Tutorials and worked examples for simulation, curve fitting, statistical analysis, and plotting.
http://www.simfit.org.uk

Cubic splines can be used for nonparametric comparison of two data sets for similarities and differences. Splines under tension are first fitted to each data set, then the areas under each curve and the absolute area between them are estimated using the trapezoidal method, integration of the best-fit curves, and Simpson's method for the absolute differences, in order to express differences as percentages.

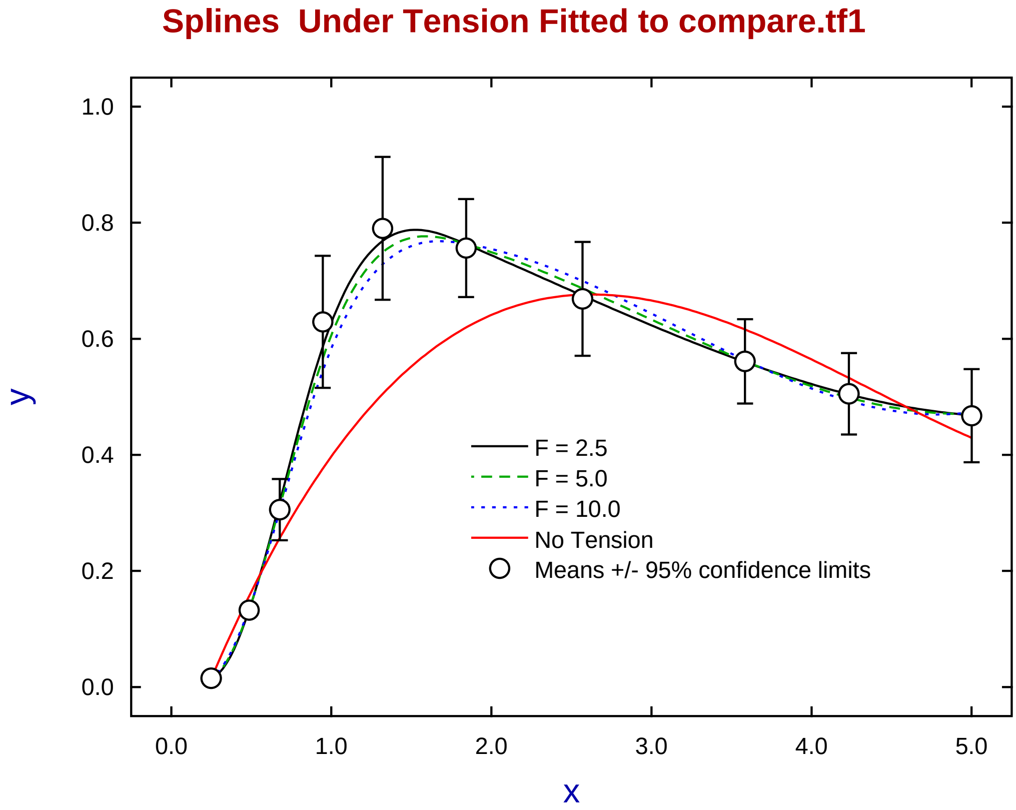
From the main SIMFIT menu select [A/Z], open program **compare** then view the default test file `compare.tf1` containing the following data.

x	y	se
0.25000	0.017267	1
0.25000	0.015585	1
0.25000	0.014268	1
0.25000	0.014136	1
0.48647	0.12861	1
0.48647	0.12536	1
0.48647	0.13339	1
0.48647	0.14230	1
0.67860	0.26261	1
0.67860	0.34277	1
0.67860	0.30364	1
0.67860	0.31373	1
0.94662	0.67252	1
0.94662	0.70382	1
0.94662	0.59192	1
0.94662	0.54850	1
1.3205	0.90417	1
1.3205	0.74158	1
1.3205	0.77353	1
1.3205	0.74208	1
1.8420	0.79030	1
1.8420	0.79384	1
1.8420	0.67971	1
1.8420	0.76176	1
2.5695	0.58575	1
2.5695	0.66178	1
2.5695	0.70023	1
2.5695	0.72772	1
3.5844	0.53286	1
3.5844	0.62744	1
3.5844	0.55484	1
3.5844	0.52923	1
4.2334	0.55003	1
4.2334	0.46641	1
4.2334	0.46840	1
4.2334	0.53647	1
5.0000	0.49920	1
5.0000	0.51847	1
5.0000	0.44355	1
5.0000	0.40895	1

The columns contain data in the following format.

1. **Column 1**: the variable x which must be in non-decreasing order.
2. **Column 2**: the response y presumed to be dependent on x.
3. **Column 3**: the value of 1 indicates that the replicates will be used to calculate the sample standard deviations at each x-replicate value to be used for weighting.
   This column can be omitted or set to a positive value se if it is wished to supply weighting factors w directly which would then be used as $w = 1/se^2$.

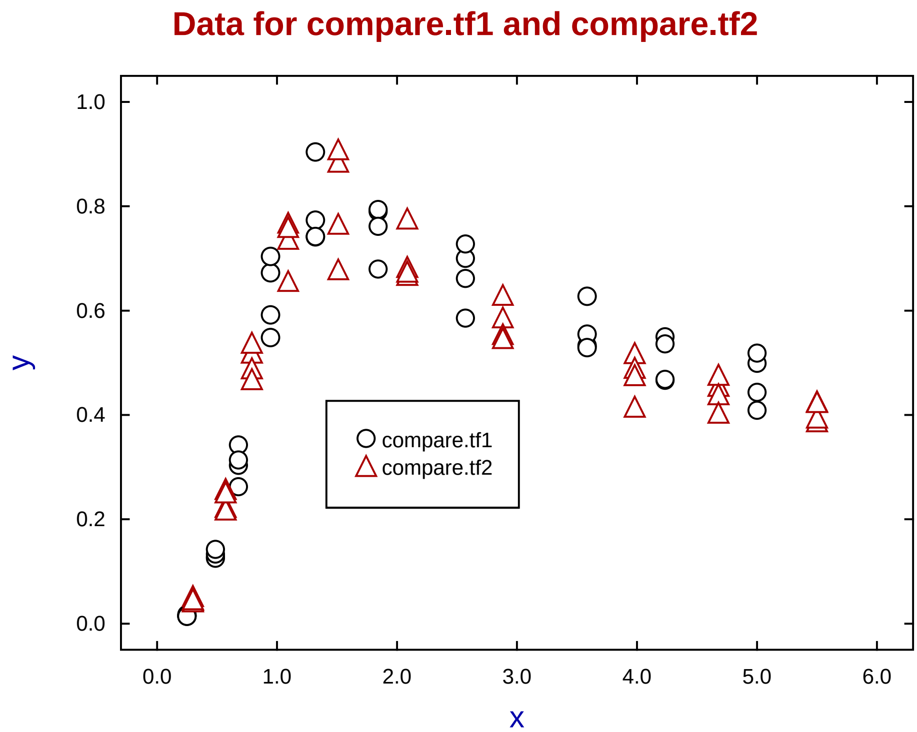
Splines were fitted with the default smoothing factor which simply fits a cubic with no internal knots, then the smoothing factor F was decreased to 10, which is the number of data points after replicates in the data were replaced by means, which gave a distinct improvement in fit. Then, as will be appreciated from the next diagram, increasing the tension by halving the smoothing factor to $F = 5$ then $F = 2.5$ gave very little subsequent improvement.

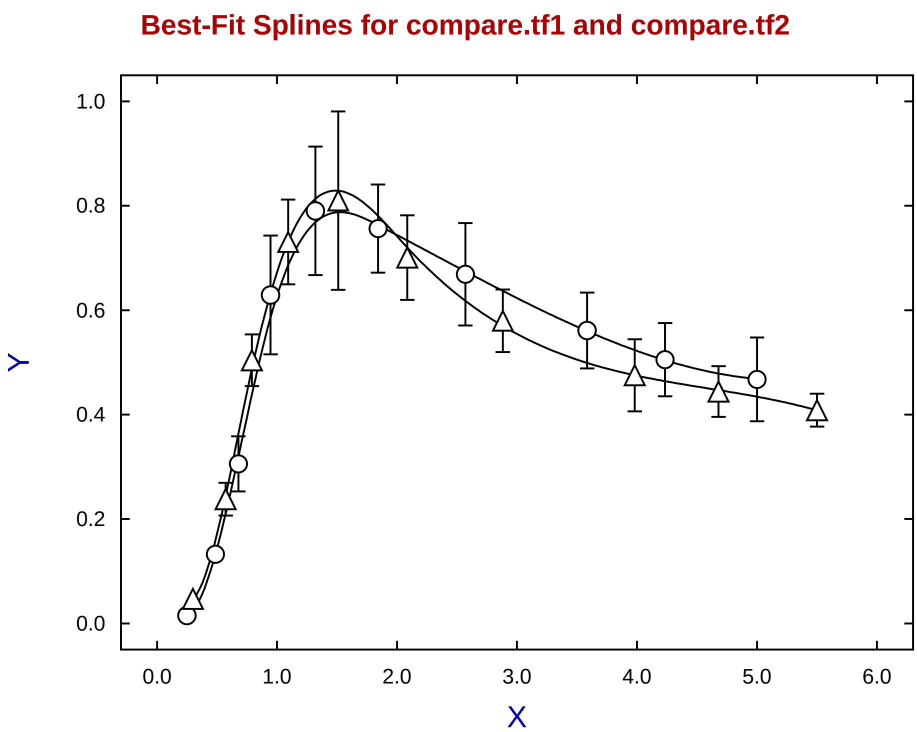


The following table summarizes the conclusion that, in this case, the trapezoidal estimate was very close to the area under the best-fit spline. Two figures are given for the percentage which depend on whether the absolute difference is scaled by the sum of areas or, perhaps more sensibly in some cases, by their average.

Area by trapezoidal rule (A)	2.7151
Area under best-fit spline (B)	2.7087
Absolute difference (C = |A - B|)	0.0063784
Fractional difference C/(A + B)	0.0012
Percent difference between A and B	0.1176% (denominator = sum)
Fractional difference C/[0.5(A + B)]	0.0024
Percent difference between A and B	0.2352% (denominator = average)

The next figures illustrate the comparison of data in test file `compare.tf1` and `compare.tf2`, first with original data, then with means, 95% confidence limits and best-fit splines.





It is clear that these data sets are very similar, and this is quantified by the next table.

Comparison of data sets and best-fit curves

Area under curve 1 (0.25 < x < 5.0) (A_1)	2.7087
Area under curve 2 (0.3 < x < 5.5) (A_2)	2.8351
For window number 1: 0.3 < x < 5.0, y_{min} = 0	
For window number 2: 0.3 < x < 5.0, y_{min} = 0.024346	
Area under curve 1 inside window 1 (B_1)	2.7077
Area under curve 2 inside window 1 (B_2)	2.6241
Integral of |curve 1 - curve 2| for the x-overlap (A_0)	0.20507
Area under curve 1 inside window 2 (C_1)	2.5933
Area under curve 2 inside window 2 (C_2)	2.5096

Estimated percentage differences between the curves

Over total range of x values: $100\lvert A_1 - A_2\rvert/(A_1 + A_2)$	2.2808%
In window 1 (with a zero baseline): $100(A_0)/(B_1 + B_2)$	3.8462%
In window 2 (with y_{min} baseline): $100(A_0)/(C_1 + C_2)$	4.0187%
Over total range of x values: $200\lvert A_1 - A_2\rvert/(A_1 + A_2)$	4.5616%
In window 1 (with a zero baseline): $200(A_0)/(B_1 + B_2)$	7.6924%
In window 2 (with y_{min} baseline): $200(A_0)/(C_1 + C_2)$	8.0374%
Conclusion: *Comparison of curves is good (likely to be identical)*	

Note that corrections may have to be made if the ranges of x are not identical for both data sets, if negative y values are encountered, or if the smallest y value is not zero, so these estimates have the following meanings.

1. **Areas under curves** A_1, A_2
   These are calculated for the best-fit spline curves over the ranges indicated without any corrections.

2. **Windows**
   These are defined as rectangles where the x ranges overlap and, if no y value is zero, or if any y values are negative, these are corrected by the parameter y_{min}. In window 1 $y_{min} = 0$, but in window 2 y_{min} is the smallest value that must be used to correct the y values to make sure the minimum y value is zero, and that all areas are for integration of nonnegative curves.

3. **Area under curves inside windows** B_1, B_2, C_1, C_2
   The values C_1, C_2 are corrected, if required, depending on y_{min}.

4. **Integral of absolute difference**
   This is evaluated by using Simpson's rule with the integrand defined as the absolute value of the difference between curves, but only over the range of x-overlap.

5. **100 or 200 in percentage calculations**
   100 is used to refer to the sum of areas in the denominator so that the value cannot exceed 100%, whereas 200, which refers to the average of areas in the denominator, can cause confusion as it can be as large as 200%.

6. **Conclusion**
   This is an arbitrary qualitative decision based on considering all the values in the above table.

The most useful values from this table are the last two giving the percentage difference between the curves with a zero baseline and when using a baseline shift, and also using the average of the two areas in the denominator when calculating the ratios, which is preferred when $B_1 \approx B_2$ and $C_1 \approx C_2$.

Archiving spline files

When a best-fit spline has been calculated the knots and coefficients can be saved to a spline file. This can be used during the running of program **compare** or can be input into program **spline** for retrospective analysis, such as using as a standard curve for calibration.