NA3 –Rayleigh Fit

My Rayleigh Fit procedure:

Let the lidar signal be:

$$P(r)r^2 \propto \left(\beta_R(r)+\beta_P(r)\right)\exp\left[-2\int_0^r\left(\alpha_R(r')+\alpha_P(r')\right)dr'\right]$$

1. Select a range in the lidar signal where clean air can be assumed (**rmax**, **rmin**).
2. Caclulate the $ß_R^{attn}$ using a "good" radiosonde, with attenuation starting at the middle rangebin (r_0, **reference range**) of the selected range (**rmin**, **rmax**).

$$\beta_R^{attn}(r,r_0)=\beta_R(r)\exp\left[-2\int_{r_0}^r\alpha_R(r')dr'\right]$$

This means negative attenuation for **r** < **r0**, and keeps the exact reference value at the reference range **r0**

$$\beta_R^{attn}(r_0,r_0)=\beta_R(r_0)$$

3. Check whether the fit is sufficiently good. A general procedure to evaluate the "goodness of fit" is under development. Ideas are welcome.
4. If the fit is not good, repeat 1. to 3. until it is good.
5. Normalize the lidar signal to the $ß_R^{attn}$ using the means of the $ß_R^{attn}$ and of the lidar signal over the fit range. This avoids an additional error due to signal noise in the fit range.

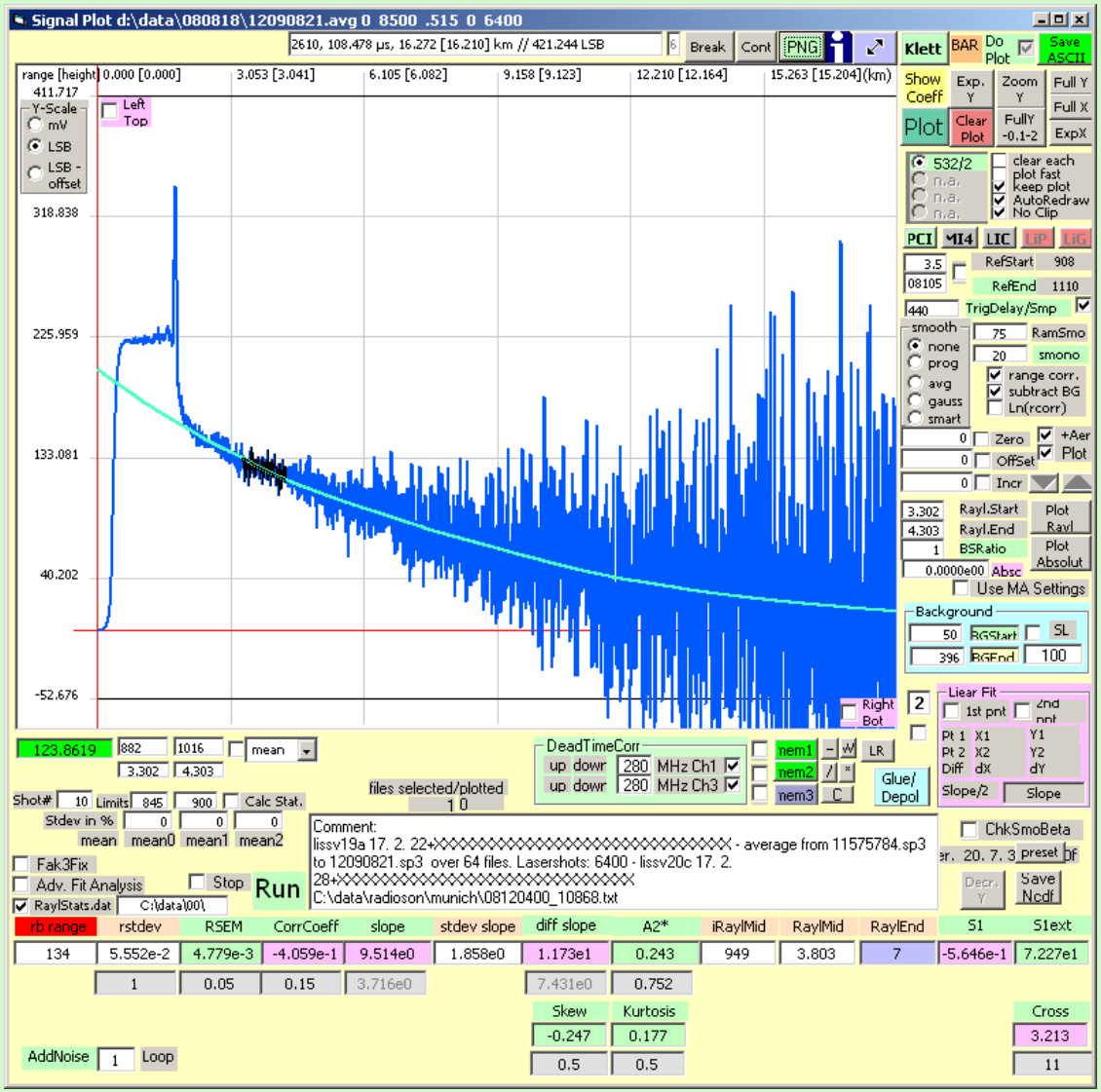
$$P^{norm}(r,r_0)r^2=P(r)r^2\frac{\int_{r\min}^{r\max}\beta_R^{attn}(r,r_0)dr}{\int_{r\min}^{r\max}P(r')r'^2dr}$$

6. Replace the value of the lidar signal at the middle rangebin **r0** with the value of the Rayleigh backscatter coefficient at this rangebin $ß_R$(**r0**). Note: this value should be the same as the one of the $ß_R^{attn}$ at this rangebin, i.e. the right reference value for Fernald/Klett.

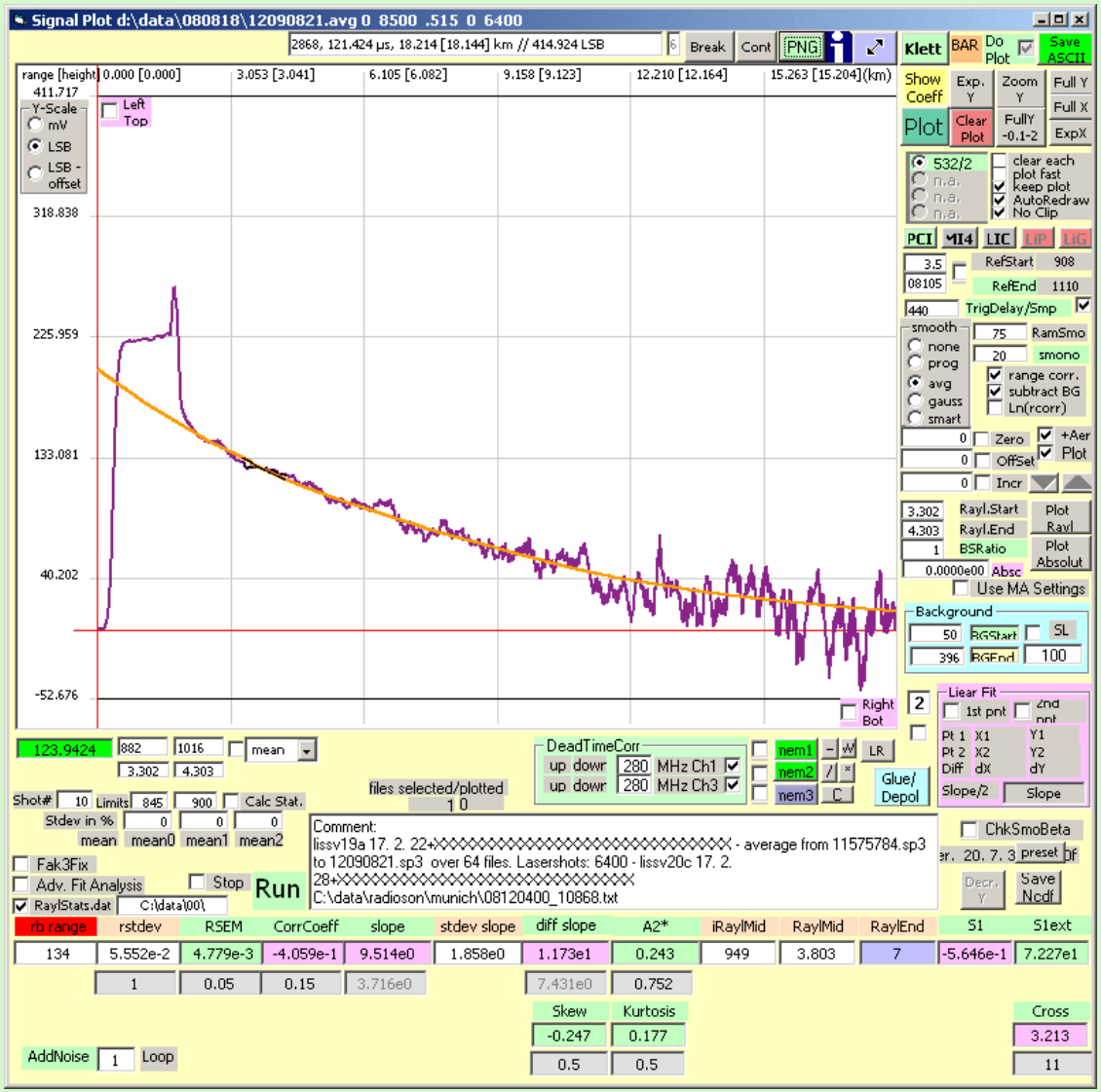
$$P^{norm}(r_0,r_0)r^2=\beta_R^{attn}(r_0,r_0)=\beta_R(r_0)$$

7. Start the Fernald/Klett inversion from this rangebin.

NA3 – But: how good is a Rayleigh Fit

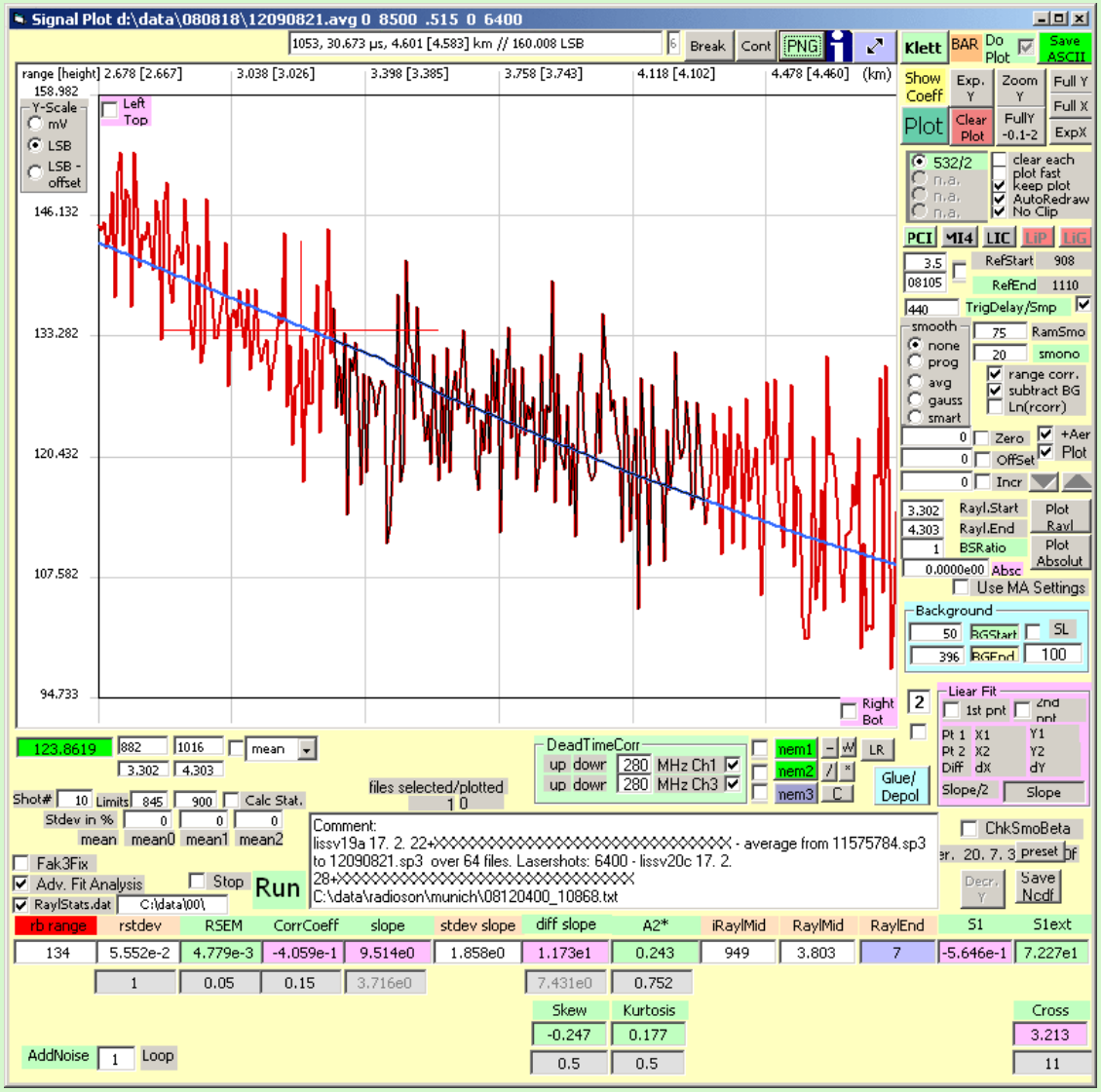


NA3 – But: how good is a Rayleigh Fit



NA3 – But: how good is a Rayleigh Fit - RSEM

Relative Standard Error of the mean (RSEM)



If noise is normal (Gaussian) distributed

Sample Standard Deviation (s)

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2},$$

Standard Error of the mean

SEM = $$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Relative Standard Error of the mean

$$RSEM = \frac{SEM}{\overline{x}}$$

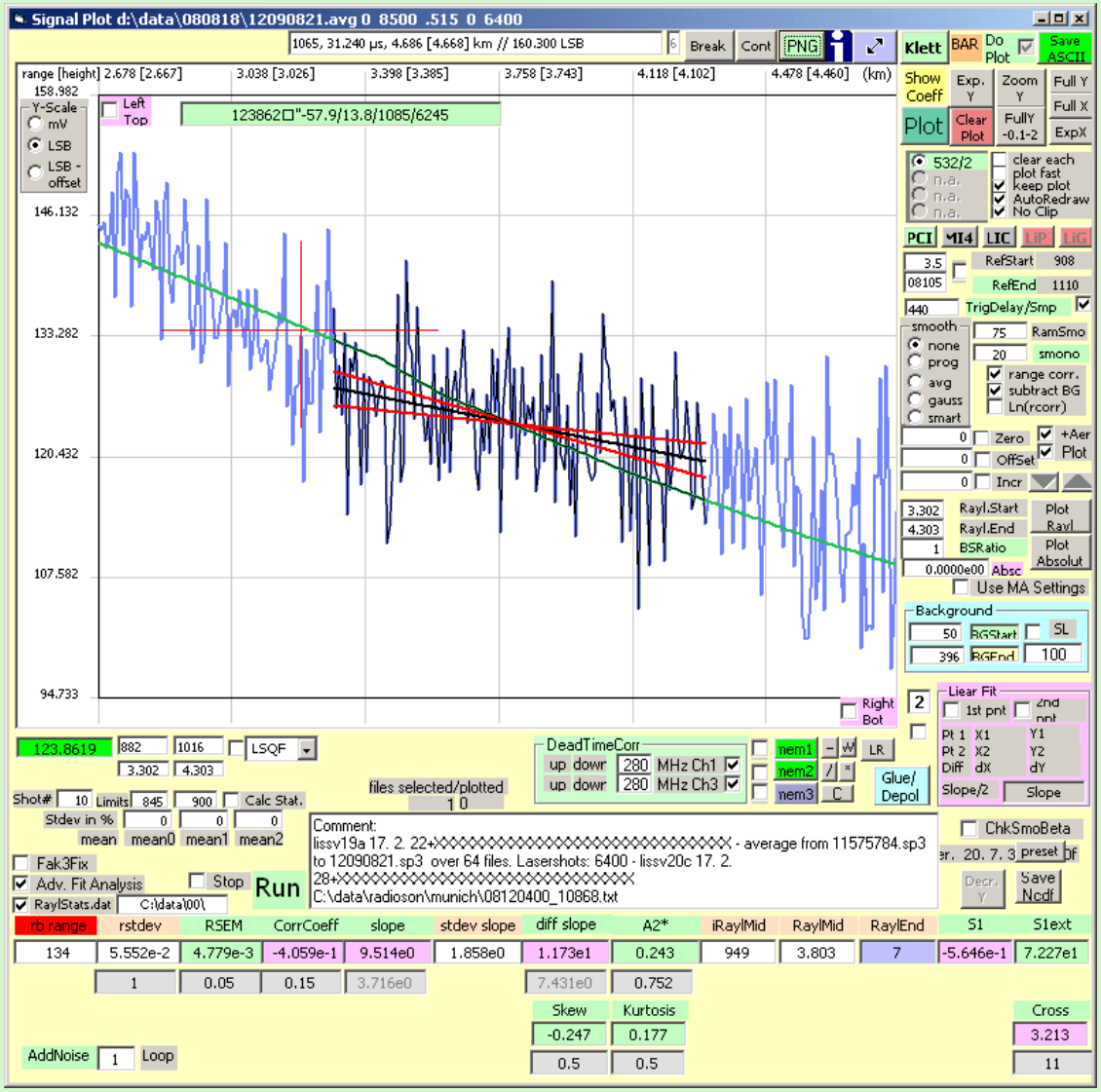
n = 134
s = 5.6%
RSEM = 0.48%

http://en.wikipedia.org/wiki/Unbiased_estimation_of_standard_deviation
http://en.wikipedia.org/wiki/Standard_deviation
http://en.wikipedia.org/wiki/Standard_error_(statistics)

68% confidence limits, $\hat{\beta} \pm \sigma$
95% confidence limits, $\hat{\beta} \pm 2\sigma$
99% confidence limits, $\hat{\beta} \pm 2.5\sigma$

NA3 – But: how good is a Rayleigh Fit - LSQF

Least square fit of a straight line



LSQF gives additionally
- Slope of the signal
- Standard uncertainty of the slope

⇒ The signal slope does not fit the slope of the Rayleigh signal ± 2 s

⇒ Possible reasons

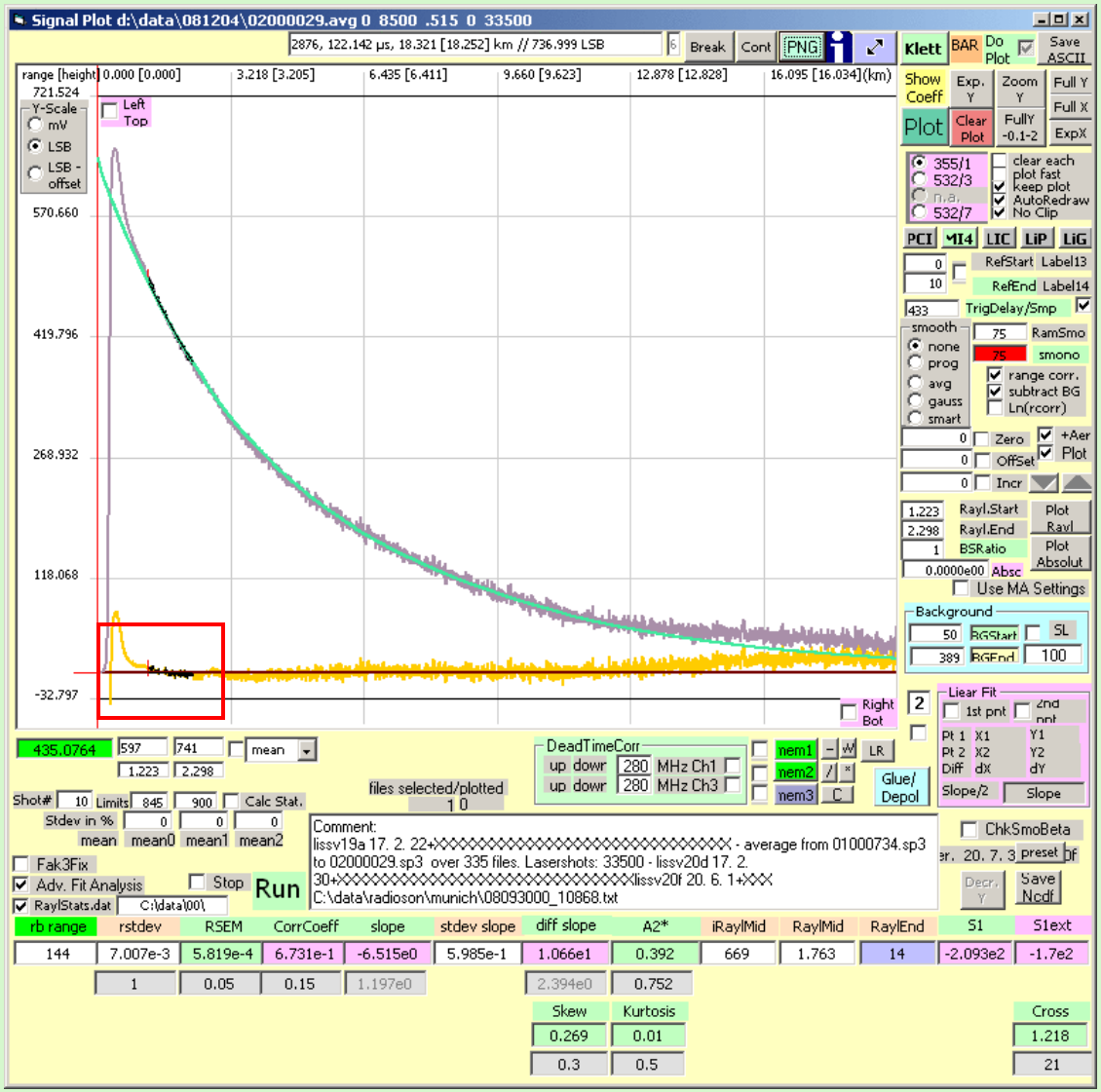
1. Not Rayleigh
2. Signal distortions

⇒ Conclusion

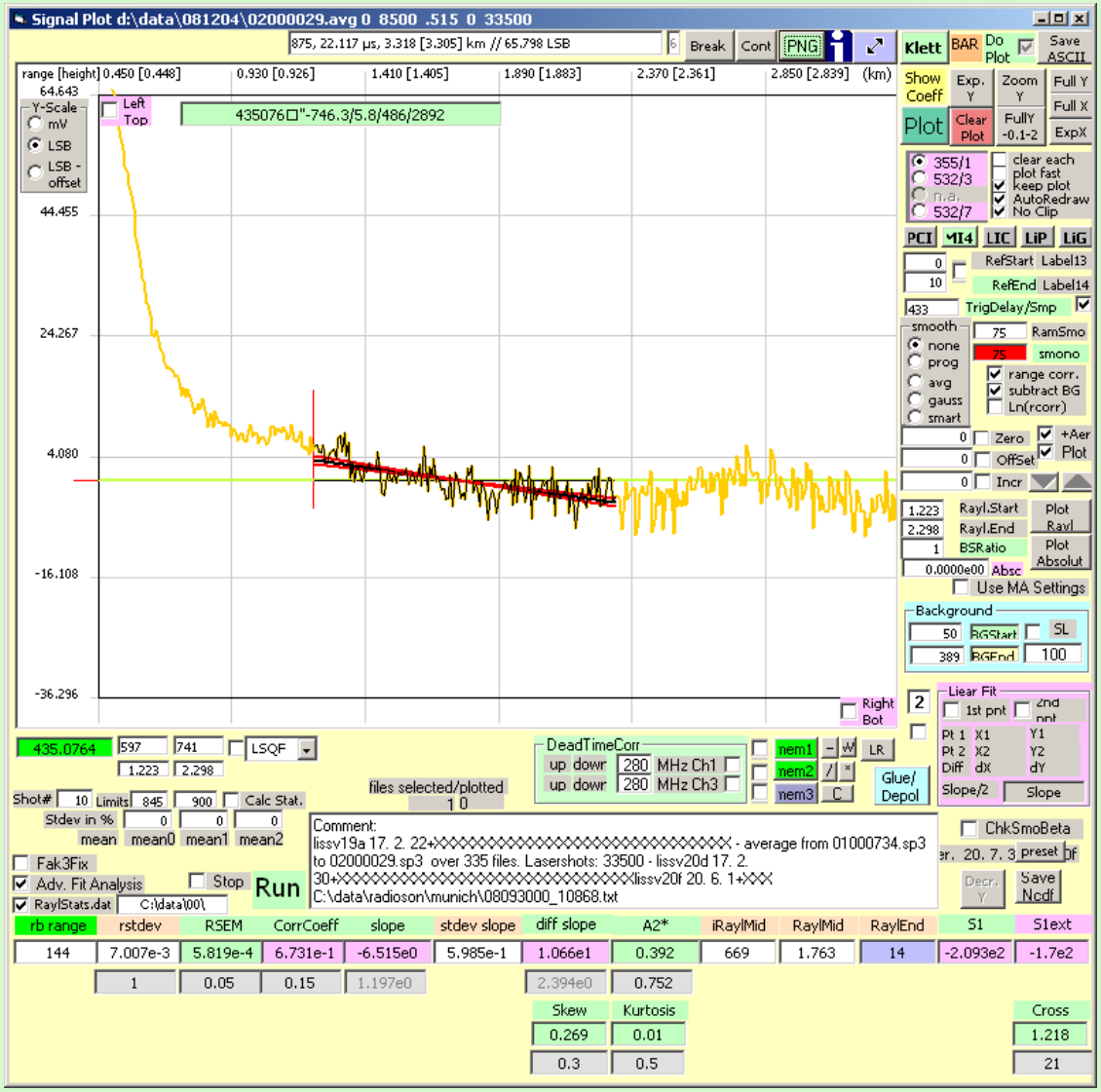
Expand the fit range
Search for another range

http://en.wikipedia.org/wiki/Unbiased_estimation_of_standard_deviation
http://en.wikipedia.org/wiki/Standard_deviation
http://en.wikipedia.org/wiki/Standard_error_(statistics)

Rayleigh Fit Criterions – Residual signal



Rayleigh Fit Criterions – Residual signal analysis – slope criterion



The slope of the linear fit to the residuals must be smaller than 2 * sigma of the fit slope

NA3 – Signal Statistics – Linear Least Squares Fit

Theory: http://en.wikipedia.org/wiki/Linear_regression
http://en.wikipedia.org/wiki/Least-squares_estimation_of_linear_regression_coefficients
good: http://www.mpi-hd.mpg.de/astrophysik/HEA/internal/Numerical_Recipes/f15-2.pdf
(official link) http://www.nrbook.com/a/bookcpdf.php

Programming: take the code from the Numerical Recipies
http://www.mpi-hd.mpg.de/astrophysik/HEA/internal/Numerical_Recipes/f15-2.pdf

Results:

Given a set of data points x(1:ndata),y(1:ndata) with individual standard deviations sig(1:ndata), fit them to a straight line y = a + bx by minimizing χ2. Returned are a,b and their respective probable uncertainties siga and sigb, the chi-square chi2, and the goodness-of-fit probability q (that the fit would have χ2 this large or larger). If mwt=0 on input, then the standard deviations are assumed to be unavailable: q is returned as 1.0 and the normalization of chi2 is to unit standard deviation on all points.

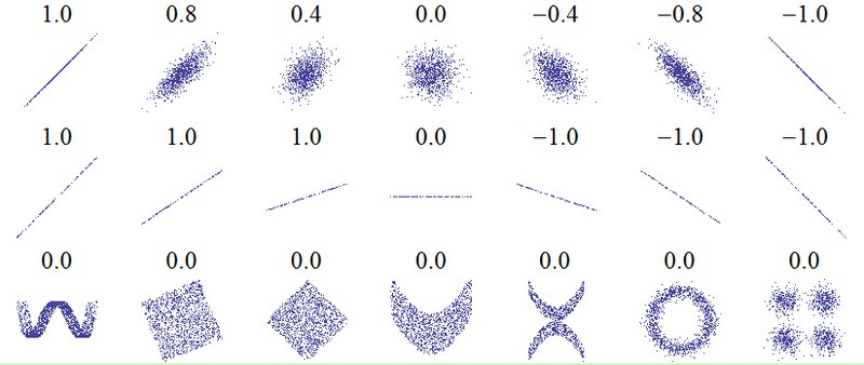
What does that mean: "*normalization of chi2 is to unit standard deviation on all points*"?
=> If you don't give samples individual SigmaY_i, the code calculates the uncertainty from the standard deviation of all Y_i . (This is what we do) That means, SigmaYi is the same for all rangebins.

What does that mean: *"probable uncertainties siga and sigb" ?*
=> These are the standard deviations in the estimates of a and b

NA3 – Signal Statistics – Correlation Coefficient

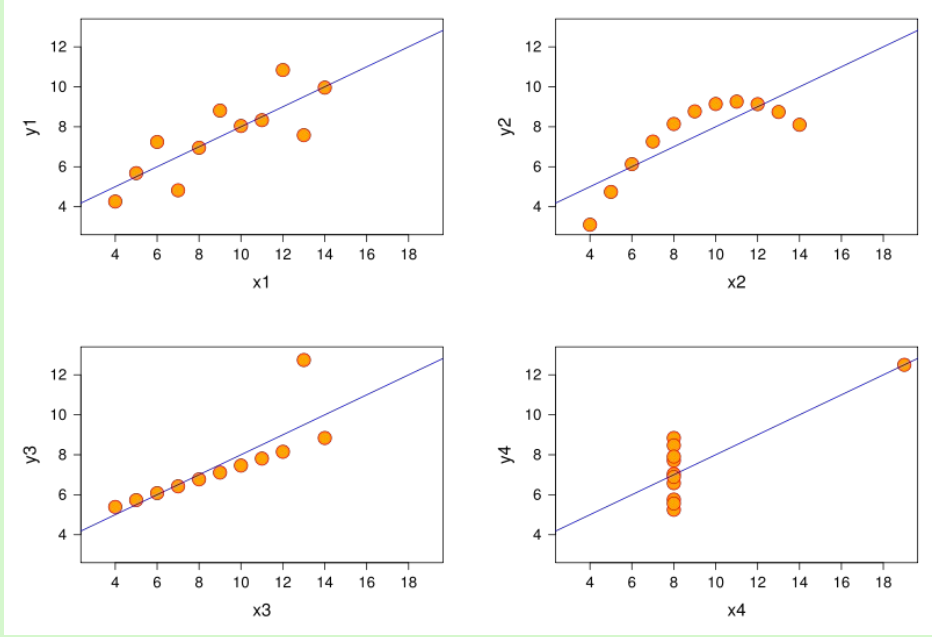
$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X \sigma_Y},$$

The correlation coefficient $\rho_{X,Y}$ between two random variables X and Y with expected values μ_X and μ_Y and standard deviations σ_X and σ_Y is defined as:



Several sets of (*x*, *y*) points, with the correlation coefficient of *x* and *y* for each set. Note that the correlation reflects the noisiness and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). N.B.: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of *Y* is zero.

http://en.wikipedia.org/wiki/Correlation



Four sets of data with the same correlation of 0.81

http://en.wikipedia.org/wiki/File:Anscombe.svg

relative easy to programm

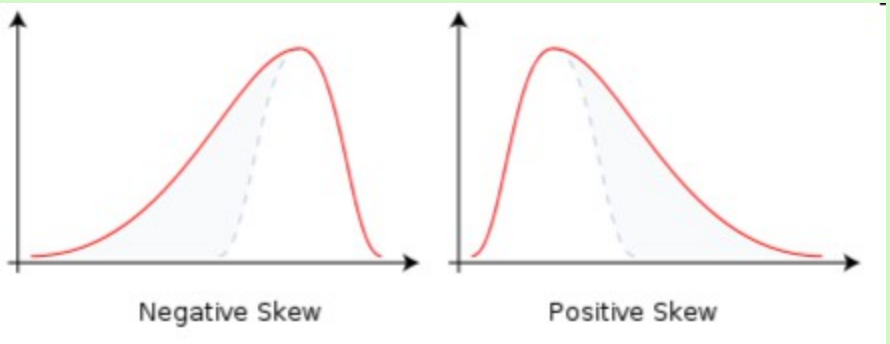
NA3- Signal Statistics - Skewness

For a sample of n values the *sample skewness* is (right image), where x_i is the i^{th} value, is the sample mean, m_3 is the sample third central moment, and m_2 is the sample variance.

$$g_1 = \frac{m_3}{m_2^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^{3/2}},$$

Given samples from a population, the equation for the sample skewness g_1 above is a biased estimator of the population skewness. The usual estimator of skewness is (right image).

$$G_1 = \frac{k_3}{k_2^{3/2}} = \frac{\sqrt{n\,(n-1)}}{n-2}\,g_1,$$



relative easy to programm

http://en.wikipedia.org/wiki/Skewness

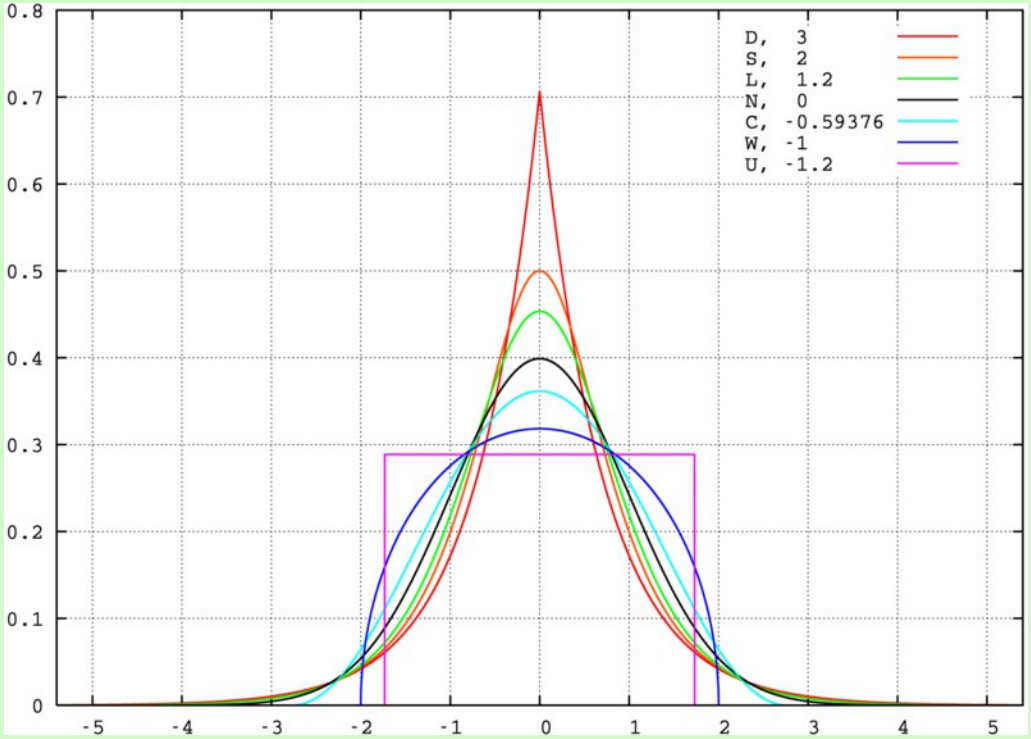
NA3- Signal Statistics - Kurtosis

For a sample of n values the **sample kurtosis** is (right image) where m_4 is the fourth sample moment about the mean, m_2 is the second sample moment about the mean (that is, the sample variance), x_i is the i^{th} value, and is the sample mean.

$$g_2 = \frac{m_4}{m_2^2} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^2} - 3$$

Given a sub-set of samples from a population, the sample kurtosis above is a biased estimator of the population kurtosis. The usual estimator of the population kurtosis (used in DAP/SAS, Minitab, PSPP/SPSS, and Excel but not by BMDP) is G_2, defined as follows:

$$G_2 = \frac{n-1}{(n-2)(n-3)}\left((n+1)\,g_2 + 6\right)$$



Probability (signal intensitiy) distribution around mean

Kurtosis of different probability distributions

- **red, kurtosis 3, Laplace (D)ouble exponential distribution**
- **orange, kurtosis 2, hyperbolic (S)ecant distribution**
- **green, kurtosis 1.2, (L)ogistic distribution**
- **black, kurtosis 0, (N)ormal distribution**
- **cyan, kurtosis −0.593762..., raised (C)osine distribution**
- **blue, kurtosis −1, (W)igner semicircle distribution**
- **magenta, kurtosis −1.2, (U)niform distribution**

relative easy to programm

http://en.wikipedia.org/wiki/Kurtosis

NA3 Signal Statistics – Anderson-Darling-Test

The **Anderson–Darling test** is a form of minimum distance estimation, and one of the most powerful statistics for detecting most departures from normality. The Anderson–Darling test assesses whether a sample comes from a specified distribution. The formula for the test statistic A to assess if data $Y_1 < .... < Y_N$ (note that the data must be put in order) comes from a distribution with cumulative distribution function (CDF) F is

A2 = – N – S where S is $S = \sum_{k=1}^{N} \frac{2k-1}{N} \left[\ln F(Y_k) + \ln\left(1 - F(Y_{N+1-k})\right)\right]$

Procedure

(If testing for normal distribution of the variable X)

1) The data X_i, for $i = 1, \ldots n$, of the variable X that should be tested is sorted from low to high.

2) The mean $\bar{X}$ and standard deviation s are calculated from the sample of X.

3) The values X_i are standardized as

$$Y_i = \frac{X_i - \bar{X}}{s}$$

4) With the standard normal CDF Φ, A^2 is calculated using

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)(\ln \Phi(Y_i) + \ln(1 - \Phi(Y_{n+1-i})))$$

or without repeating indices as

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}\left[(2i-1)\ln \Phi(Y_i) + (2(n-i)+1)\ln(1-\Phi(Y_i))\right].$$

5) A^{*2}, an approximate adjustment for sample size, is calculated using

$$A^{*2} = A^2\left(1 + \frac{0.75}{n} + \frac{2.25}{n^2}\right)$$

6) If A^{*2} exceeds 0.752 then the hypothesis of normality is rejected for a 5% level test.

Note:

1. If $s = 0$ or any $\Phi(Y_i) = (0 \text{ or } 1)$ then A^2 cannot be calculated and is undefined.

relative easy to programm

but
- Yi must be sorted
- the normal CDF is needed.

See next page

See also
- Kolmogorov–Smirnov test
- Shapiro–Wilk test
- Smirnov–Cramér–von-Mises test
- Jarque–Bera test
- Goodness of fit
- US NIST Handbook of Statistics

http://en.wikipedia.org/wiki/Anderson-Darling_test

NA3 - Anderson-Darling-Test Limits

1. ' ANDERSON-DARLING 1-SAMPLE TEST
2. ' THAT THE DATA CAME FROM A NORMAL DISTRIBUTION

3. ' 2. CRITICAL VALUES:
4. ' 90 % POINT = 0.631
5. ' 95 % POINT = 0.752
6. ' 97.5 % POINT = 0.837
7. ' 99 % POINT = 1.035

8. from http://statisticalengineering.com/goodness.htm

9. 'Actually, the null hypothesis to be tested is that the POPULATION can be adequately modeled with a normal distribution.
10. 'If A2 * exceeds 0.752 then the hypothesis of normality is rejected for a 5% level test.

11. See also
    1. http://www.itl.nist.gov/div898/software/dataplot/homepage.htm (Dataplot: free statistical software)
    2. http://www.itl.nist.gov/div898/handbook/index.htm (NIST statistical handbook)
    3. http://www.itl.nist.gov/div898/handbook/eda/section3/eda35e.htm (Anderson_Darling_Test)

NA3 - Anderson-Darling-Test Limits

```
1.    ' ANDERSON-DARLING 1-SAMPLE TEST
2.    '            THAT THE DATA CAME FROM A NORMAL DISTRIBUTION

3.    '    2. CRITICAL VALUES:
4.    '       90        % POINT    = 0.631
5.    '       95        % POINT    = 0.752
6.    '       97.5      % POINT    = 0.837
7.    '       99        % POINT    = 1.035

8.    from http://statisticalengineering.com/goodness.htm

9.    'Actually, the null hypothesis to be tested is that the POPULATION can be adequately modeled with a normal distribution.
10.   'If A2 *  exceeds 0.752 then the hypothesis of normality is rejected for a 5% level test.


11.   See also
        1.    http://www.itl.nist.gov/div898/software/dataplot/homepage.htm   (Dataplot: free statistical software)
        2.    http://www.itl.nist.gov/div898/handbook/index.htm   (NIST statistical handbook)
        3.    http://www.itl.nist.gov/div898/handbook/eda/section3/eda35e.htm   ( Anderson_Darling_Test)
```

NA3 Quality Assurance – Sort and normal CDF

Yi must be sorted

=> search the www for aome code, e.g.

=> VB6 MedianThreeQuickSort1(Y) from sorting algorithms: http://www.vbforums.com/showthread.php?t=473677)

Copied at the end of this talk

normal CDF Φ is needed

can be caclulated from Error Function Erf2 : $\Phi(z) = 0.5 * (1.0 + Erf2(z / sqrt(2.0)))$

http://en.wikipedia.org/wiki/Error_function

http://www.cs.princeton.edu/introcs/21function/ErrorFunction.java.html

// fractional error less than x.xx * 10 ^ -4. // Algorithm 26.2.17 in Abromowitz and Stegun, Handbook of Mathematical.

(I can confirm this accuracy from comparision with a even better approximation of the error function.

I transferred the code from Java to VisualBasic6)

```
Public Function Erf2(z As Double) As Double
    Dim t As Double
    Dim poly As Double
    Dim ans As Double

    t = 1 / (1# + 0.47047 * Abs(z))
    poly = t * (0.348024 + t * (-0.0958798 + t * 0.7478556))
    ans = 1# - poly * Exp(-z * z)
    If z >= 0 Then
        Erf2 = ans
    Else
        Erf2 = -ans
    End If
End Function
```

relative easy to copy

http://en.wikipedia.org/wiki/Anderson-Darling_test

NA3 Quality Assurance – Quicksort (Visual Basic 6)

```
1.    ' Omit plngLeft & plngRight; they are used internally during recursion
2.    ' there was an error in the code using a wrong variable => corrected in red
3.    Public Sub MedianThreeQuickSort1(ByRef pvarArray As Variant, Optional ByVal plngLeft As Long, Optional ByVal plngRight As Long)
4.       Dim lngFirst As Long
5.       Dim lngLast As Long
6.       Dim varMid As Variant
7.       Dim lngIndex As Long
8.       Dim varSwap As Variant
9.       Dim a As Long
10.      Dim b As Long
11.      Dim c As Long
12.
13.      If plngRight = 0 Then
14.         plngLeft = LBound(pvarArray)
15.         plngRight = UBound(pvarArray)
16.      End If
17.      lngFirst = plngLeft
18.      lngLast = plngRight
19.      lngIndex = plngRight - plngLeft + 1
20.      a = Int(lngIndex * Rnd) + plngLeft
21.      b = Int(lngIndex * Rnd) + plngLeft
22.      c = Int(lngIndex * Rnd) + plngLeft
23.      If pvarArray(a) <= pvarArray(b) And pvarArray(b) <= pvarArray(c) Then
24.         lngIndex = b
25.      Else
26.         If pvarArray(b) <= pvarArray(a) And pvarArray(a) <= pvarArray(c) Then
27.            lngIndex = a
28.         Else
29.            lngIndex = c
30.         End If
31.      End If
32.      varMid = pvarArray(lngIndex)
33.      Do
34.         Do While pvarArray(lngFirst) < varMid And lngFirst < plngRight
35.            lngFirst = lngFirst + 1
36.         Loop
37.         Do While varMid < pvarArray(lngLast) And lngLast > plngLeft
38.            lngLast = lngLast - 1
39.         Loop
40.         If lngFirst <= lngLast Then
41.            varSwap = pvarArray(lngFirst)
42.            pvarArray(lngFirst) = pvarArray(lngLast)
43.            pvarArray(lngLast) = varSwap
44.            lngFirst = lngFirst + 1
45.            lngLast = lngLast - 1
46.         End If
47.      Loop Until lngFirst > lngLast
48.      If lngLast – plngLeft < plngRight – lngFirst Then
49.         If plngLeft < lngLast Then MedianThreeQuickSort1 pvarArray, plngLeft, lngLast
50.         If lngFirst < plngRight Then MedianThreeQuickSort1 pvarArray, lngFirst, plngRight
51.      Else
52.         If lngFirst < plngRight Then MedianThreeQuickSort1 pvarArray, lngFirst, plngRight
53.         If plngLeft < lngLast Then MedianThreeQuickSort1 pvarArray, plngLeft, lngLast
54.      End If
55.   End Sub
```

http://www.vbforums.com/showpost.php?p=2909260&postcount=14

relative easy to copy

NA3 Quality Assurance – Anderson Darling Test VB6 code

```
1.   Public Function ADT(ByRef X() As Double, N As Long, Optional ByVal m As Double, Optional ByVal s As Double) As Double
2.   ' Anderson Darling test according to http://en.wikipedia.org/wiki/Anderson-Darling_test
3.   ' X ist ein Datenarray von Länge N mit der gesasmt standartabw S und dem Mittelwert M
4.   Dim i As Integer
5.   Dim Y() As Double
6.   Dim ifile As Integer
7.   Dim Psi() As Double
8.   'Dim Psi3 As Double
9.   Dim help As Double
10.  ReDim Y(1 To N)
11.  ReDim Psi(1 To N)
12.  For i = 1 To N
13.      Y(i) = X(iRaylStart + i - 1) / s
14.  Next i

15.  ' Aufsteigend Sortieren
16.  Call MedianThreeQuickSort1(Y)
17.     ADT = 0
18.
19.     For i = 1 To N
20.        'Psi = 0.5* Erfcc(-Y(i) * DW2)
21.        Psi(i) = 0.5 * (1 + Erf2(Y(i) * DW2))
22.        If Psi(i) = 0 Then Psi(i) = 0.001
23.        If Psi(i) = 1 Then Psi(i) = 0.999
24.     Next i
25.
26.     For i = 1 To N
27.        help = help + (2 * i - 1) * (Log(Psi(i)) + Log(1 - Psi(N - i + 1)))    ' in visual basic Log is the logarithm to the basis e
28.     Next i
29.
30.     ADT = (-N - help / N) * (1 + 0.75 / N + 2.25 / N ^ 2)
31.
32.  End Function
```

http://www.vbforums.com/showpost.php?p=2909260&postcount=14

relative easy to copy

Rayleigh Fit Criterions

1. Make a Rayleigh Fit to the raw, unsmoothed (!) range corrected lidar signal and calculate the differences (residuals; statistics in the signal is different from that in the residuals)
2. Calculate from the residual signal over an appropriate Rayleigh fit range (~1km):

Local estimators

- Is noise normal distributed? => Anderson-Darling-Test (ADT), Kurtosis, Skewness..... a prerequisit for all statistical calculations, which are based on the assumtion of normal noise distribution.
- Slope: local slope must not "significantly" deviate from Zero (i.e. the Rayleigh signal).
- Correlation coefficient (?): (seems to give similar information as slope)
- Curvature => Fit to polynomial (at least quadratic, advaced programming), or
- differential slope DiffSlope between first and second half of the fir range
- Relative standard error of the mean (RSEM):

Global estimator

- Lowest Good Fit of all (LGF): .....
- Cross: below the fitting range, residuals may not be smaller than Zero (regarding local noise)

Rayleigh Fit Criterions

use available or simple programming

search for other useful estimators
determine minimal set of estimators to best characterise the fit
determine useful limits for each estimator

fit quality (depends on effect on the accuracy of the scattering coefficients):
determine best local fit with a minimum fit length (maybe 1 km ?)
where can pure Rayleigh be assumed? => statistical estimators sufficient? How much uncertainty?

Determination of the optical quality of the lidar system

1. use best fit on a perfect day/night
2. calc relative residuals = (Pr2 – Rayl) / Rayl
3. => relative systematic error of determination of the reference values for Fernald/Klett- and Raman retrivals

Literature, References

Literature

http://www.itl.nist.gov/div898/handbook/index.htm

http://en.wikipedia.org/wiki/Statistics

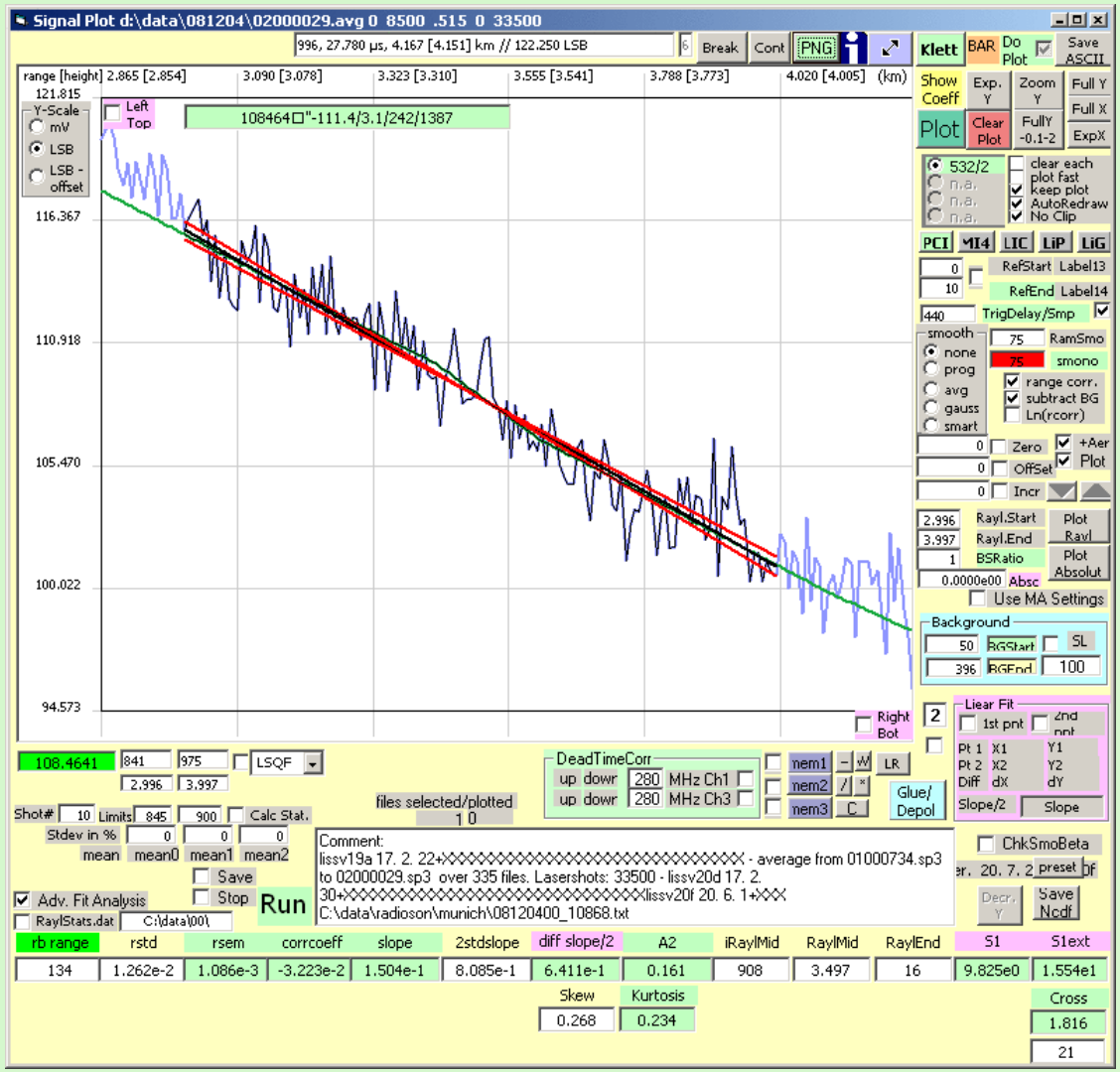
http://en.wikipedia.org/wiki/Errors_and_residuals_in_statistics

Numerical Recipies 15.2 Fitting Data to a Straight Line

(official link) http://www.nrbook.com/a/bookcpdf.php
(inofficial link) http://www.mpi-hd.mpg.de/astrophysik/HEA/internal/Numerical_Recipes/f15-2.pdf

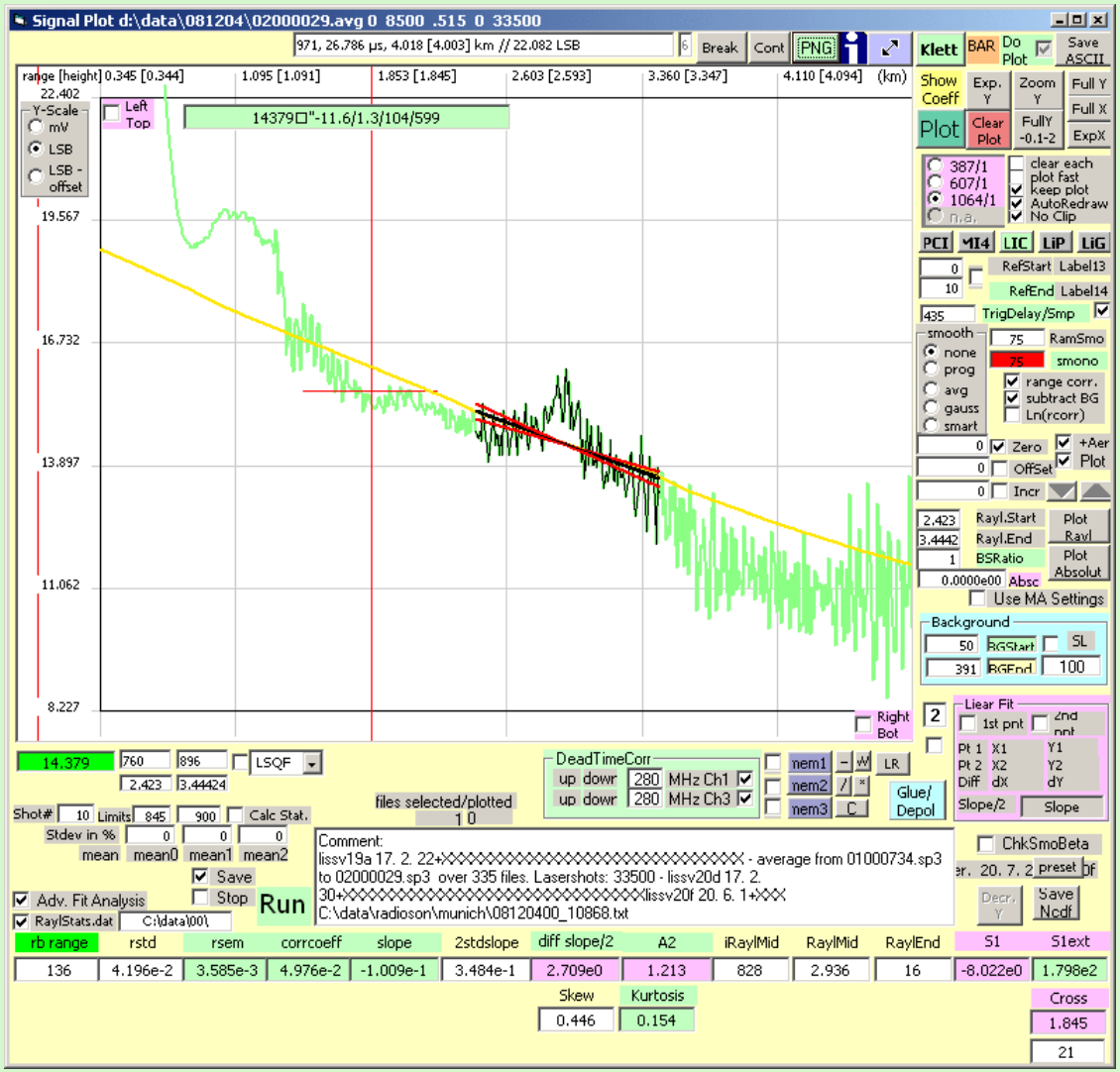
Rayleigh fit - slope criterion, correlation coefficient

The slope of the linear fit to the residuals must be smaller than 2 * sigma of the fit slope



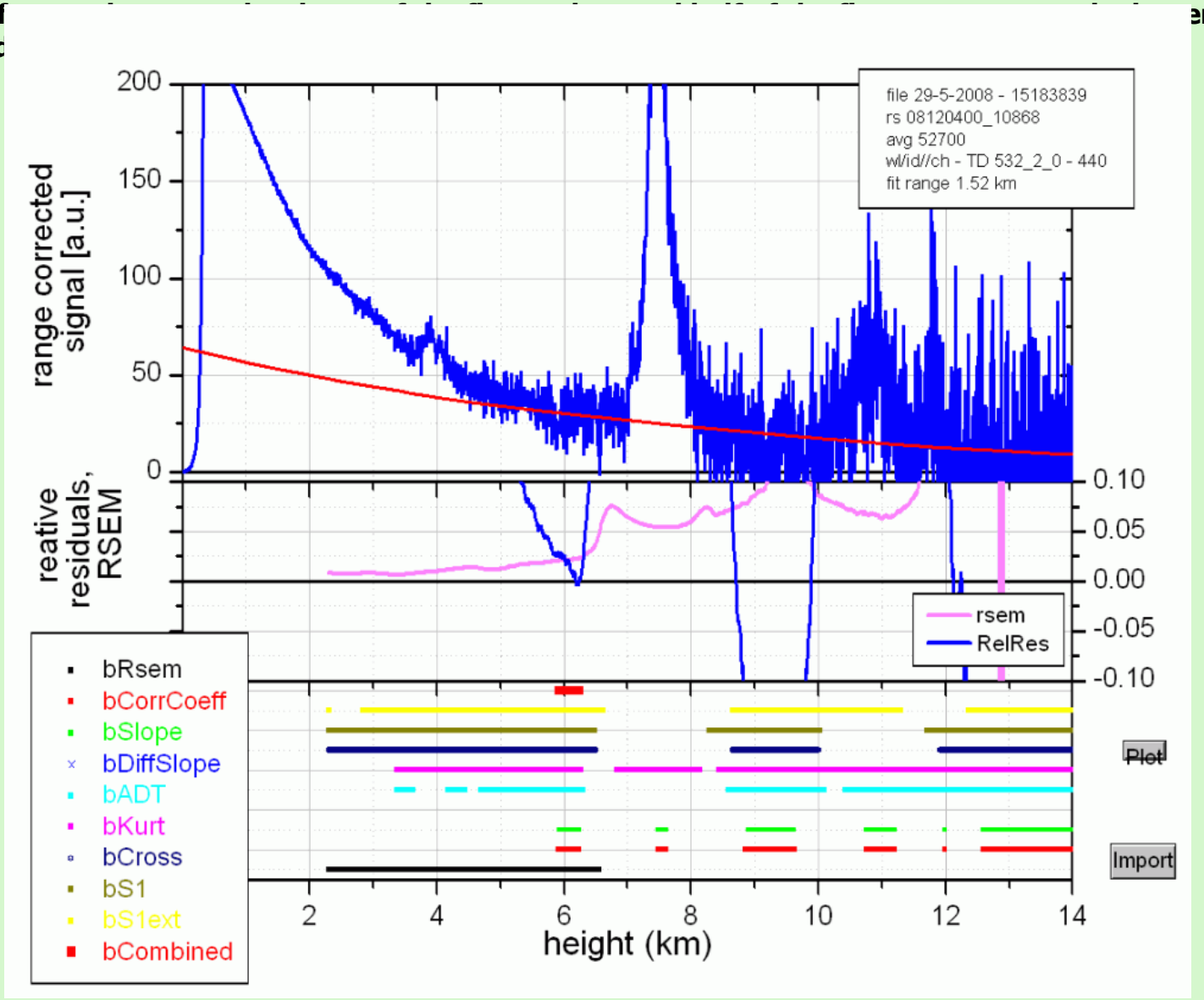
Rayleigh fit – curvature / differential slope criterion

The difference between the slopes of the first and second half of the fit range must not be larger than allowed from statistics



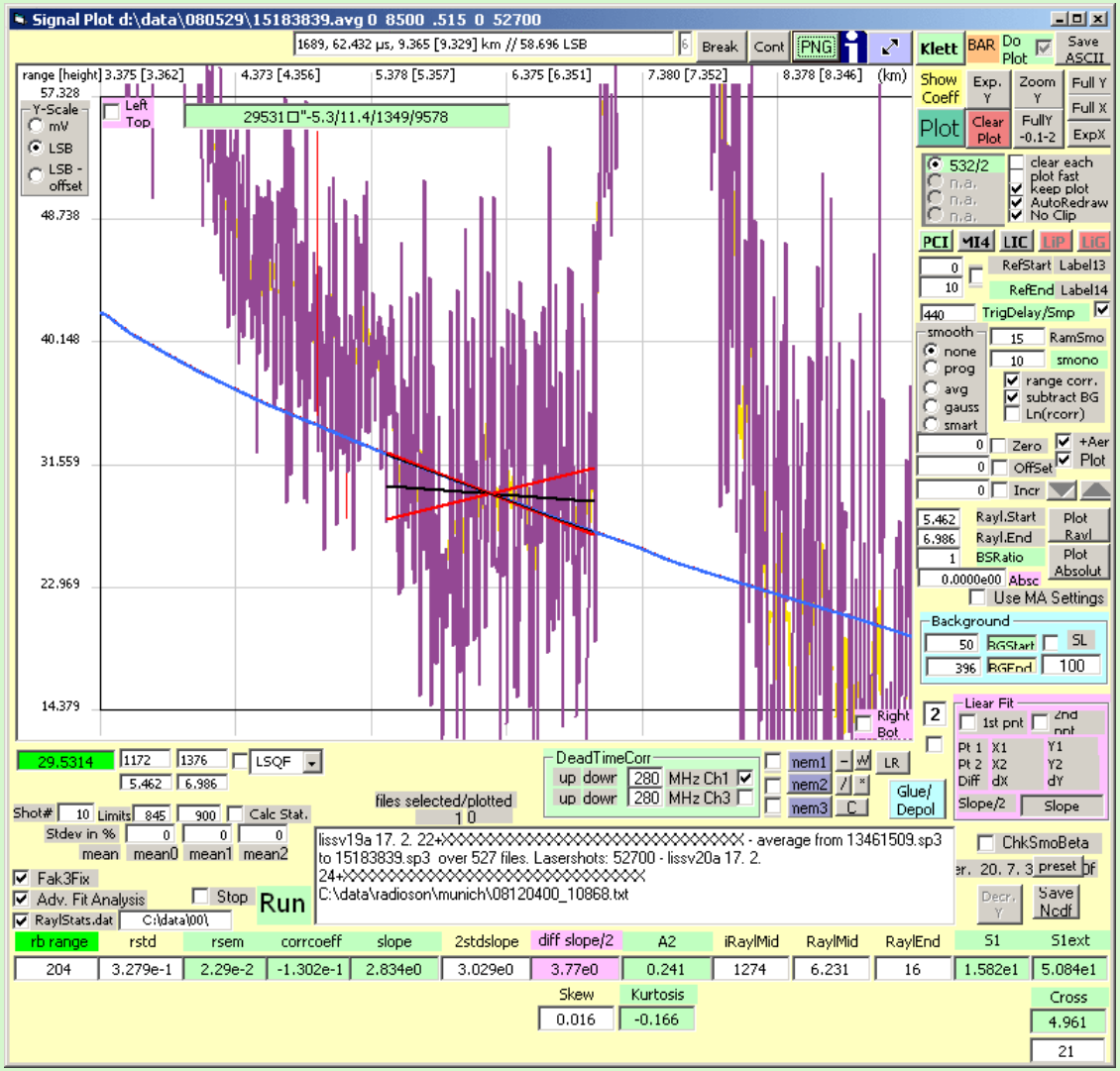
Rayleigh fit – curvature / differential slope criterion

The diff … er than allowed



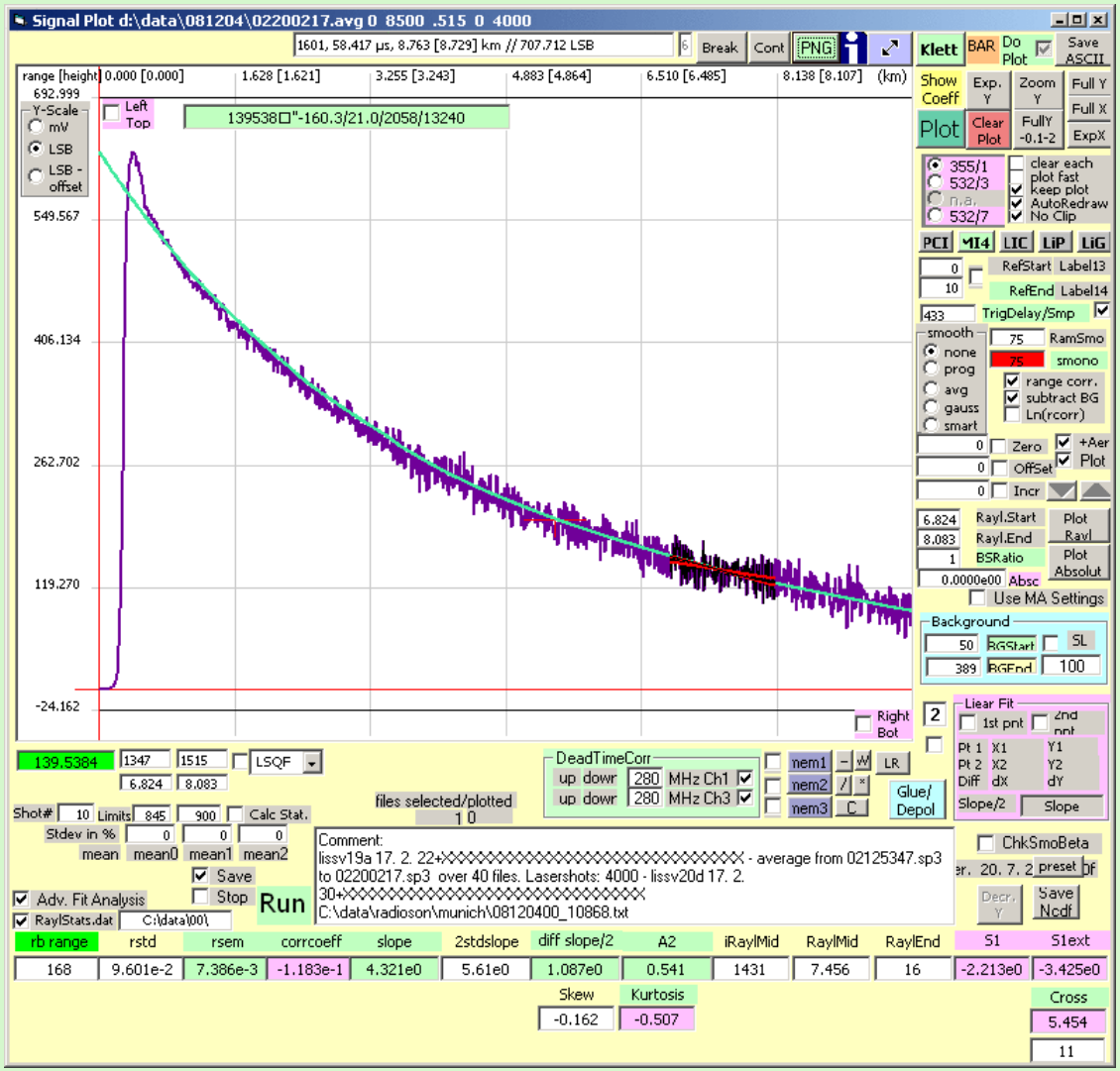
Rayleigh fit – curvature / differential slope criterion

The difference between the slopes of the first and second half of the fit range must not be larger than allowed from statistics



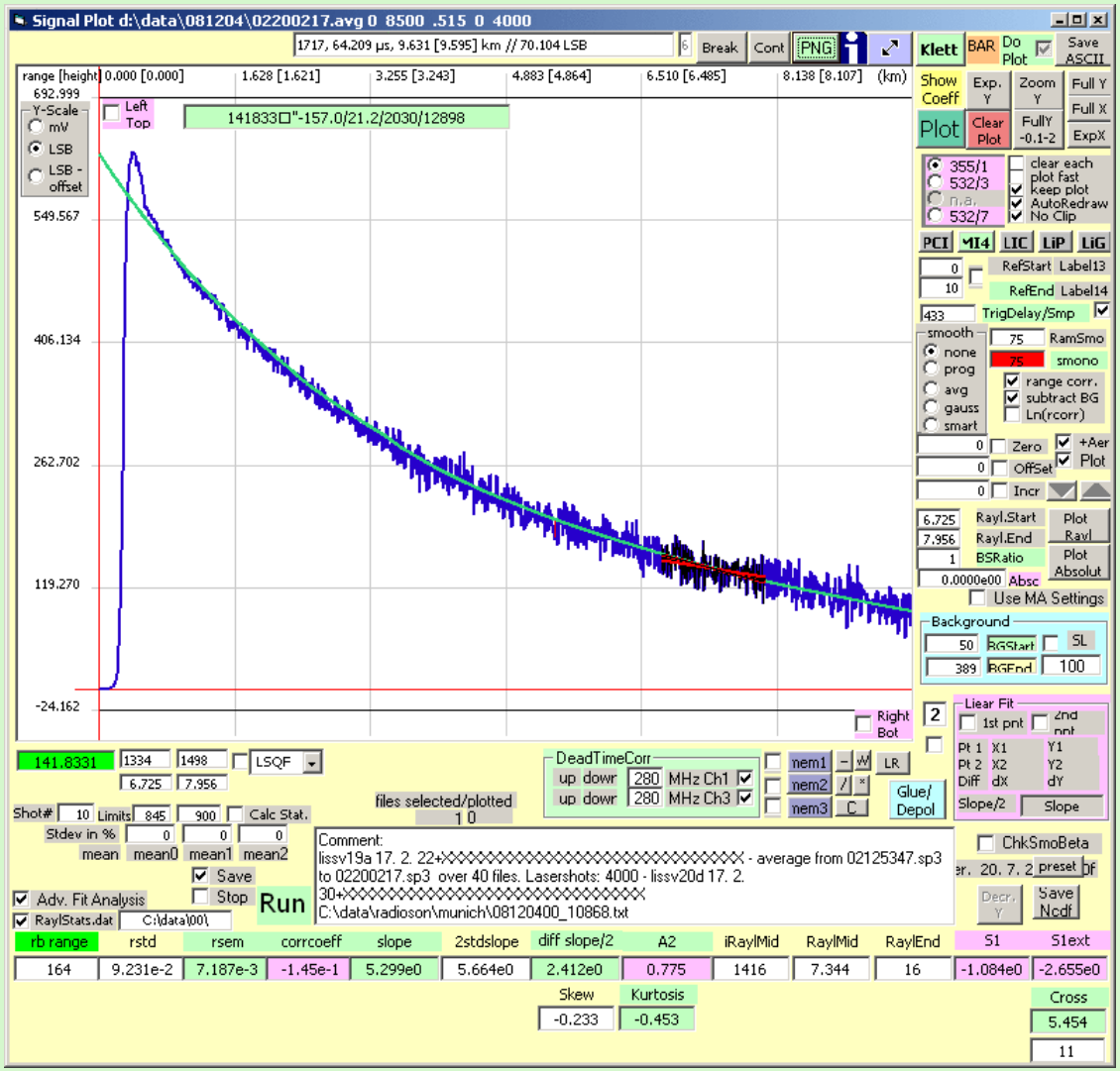
Rayleigh fit – cross criterion

The fitted attenuated Rayleigh signal must not be larger than the lidar signal at ranges lower than the fit range. What "larger" means depends on stastistical properties at the "crossing" point.



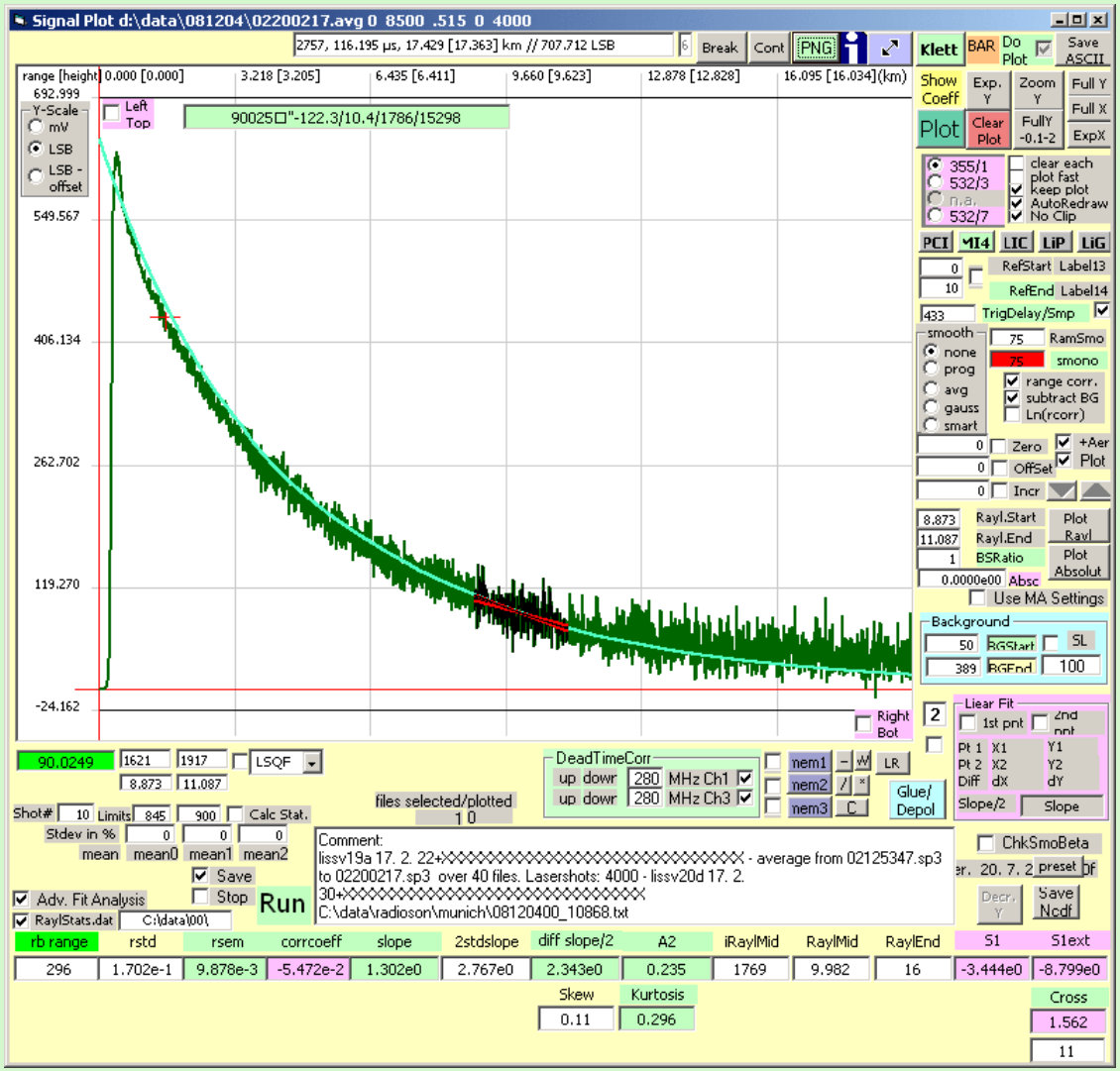
Rayleigh fit – cross criterion

The fitted attenuated Rayleigh signal must not be larger than the lidar signal at ranges lower than the fit range. What "larger" means depends on stastistical properties at the "crossing" point.



Rayleigh fit – relative standart error of the mean (RSEM)

The statistical fit accuracy of the signal to Rayleigh must be better than a limit (1% RSEM). RESM is a direct measure of the statistical uncertainty of the reference value.



Rayleigh fit – relative standart error of the mean (RSEM)

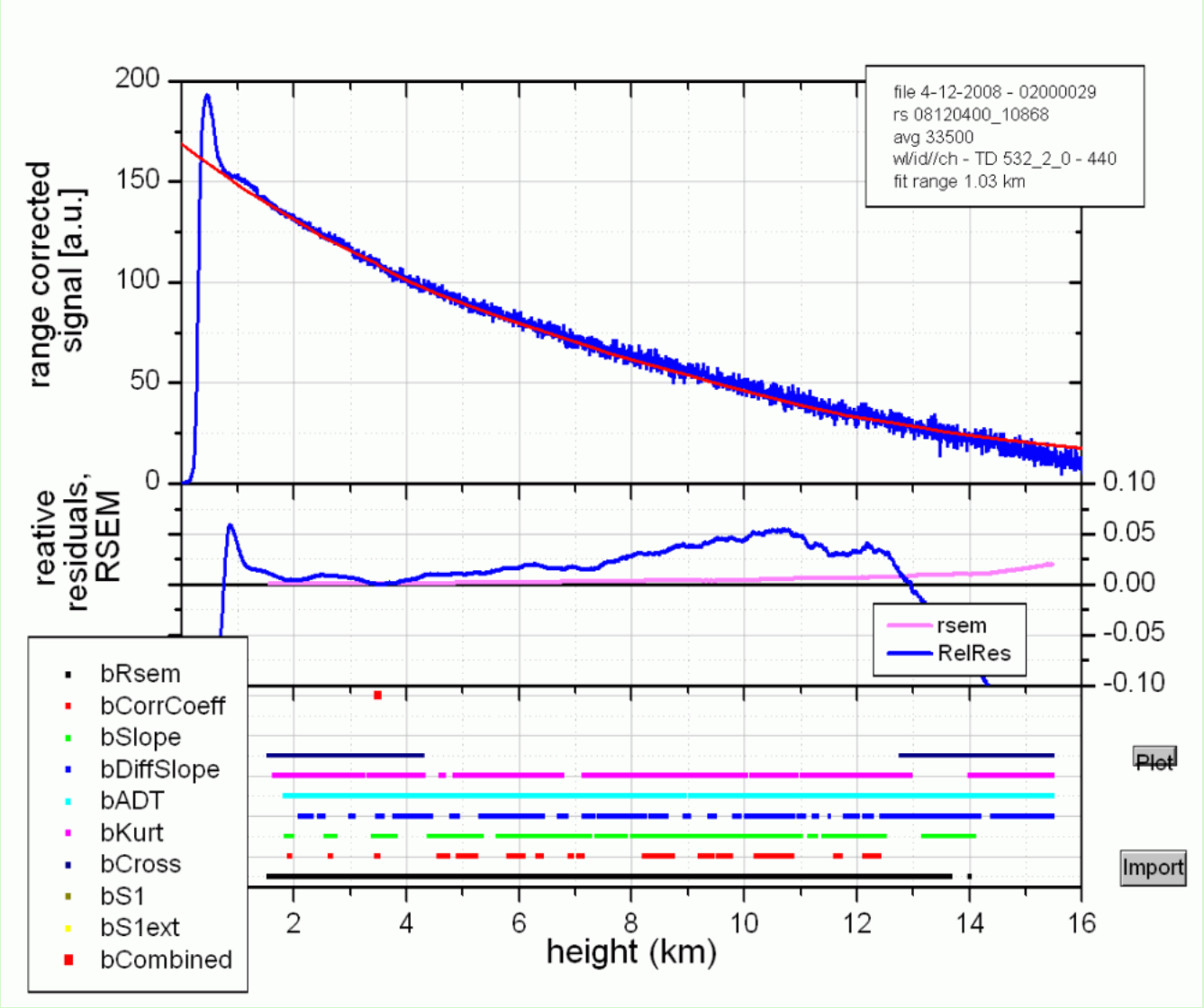
RESM limit determines

1. **standart deviation of the slope of the fit**
2. **allowed slope deviation**
3. **allowed differential slope deviation**

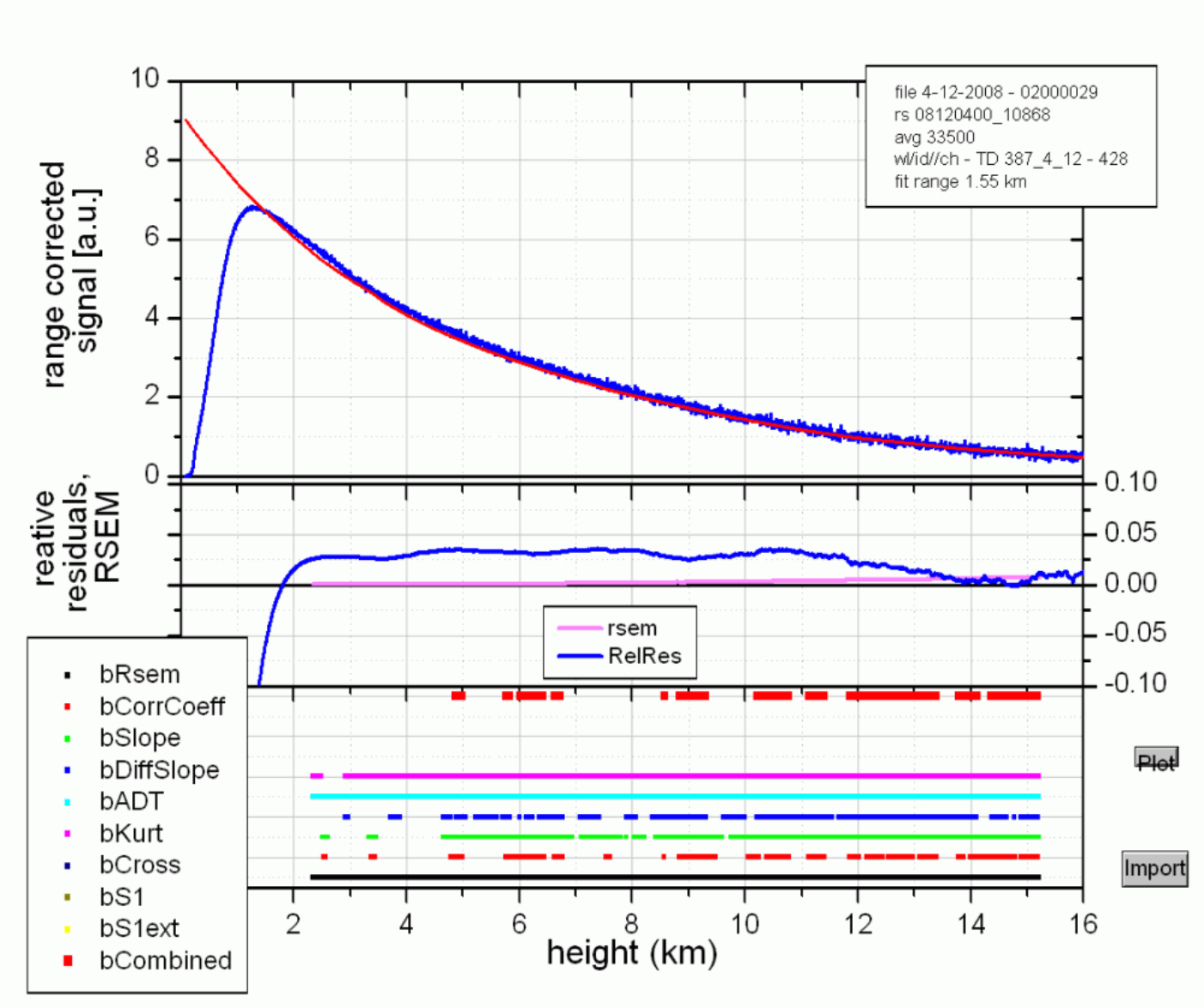
$$\beta_p(r,\lambda_0) = [\mathcal{P}(r)\,E(r)\,R(r_0) \;\; - 1] \;\; \beta_m(r,\lambda_0)$$

$$\mathcal{P}(r) = \frac{P(r,\lambda_0)P(r_0,\lambda_R)}{P(r_0,\lambda_0)P(r,\lambda_R)}$$

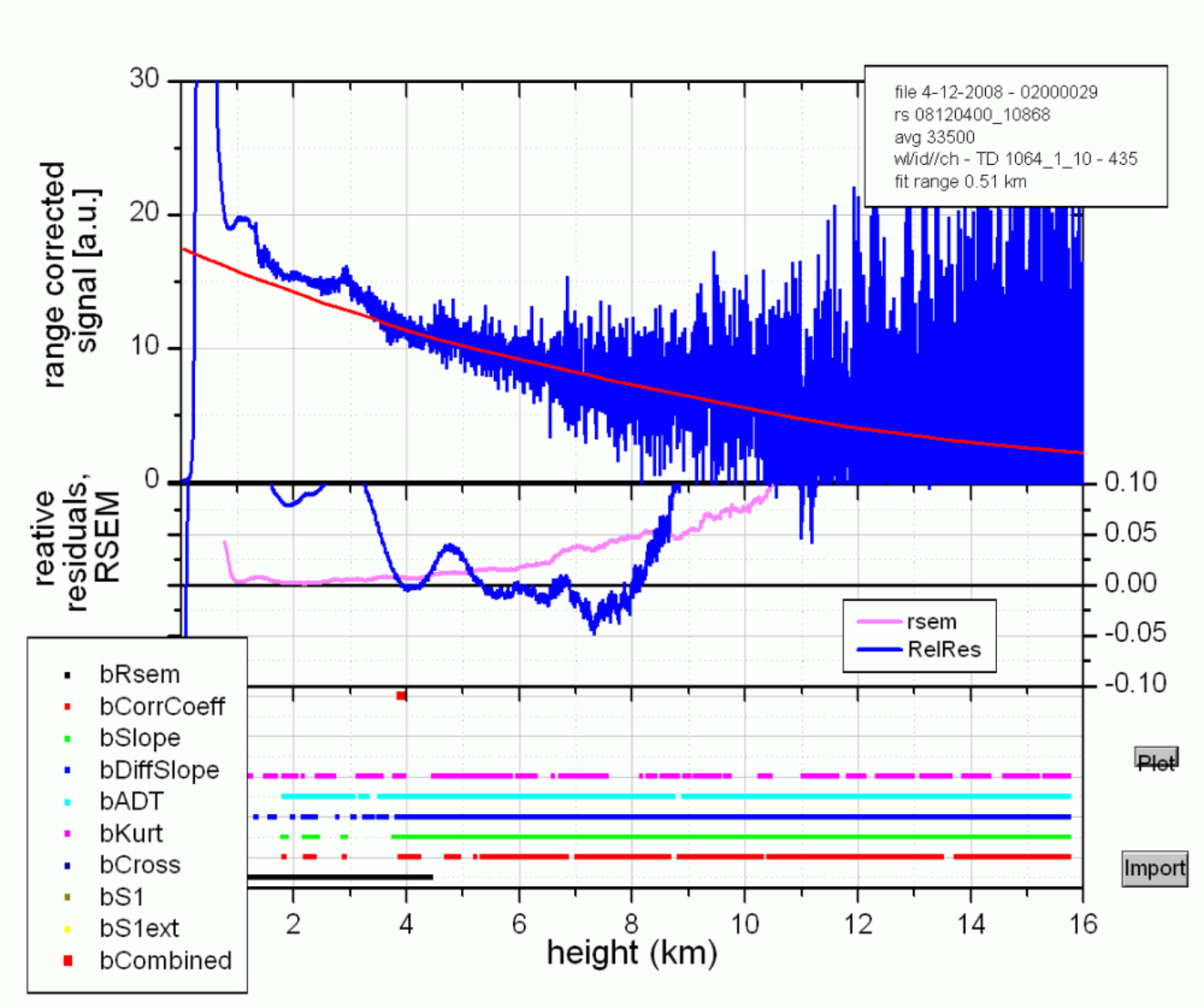
Rayleigh Fit – Residuals and Combined Criterion – 532 analog



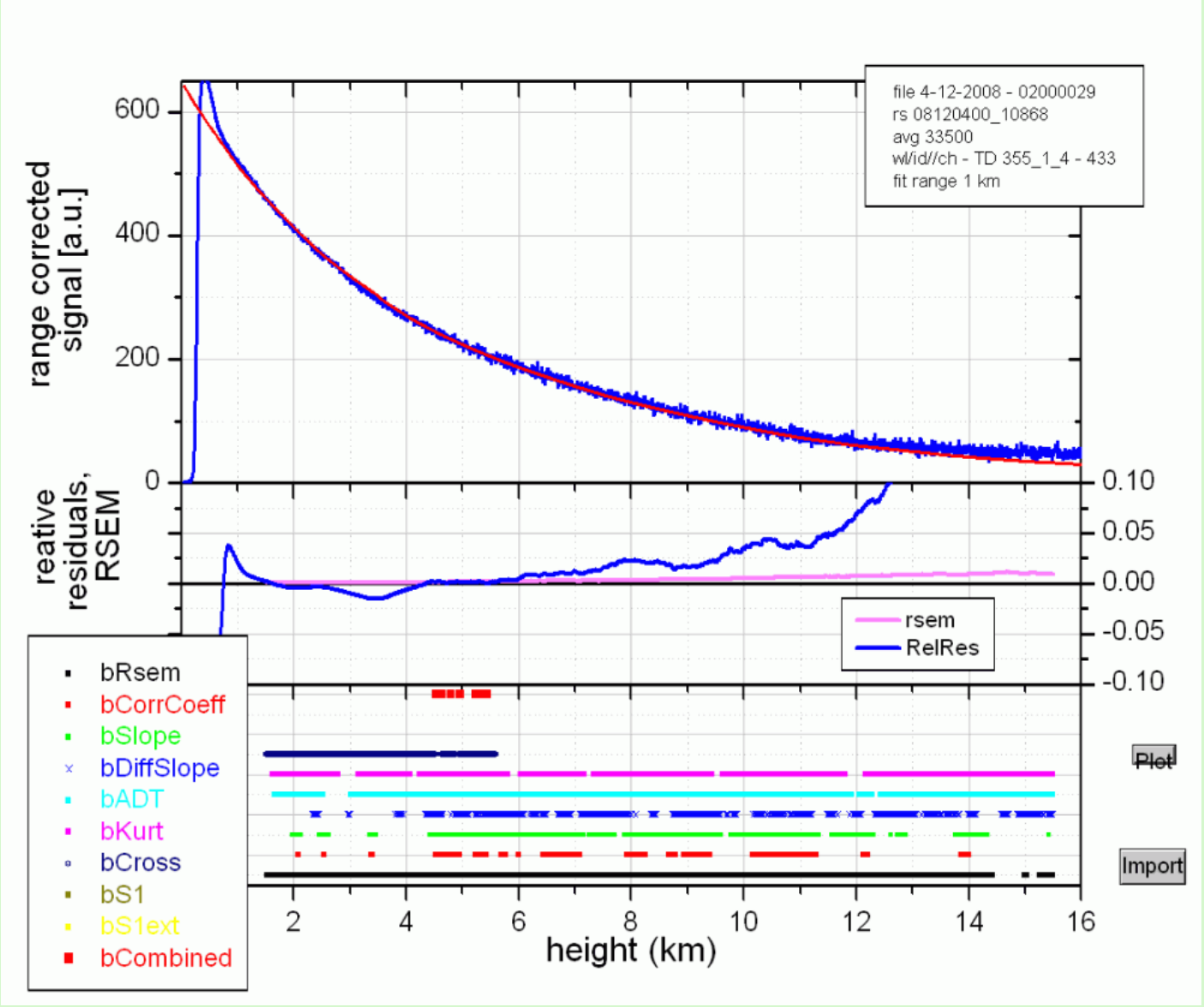
Rayleigh Fit – Residuals and Combined Criterion – 387 photon-counting



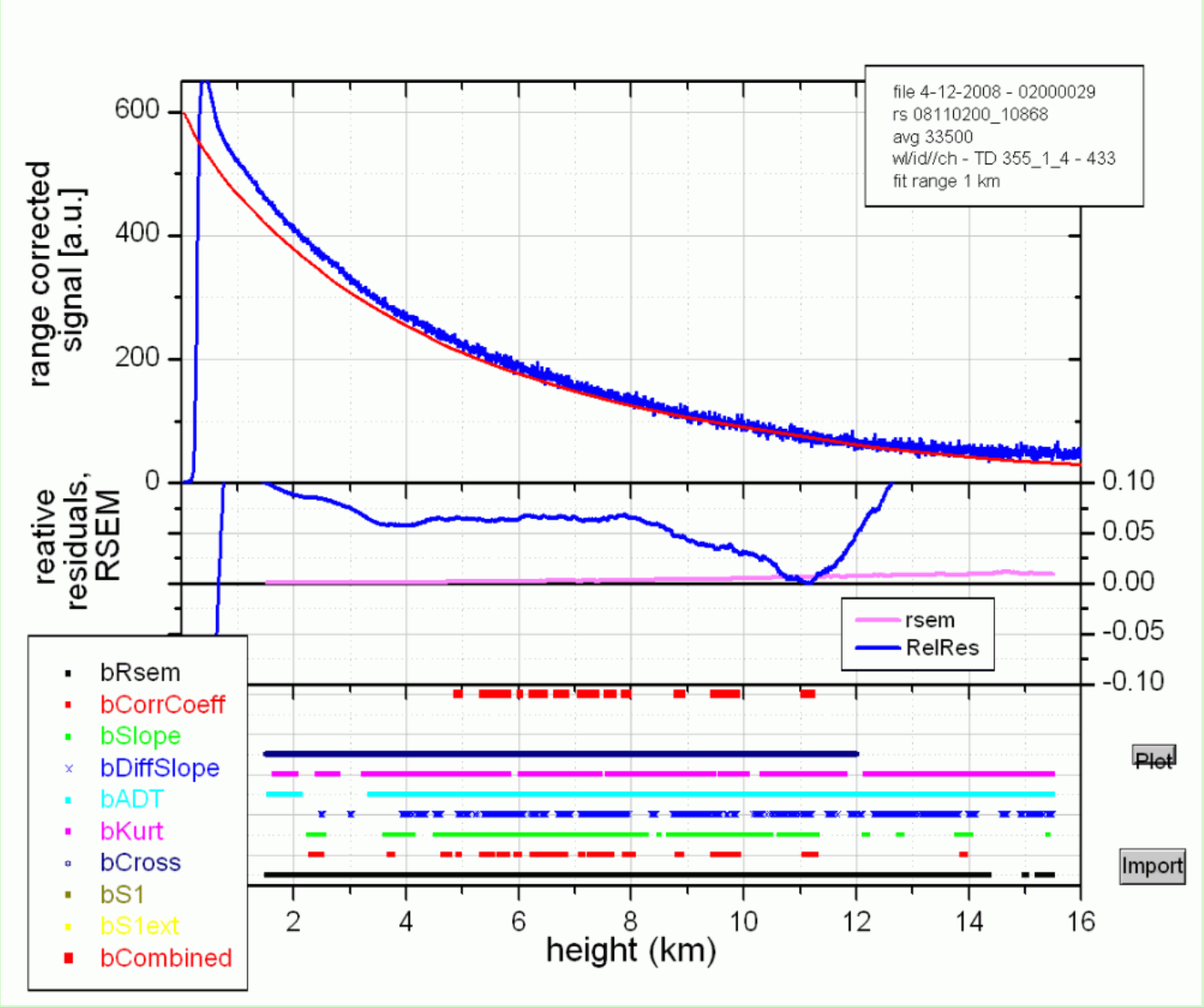
Rayleigh Fit – Residuals and Combined Criterion – 1064 analog



Rayleigh Fit – Radiosonde (355 analog)



Rayleigh Fit – Radiosonde (355 analog)



Acknowledgement

This work has been supported by the European Commission under grant RICA-CT-2006-025991