

## Cubic smoothing spline

Menu:	QCExpert	Regression	Cubic spline
-------	----------	------------	--------------

The module *Cubic Spline* is used to fit any functional regression curve through data with one independent variable  $x$  and one dependent random variable  $y$ . Number of points  $(x_i, y_i)$  is  $n$ . The regression model  $y = f(x) + \varepsilon$  is composed of  $p$  cubic curves defined on  $p$  adjacent segments  $(-\infty, x_{u,1}) \cup (x_{u,1}, x_{u,2}) \cup \dots \cup (x_{u,p-1}, +\infty)$  of the  $x$ -axis separated by  $p-1$  knots. The values  $x_{u,i}$  are knots and can be defined by the user. Analytical properties of the curve (like smoothness, curvature, continuity) and statistical properties (residual variance, prediction variance) are subject to modeling. The polynomial regression spline can be written generally as

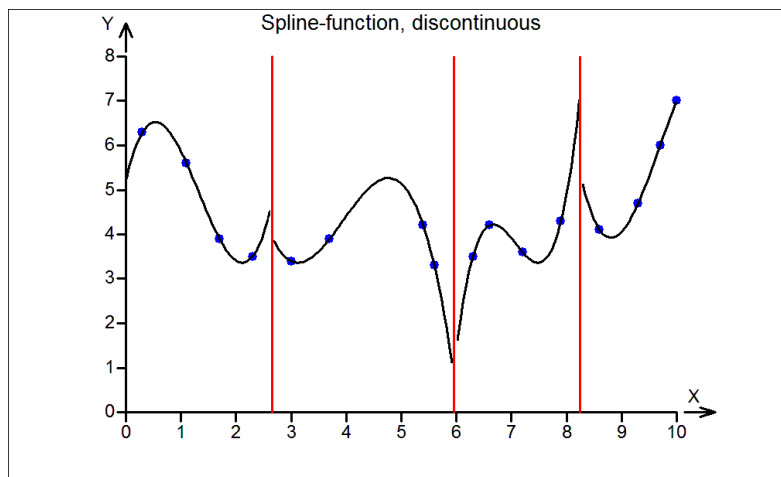
$$y(x) = [S^h(x)]^T \mathbf{a}_k,$$

$$[S^h(x)]^T = [1, x, x^2, \dots, x^h]; h \geq 0$$

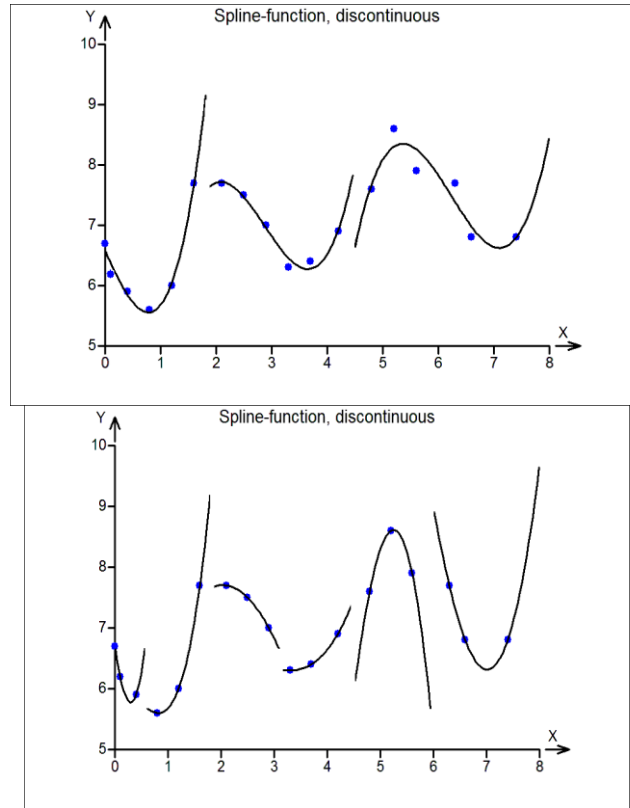
where  $h$  is the polynomial order, in case of  $h=3$  it is a cubic spline.  $S^h(x)$  is the vector or predictors and  $\mathbf{a}_k$  is vector of regression coefficients in the  $k$ -th segment. If the measured value at  $x_i$  is  $y_i$ , and  $x_i$  is in the  $k$ -th interval ( $k$ -th segment) which is the interval  $(x_{u,k-1}; x_{u,k})$ , then the predicted value  $f(x_i)$  is given by

$$f(x_i) = a_{k,1} + a_{k,2}x_i + a_{k,3}x_i^2 + a_{k,4}x_i^3$$

Parameters  $a_{k,i}$  are computed by linear least squares regression. If no continuity for the function at knots was required, the model would be  $p$  independent regression polynomials as shown on Fig. 1. The shape of the regression curve depends heavily on the chosen number of segments as shown on Fig. 2a, b.



**Fig. 1 Example – four independent cubic regression polynomials (discontinuous spline)**



**Fig. 2 a, b Discontinuous regression splines for  $p = 3$  and  $p = 6$**

The above models are discontinuous, they show flexibility of regression splines, they are of little use however. Usually, we want the model to be continuous in  $f(x)$ , so we impose the condition

$$\left[ S^h(x) \right]^T \mathbf{a}_{k-1} = \left[ S^h(x) \right]^T \mathbf{a}_k.$$

Further, we can make use of easy differentiability of polynomials and introduce conditions of continuous first and second derivatives at knots. For our cubic splines we have  $h = 3$ .

$$\left[ S^{h'}(x) \right]^T \mathbf{a}_{k-1} = \left[ S^{h'}(x) \right]^T \mathbf{a}_k, \quad \left[ S^{h''}(x) \right]^T \mathbf{a}_{k-1} = \left[ S^{h''}(x) \right]^T \mathbf{a}_k$$

where the first and second derivative of the polynomial are defined as follows:

$$\left[ S^{h'}(x) \right]^T = \frac{\partial \left[ S^h(x) \right]^T}{\partial x} = \left[ 0, 1, 2x, 3x^2, \dots, hx^{h-1} \right]$$

$$\left[ S^{h''}(x) \right]^T = \frac{\partial^2 \left[ S^h(x) \right]^T}{\partial x^2} = \left[ 0, 0, 2, 6x, \dots, h(h-1)x^{h-2} \right].$$

Imposing these conditions will result in a curve with second order differentiability. Such a spline curve will be suitable for a wide class of applications. This smooth curve will

have continuous second derivative and smooth first derivative in the sense that  $|S^{3''}(x) - S^{3''}(x+\delta)| < \epsilon$  for any  $\epsilon > 0$  and sufficiently small nonzero  $\delta$ .

Compared to kernel smoothers or moving average (see Basic Statistics module), cubic splines eliminate bias in estimating mean, especially near maxima or minima. Thanks to using a linear regression model it is possible to compute statistical characteristics such as confidence intervals, standard deviation, covariance matrices, error estimates. Curves on Fig. 3 through Fig. 5 show the same data fitted with splines with continuous  $f(x)$ , continuous  $f(x)$  and  $f'(x)$ , and continuous  $f(x)$ ,  $f'(x)$  and  $f''(x)$ . The data are the same as in Fig. 1. No boundary conditions are posed at the beginning or the end of the curve in the Cubic Spline module.

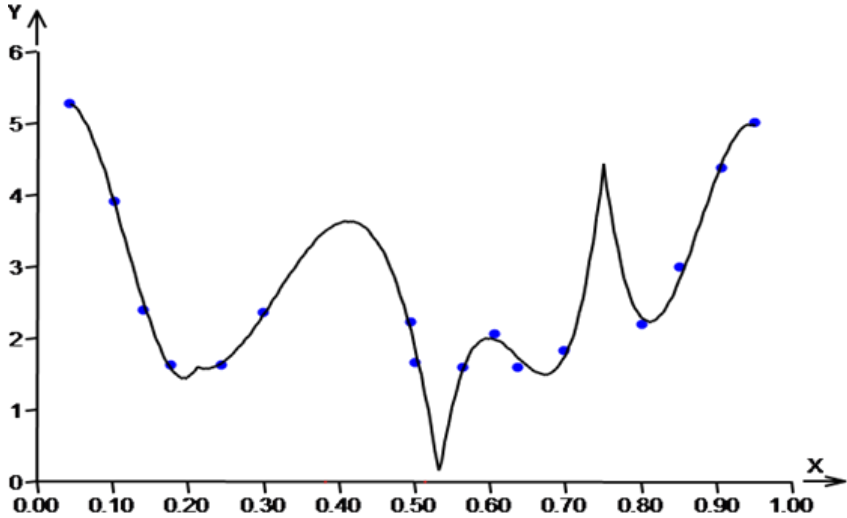


Fig. 3 Fit the data with a continuous spline function  $f(x)$

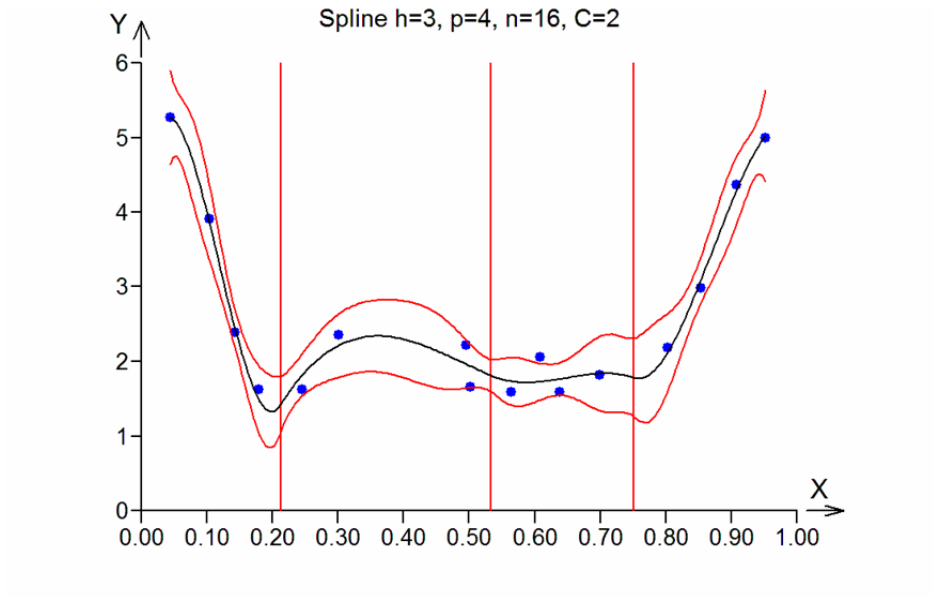
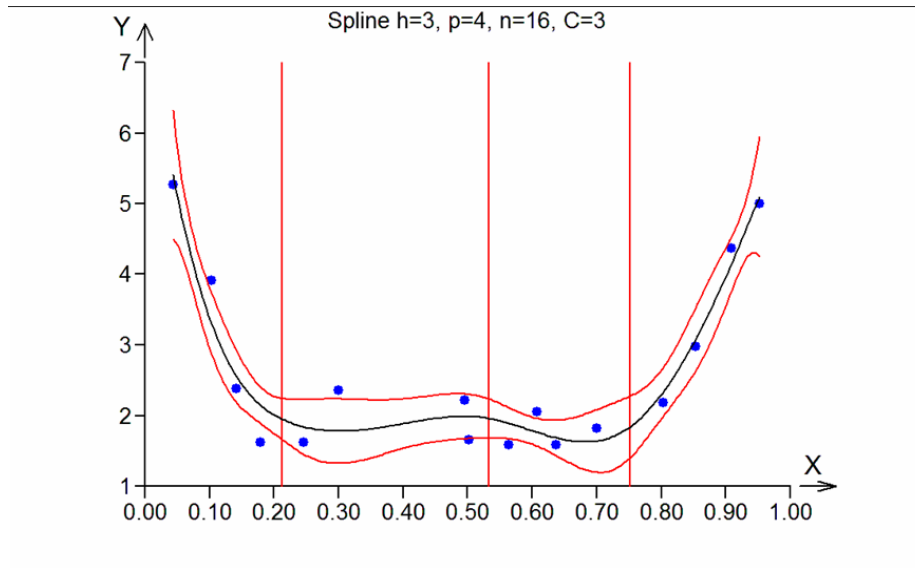
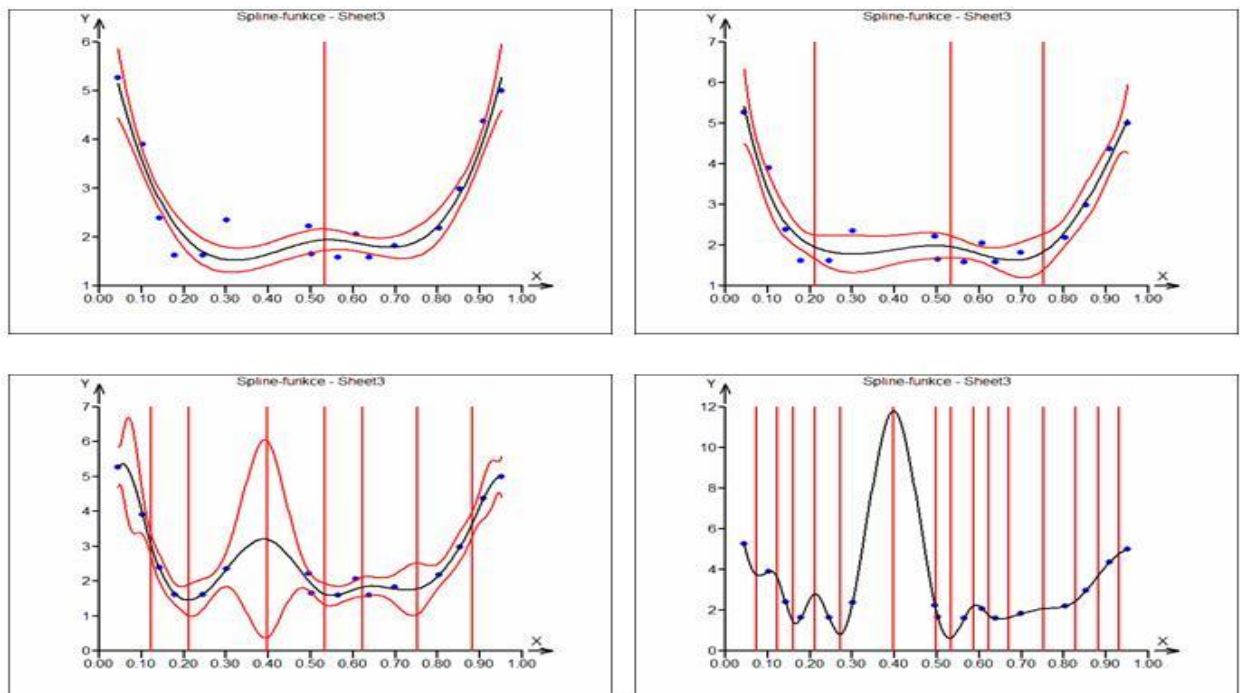


Fig. 4 Fit the data with continuous  $f(x)$  and  $f'(x)$



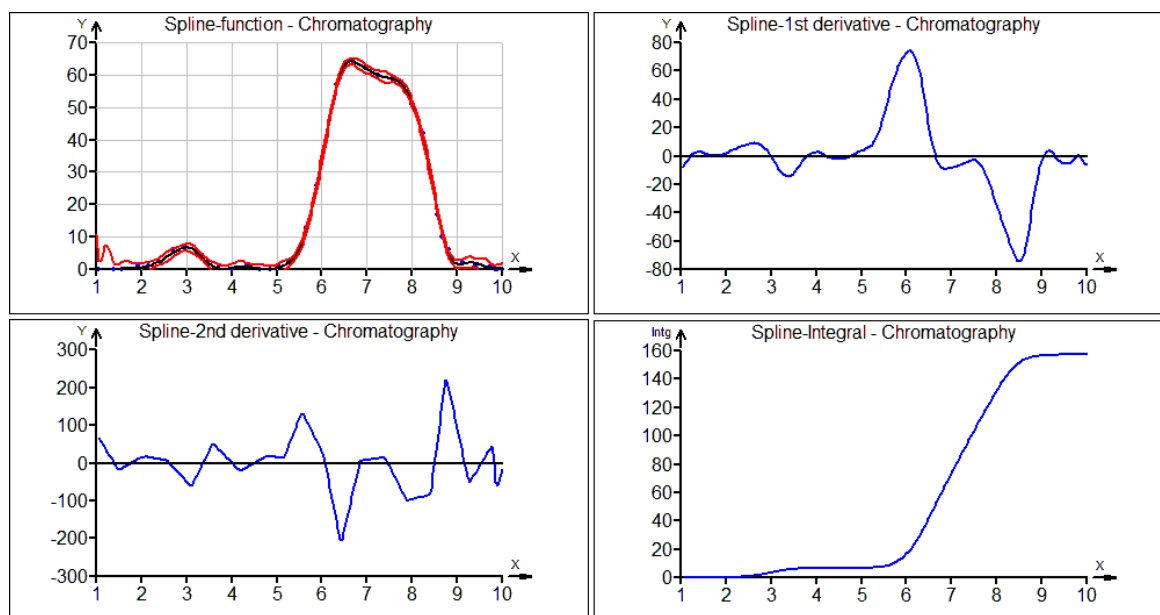
**Fig. 5 Fit the data with continuous  $f(x)$  and  $f'(x)$  and  $f''(x)$  (the “smoothest” curve)**

Knots can be defined by the user (by default, the knots are spaced on  $x$  so that the number of points in each segment be the same if possible). The influence of number and position of knots on the spline curve is illustrated on Fig. 6. The smallest number of knots is 1, dividing the spline into two segments (number of segments  $p = 2$ ). Maximum number of knots is  $n - 1$  (not generally recommended), where  $n$  is number of data points. In this case, the spline can provide no statistical information (such as variances or confidence intervals) as it passes through all points. Since the splines can sometimes be overdetermined a generalized Moore-Penrose pseudoinverse is used to compute the regression coefficients to ensure correct solution.

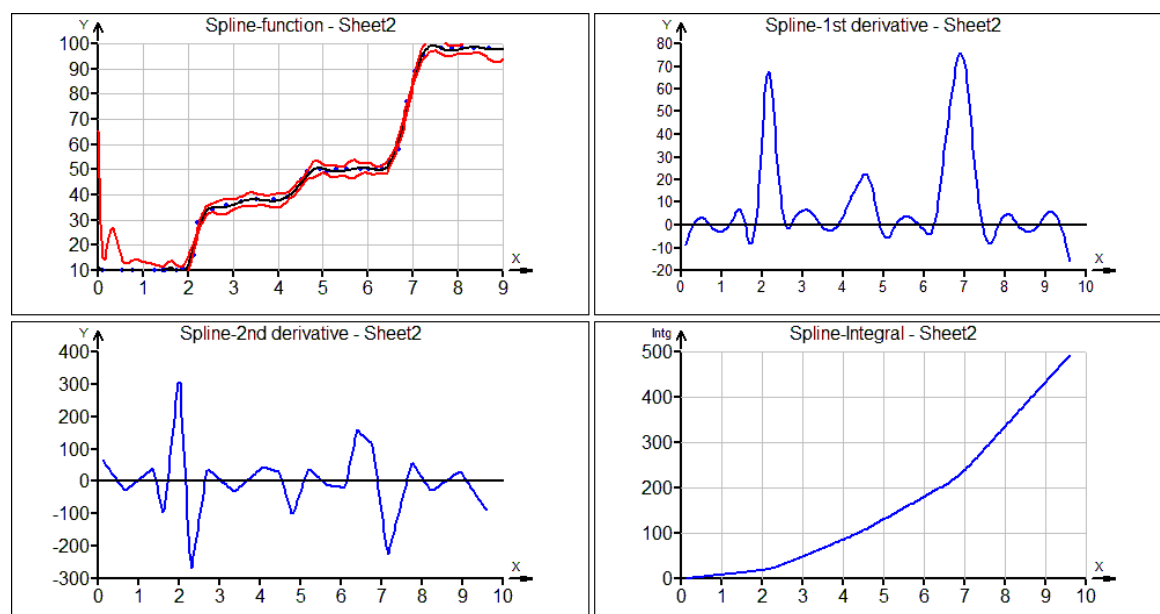


**Fig. 6 Influence of knots selection (same data in all 4 plots)**

Polynomial form of the spline curve allows analytical evaluation of derivatives and integrals of the curve which can be used for analysis of physical or chemical processes as suggested on Fig. 7 and Fig. 8.



**Fig. 7 Using spline to find area under curve and derivatives**



**Fig. 8 Using Spline module to determine inflex points on a titration curve**

## Data and parameters

Data X and Y are expected in two columns in the Data sheet. Further columns may contain positions of user-defined knots, new x-data for prediction and weights for individual measurements. An example of data structure is shown on Fig. 9 Fig. 10. It is good to keep in mind that each segment may contain at most one maximum, one minimum and one inflex

point. If the data for prediction is defined, the predicted y-values, their  $1 - \alpha$  confidence interval and analytical derivatives and integrals are computed in the protocol.

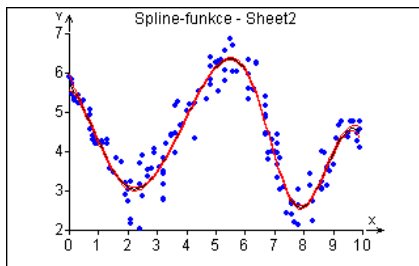
Note: The number  $m$  of columns of the characteristic matrix is  $m = 4p$ . There may be an issue of numerical instability in case of too many segments. It is recommended to use rather small number of segments, say 2 to 10, but not more than 50. A possible consequence of too many segments is illustrated on Fig. 11.

X	Y
0.2	5.91
0.7	4.77
1.4	2.93
1.9	2.93
2.6	2.31
3.5	2.67
3.9	2.46
4.7	2.9
5.7	4.71
6.7	5.35
7.1	5.5
8.6	5.97

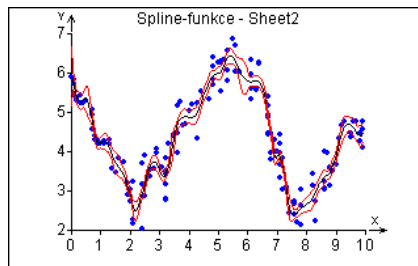
Fig. 9 Basic data structure for spline

X	Y	knots	predict	weight
0.2	5.91	2	0	1
0.7	4.77	4	1	1
1.4	2.93	6	2	2
1.9	2.93		3	4
2.6	2.31		4	5
3.5	2.67		5	5
3.9	2.46		6	4
4.7	2.9		7	4
5.7	4.71		8	3
6.7	5.35		9	2
7.1	5.5			2
8.6	5.97			2

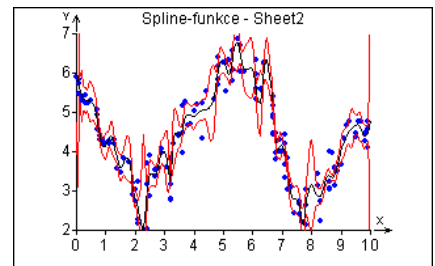
Fig. 10 Full possible structure of spline data



A:  $p=5$

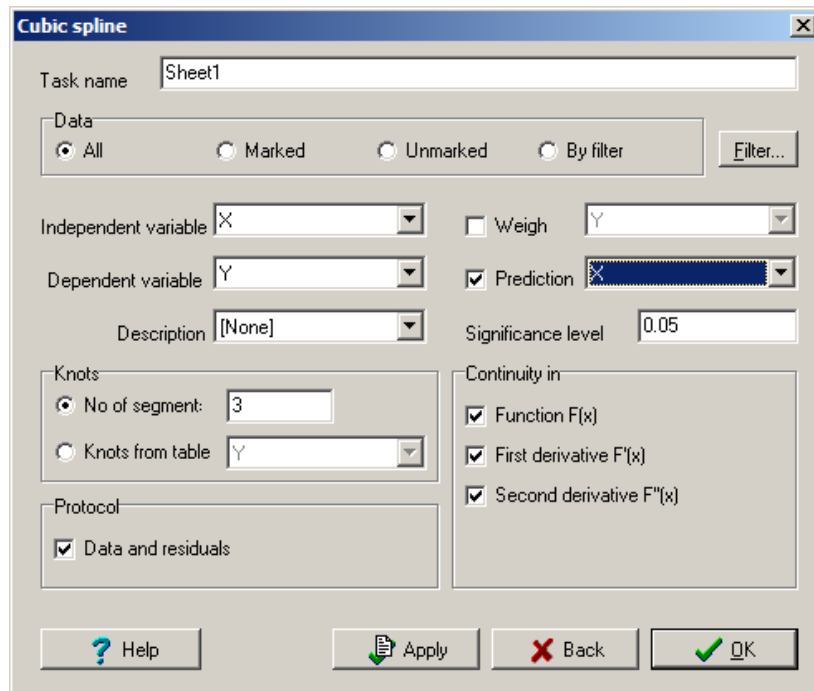


B:  $p=30$



C:  $p=50$

Fig. 11 Suitable (A, B) and too big (B, C) number of segments  $p$  for  $n=150$



**Fig. 12 Dialog window for Cubic spline module**

Open the Cubic Spline dialog window (*Menu: QC.Expert – Regression – Cubic Spline*). Choose the columns with independent and dependent variable. Optionally, choose the column with description of individual cases, check the *Weights* and *Prediction* check boxes and select appropriate columns. In the group *Knots* select *No of segments* (knots will be chosen automatically), or *Knots from table* (knots must be given in the selected column of the data table). It is recommended to choose number of segments  $p$  so that number of points  $n$  is divisible by  $p$ . Generally, number of points in the first  $(p - 1)$  segments is equal to  $\text{INT}(n/p)$ , the last  $p^{\text{th}}$  segment contains the rest of the points. For example, if  $n=12$  and  $p=7$  then first six segments will contain one point while the last segment will contain 6 points.

Continuity requirements are set in the group *Continuity in*. The user may select any combination of continuity in function values, first and second derivatives. By default, all boxes are checked (recommended). Checking the box *Data and residuals* in the group *Protocol* will produce a table of data and residuals. This table has as many rows as the original data and may be avoided if not necessary. Clicking *OK* will run the computation.

### Protocol

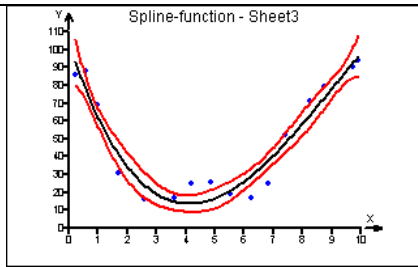
Task name	Task name from the dialog window
Data	Selected data subset (All, Marked, Unmarked, or user-defined filter)
Independent variable	Selected independent variable column
Dependent variable	Selected dependent variable column
No of points	Number of data rows $n$
No of knots	Number of knots $(p - 1)$
No of sections	Number of segments $p$
Continuous function	Required continuity in function values (Yes / No)
Continuous first derivatives	Required continuity in first derivatives (Yes / No)

Continuous second derivatives Knots positions	Required continuity in second derivatives (Yes / No) Number and x-position of knots
Model parameters	Number of data points in individual segments and estimated regression coefficients of the polynomial $y = a_0 + a_1x + a_2x^2 + a_3x^3$ ; A(0)=absolute term, A(1)=linear term, A(2)=quadratic term, A(3)=cubic term
Table of predicted values	If <i>Prediction</i> was selected in the Cubic Spline dialog window this table gives predicted value ( <i>Y-pred</i> ) for all a-values given in the <i>Prediction</i> column ( <i>X-value</i> ) together with confidence interval of prediction ( <i>Lower bound, Upper bound</i> ), and values of first and second derivatives.
Residual squares sum	Sum of the squared residuals (estimated errors)
Residual std. deviation	Standard deviation of the residuals
Residual variance	Variance of the residuals
Mean std. deviation	Residual standard deviation
Eff. degrees of freedom	Effective degrees of freedom, $v_{eff} = n + c*(p - 1) - 4p$ , where $p$ is number of segments and $c$ is number of continuity requirements. If the degrees of freedom is less than 1 no statistics is computed.
Tables of extremes and inflexes	Consists of two tables – a table of extremes and a table of inflexes in the interval of x-data ( $x_{min}, x_{max}$ ). If no extremes and inflexes are found this table is not created.
No of extremes	This table provides analytically computed extremes (minima and maxima) on the regression curve. Columns in the table include: Type of extreme (MIN or MAX), X- and Y-coordinate and second derivative (first derivative is always zero at the extreme).
Table of inflexes	This table provides analytically computed inflexes on the regression curve. Columns in this table include: X- and Y-coordinate and first derivative (second derivative is always zero at the inflex point).
Table of data and residuals	If the checkbox Data and residuals was checked in the Cubic Spline dialog window this table provides original data (“measured” X and Y values), predicted Y-values and residuals (difference between measured and predicted Y values).

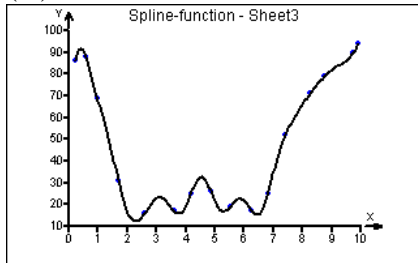
## Graphs

Spline – function	Plot of regression spline function. If degrees of freedom $v_{eff} > 0$ confidence intervals of the mean are plotted as red lines around the spline curve (plot A). If $v_{eff} = 0$ the curve is called interpolation spline as it passes through data (plots B-D). Plots (C) and (D) are splines with loosened continuity conditions. However, in most real
-------------------	---

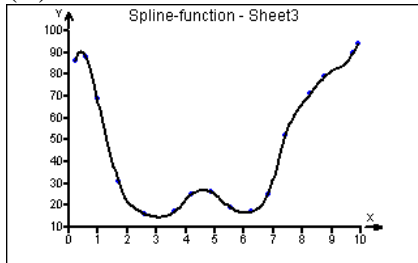




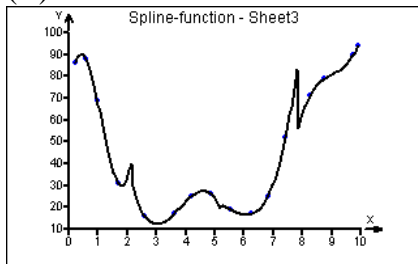
(A)



(B)

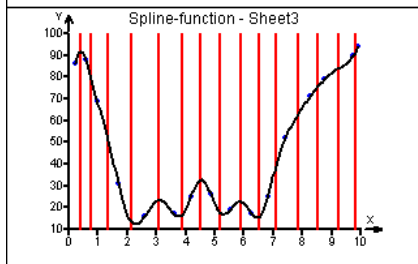
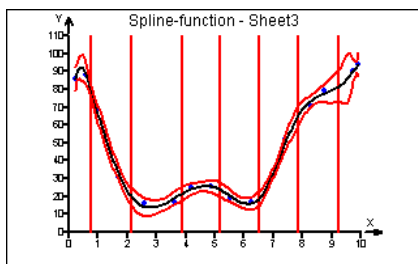


(C)



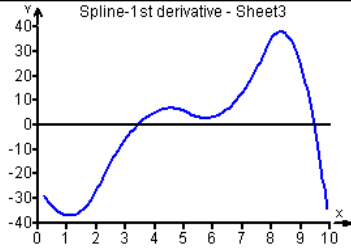
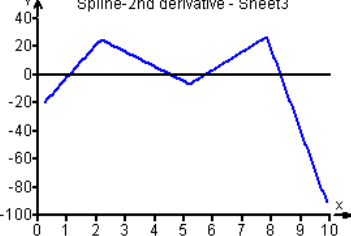
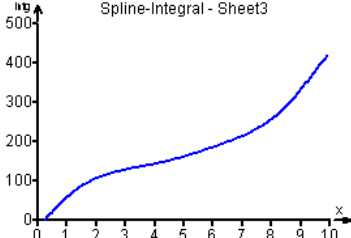
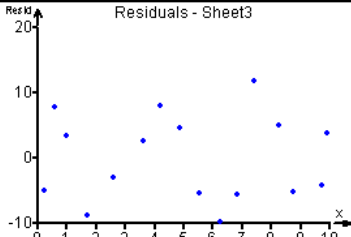
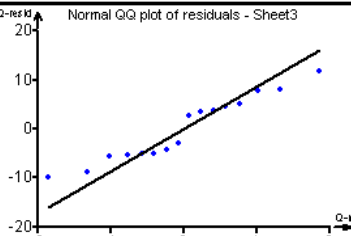
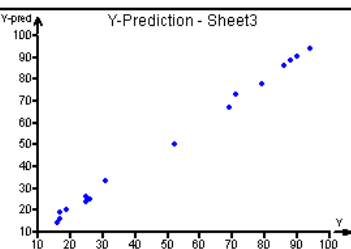
(D)

Spline – function with knots



cases,  $v_{eff} > 0$  and  $p \ll n$  with all continuity conditions imposed will be advisable.

Same as the previous plot with visible knot values.

	<p>First derivative <math>f^{(1)}(x)</math>. of the spline function. Zero-points correspond to local minima and maxima of the spline function. Numerical values of the first derivative are also available in Protocol in “Table of predicted values” and “Tables of extremes and inflexes”.</p>
	<p>Second derivative <math>f^{(2)}(x)</math>. of the spline function. Zero-points correspond to inflex points of the spline function. Numerical values of the second derivative are also available in Protocol in “Table of predicted values” and “Tables of extremes and inflexes”.</p>
	<p>Integral of the spline function.</p> $\int_{x_{\min}}^x f(z) dz$
	<p>Plot of residuals <math>y_i - f(x_i)</math>. If there is a trend in the residuals the knots may have been chosen improperly and changes in the model should be considered.</p>
	<p>Q-Q plot of the residuals. If the points lie close to the line the residuals (not necessarily the actual errors) have approximately normal distribution.</p>
	<p>Plot of Y-prediction plots the measured data against predicted y-values. The closer lie the points to a line the closer fit through the data. It is not generally the goal to have the closest fit (like interpolation spline). Rather we try to find reasonably complex model to describe the data realistically.</p>