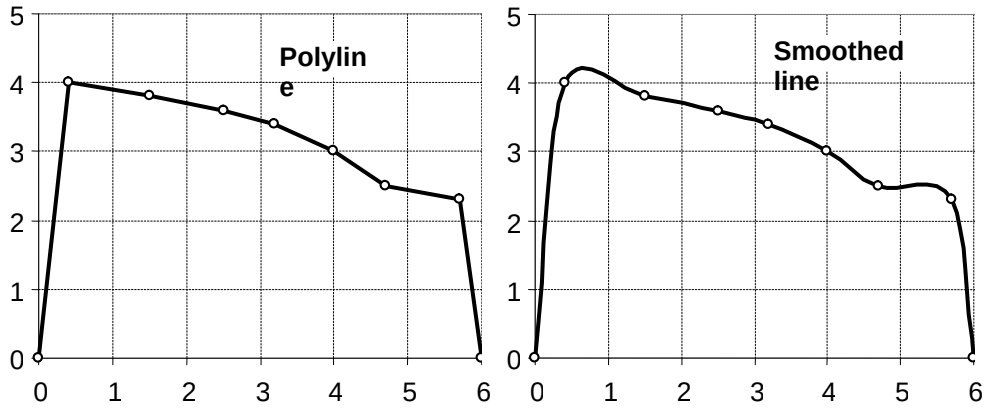

Fig. 2 Polyline or smoothed line in the XY(Scatter) chart

Precise calibration of the chart is impossible, therefore the shape of the plotted characteristic may be deformed with any change of the shape and size of all objects in the graph (e.g. legend, title, axes labels, etc.). If the main goal of graphical representation is to obtain the characteristic in its true shape (without distortion caused by different size of the axes division), it is necessary to be very careful during editing of these objects.

2) Analytic representation

The user has no access to the information either about the actual mathematical model used to plot the smoothed line (Fig. 2) or about its analytical expression. This fact can be considered a big disadvantage for further data processing.

A possible analytical expression of the dependence between individual variables can be determined using the methodology of regression analysis. The user can add a regression curve – trendline – to the selected characteristic in the chart. The types of regression curve which are offered in Microsoft Excel is a straight line (in the case of linear dependence) or polynomial, power, logarithmic and exponential curve (in the case of nonlinear dependence). All these curves are fitted using the least squares method. Type of the regression curve can be chosen by the user interactively according to the form of the input data.

The corresponding correlation equation can be plotted together with the regression curve. The example of the obtained polynomial trendline of 2nd and 6th degree and their equations are shown in Fig. 3. The set of input points is the same as the set of points depicted in Fig. 2.

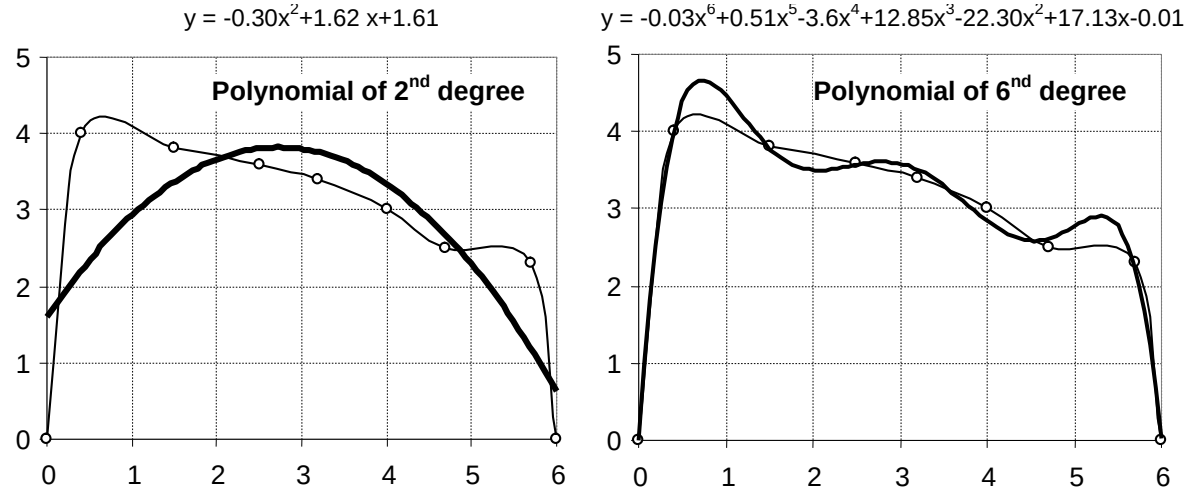

Fig. 3 Polynomial trendline

There are disadvantages of regression analysis methodology as follows. The user has to know that the calibration of the chart axes cannot be performed, therefore the shape of the displayed curves can be deformed. The methodology of regression analysis can be used only in a limited number of simpler cases. If the input data is to be fitted by the appropriate curve, the regression analysis can not be used.

3. Non-standard tools in Microsoft Excel

In order to overcome the above mentioned limitations of standard tools of Microsoft Excel, a procedure for graphical and analytical solution of interpolation problem was developed. This procedure allows a user without specialized mathematical skills to find the appropriate interpolation model for processed data. The procedure is programmed in Visual Basic for Applications as Interpolation add-in [3]. The interpolation method used in Interpolation add-in is based on so called Hermite interpolation [1] which is a special case of NURBS (Non-Uniform Rational B-Spline) representation [4] widely used in computer graphics. Here, the mathematical theory of this method is described.

The Hermite interpolation curve is defined by the series of given points (input data) $\mathbf{P}_0, \mathbf{P}_1, \ldots, \mathbf{P}_n$ and geometric conditions in these points – e.g. the magnitude and direction of tangent vectors as well as the vectors of second derivative at these points.

Here, the segment of designed interpolation curve of 3rd degree is defined by two consecutive endponints $\mathbf{P}_i$, $\mathbf{P}_{i+1}$ from input data and tangent vectors $\mathbf{P'}_i$, $\mathbf{P'}_{i+1}$ at these endpoitns. Vector equation of one segment of the resulted curve is

$$\mathbf{P}_i(t) = \begin{bmatrix} t^3 & t^2 & t & 1 \end{bmatrix} \cdot \begin{bmatrix} 2 & -2 & 1 & 1 \\ -3 & 3 & -2 & -1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} \mathbf{P}_i \\ \mathbf{P}_{i+1} \\ \mathbf{P}' \\ \mathbf{P}'_{i+1} \end{bmatrix}, t \in [0,1],\ i = 0,1,\ldots,n-1. \tag{1}$$

The magnitude and direction of tangent vectors significantly influences the shape of the resulted curve. The Interpolation add-in offers four methods different in the way of calculating tangent vectors: L-interpolation, N-interpolation, D-interpolation and natural spline interpolation.

1) L-interpolation method

This method is suitable in the case of uneven distribution of input data. The shape of the curve obtained by this method is satisfied in majority of practical cases. Unknown tangent vectors required to calculate Eq. (1) are determined as follows [2]:

$$\begin{aligned} \mathbf{P}'_0 &= -\tfrac{2}{3}\mathbf{P}_0 + 2\mathbf{P}_1 - \tfrac{1}{2}\mathbf{P}_2, \\ \mathbf{P}'_i &= d_i \cdot \tfrac{1}{2}\left(\mathbf{P}_{i+1} - \mathbf{P}_{i-1}\right),\ i = 1,2,\ldots,n-1, \\ \mathbf{P}'_n &= \tfrac{1}{2}\mathbf{P}_{n-2} - 2\mathbf{P}_{n-1} + \tfrac{3}{2}\mathbf{P}_n. \end{aligned} \tag{2}$$

Coefficients d_i represent correction of the length of calculated tangent vectors. It is a linear function (therefore "L"-interpolation) of the angle γ_i, which is an angle between two consecutive straight line segments $\mathbf{P}_{i-1}\mathbf{P}_i$ and $\mathbf{P}_i\mathbf{P}_{i+1}$:

$$d_i = \frac{d_{\min}}{d_{\max}} + \frac{d_{\max} - d_{\min}}{d_{\max}} \cdot \frac{\gamma_i}{180},\ i = 1,2,\ldots,n-1, \tag{3}$$

where $d_{\min} = 1$ and $d_{\max} = 3$ are empirically specified. d_i reaches the maximum value 1 which corresponds to the angle $\gamma_i = 180°$ and minimum value 1/3 for the angle $\gamma_i = 0°$.

The resulting characteristic obtained by L-interpolation method is shown in Fig. 4. Its shape is very similar to the shape of the smoothed line from Fig. 2. L-interpolation method is suitable in majority of cases in economic practice.

2) N-interpolation method

Tangent vectors for N-interpolation method are calculated according Eq. (2), but correction of their length is a non-linear (therefore "N"-interpolation) function of the angle between two consecutive straight segments $\mathbf{P}_{i-1}\mathbf{P}_i$ and $\mathbf{P}_i\mathbf{P}_{i+1}$:

$$d_i = \frac{1}{1 + \frac{d_{\max} - d_{\min}}{180}\left(180 - \gamma_i\right)},\ i = 1,2,\ldots,n-1. \tag{4}$$

The minimum and maximum values $d_{\min} = 1$ and $d_{\max} = 3$ are empirically specified too. The introduction of non-linearity into Eq. (4) has a great influence on the resulting shape of the characteristic, see Fig. 5. The overshootings of the curve are strongly reduced, therefore this model can be used in the case of highly uneven distribution of the processed data.

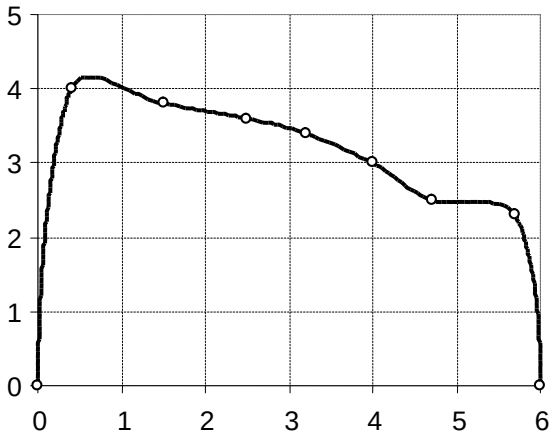

Fig. 4 L-interpolation

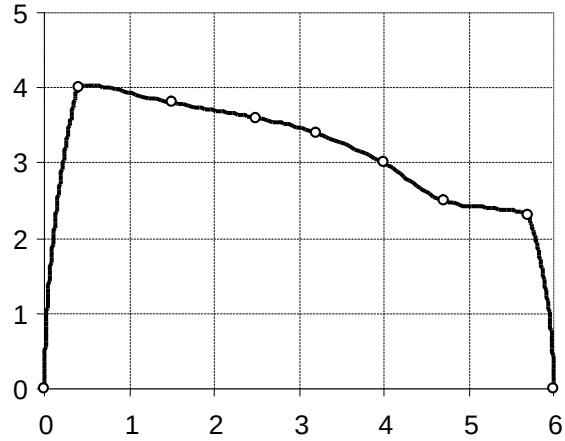

Fig. 5 N-interpolation

3) D-interpolation

D-interpolation method is a method proposed in [5]. The unknown tangent vectors are calculated as follows: the tangent vectors $\mathbf{P'}_i$, $i = 1, 2, \ldots, n\text{-}1$, at all internal points is given as the image of the vector $\mathbf{b}_i = \overrightarrow{\mathbf{P}_{i-1}\mathbf{P}_i}$ in the orthogonal mapping to the bisector a_i of the angle φ_i between vectors $\mathbf{b}_i = \overrightarrow{\mathbf{P}_{i-1}\mathbf{P}_i}$ and $\mathbf{c}_i = \overrightarrow{\mathbf{P}_i\mathbf{P}_{i+1}}$, see Fig. 6.

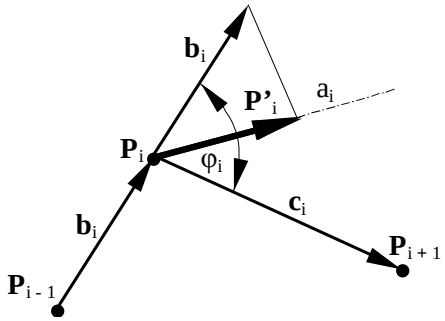

Fig. 6 Determination of tangent vector in D-interpolation

The curve which is constructed according D-interpolation method [5] does not pass through the initial point $\mathbf{P}_i$ and the end point $\mathbf{P}_n$, because the tangent vectors at these points are not defined. In Interpolation add-in, the tangent vectors $\mathbf{P}_i$ and $\mathbf{P}_n$ are calculated according Eq. (2). The result obtained for the same data as in L- and N-interpolations is shown in Fig. 7. The D-interpolation method is possible in the case of highly even distribution of the processed data.

4) Natural spline interpolation

The above mentioned methods L-, N- and D-interpolation satisfy the condition for continuity of first degree (first derivative is continuous along the whole curve). To cover more strict continuity requirements, the natural spline method [1] is available in Interpolation add-in. The natural spline interpolation satisfy two end conditions

$$\mathbf{P}''_0 = 0, \mathbf{P}''_n = 0 \tag{6}$$

and tangent vectors at all internal points $\mathbf{P'}_i$, $i = 1, 2, \ldots, n\text{-}1$, are calculated from the condition for continuity of second degree (second derivative is continuous along the whole curve). The resulted curve obtained by natural spline interpolation is depicted in Fig. 8.

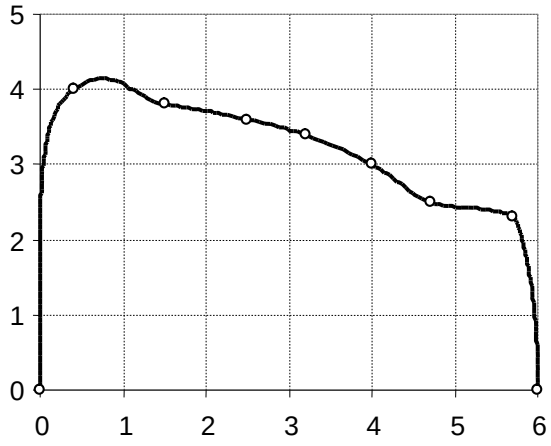

Fig. 7 D-interpolation

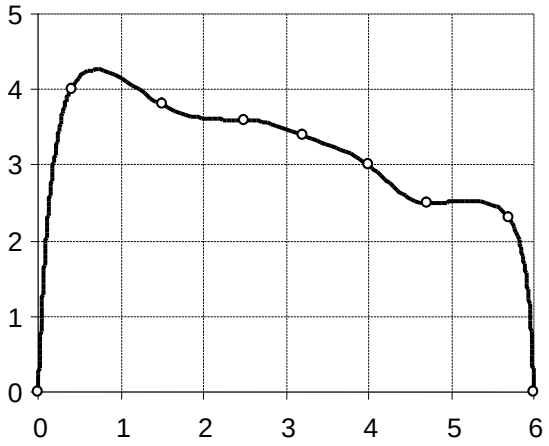

Fig. 8 Natural spline interpolation

4. Conclusion

The graphical and analytical representation of the data from information systems by means of standard and non-standard tools of Microsoft Excel are described in this paper. In simpler cases, it is possible to find analytical representation of processed data using methods of regression analysis. In general, this approach is not applicable and Microsoft Excel can not be straightforwardly used. To overcome this limitation, the mathematical theory of four interpolation methods is presented and graphical representation of the same set of processed data is shown. These mathematical models can to be recommended to use when the detailed short-term trend is pursued and all deviations in processed data are important.

References:

[1] FOLEY, J. D., VAN DAM, A., FEINER, S. K., et.al, *Computer Graphics.* 1990. Addison-Wesley Publishing Company.

[2] LINKEOVÁ, I. *Konstrukce, výroba a měření obecných tvarových ploch*, disertační práce, Fakulta strojní Českého vysokého učení technického v Praze, Ústav strojírenské technologie, 1999.

[3] LINKEOVÁ, I. Interpolation Add-In Interpolation Add-In for MS Excel, In: *Mathematical and Computer Modelling in Science and Engineering*. Prague: CTU, 2003, vol. 1, p. 211-215.

[4] LINKEOVÁ, I. *NURBS křivky (NeUniformní Racionální B-spline křivky).* 1. vydání. Praha: Vydavatelství ČVUT, 2007, 208 s. ISBN 978-80-01-03893-2.

[5] VELICHOVÁ, D. *Konštrukčná geometria*. 1. vydání. Vydavateľstvo Slovenskej technickej univerzity v Bratislave, 1999, 201 s. ISBN 80-227-0904-2.

Contact address:

Ing. Ivana Linkeová, Ph.D.
Czech Technical University in Prague
Faculty of Mechanical Engineering – Department of Technical Mathematics
Karlovo nám. 13, 121 35 Praha 2 – Nové Město
Email: Ivana.Linkeova@fs.cvut.cz
+420 224 357 534