Meteorologisches Institut
Universität Bonn

On the reliability of using the maximum explained variance as criterion for optimum segmentations

Ralf Lindau & Victor Venema
University of Bonn
Germany

Meteorologisches Institut
Universität Bonn

Internal and External Variance

Consider the differences of one station compared to a neighbor reference.

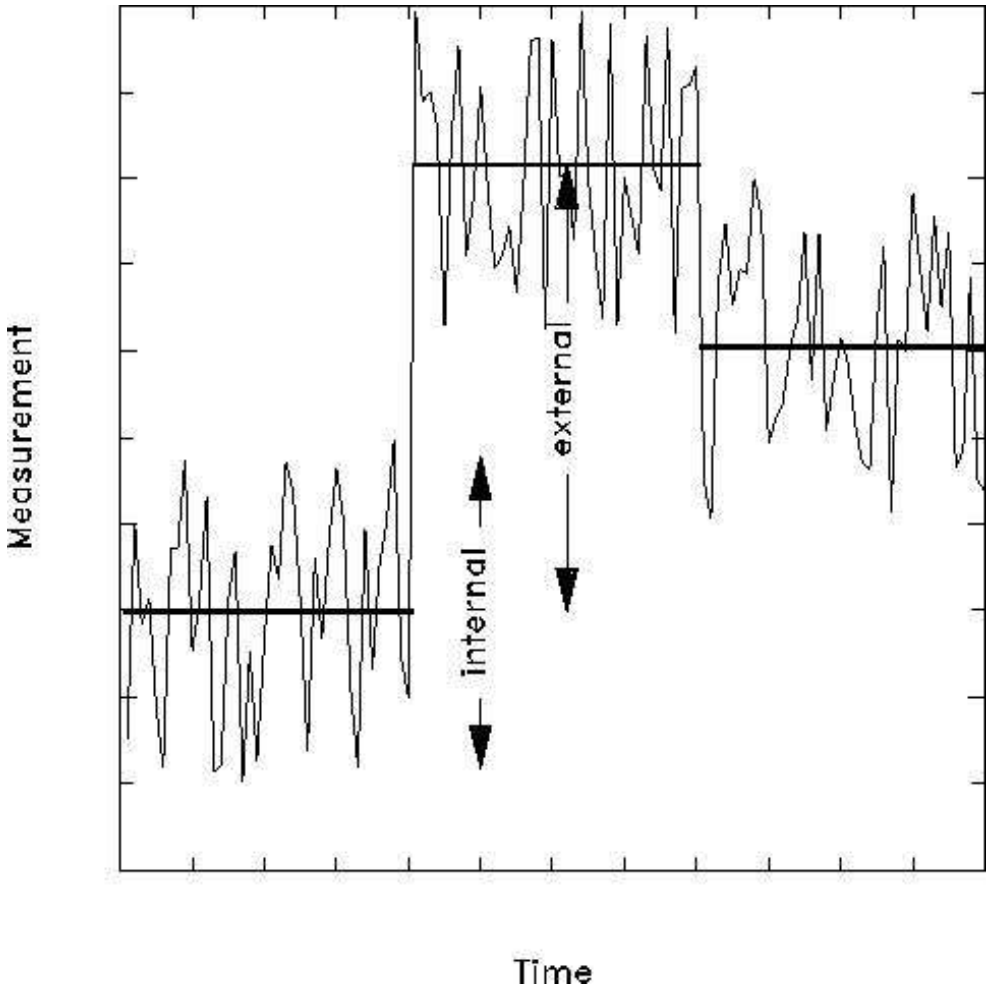
The dominating natural variance is cancelled out, because it is very similar at both stations.

Breaks become visible by abrupt changes in the station-reference time series.

Internal variance
within the subperiods

External variance
between the means of different subperiods

Break criterion:
Maximum external variance



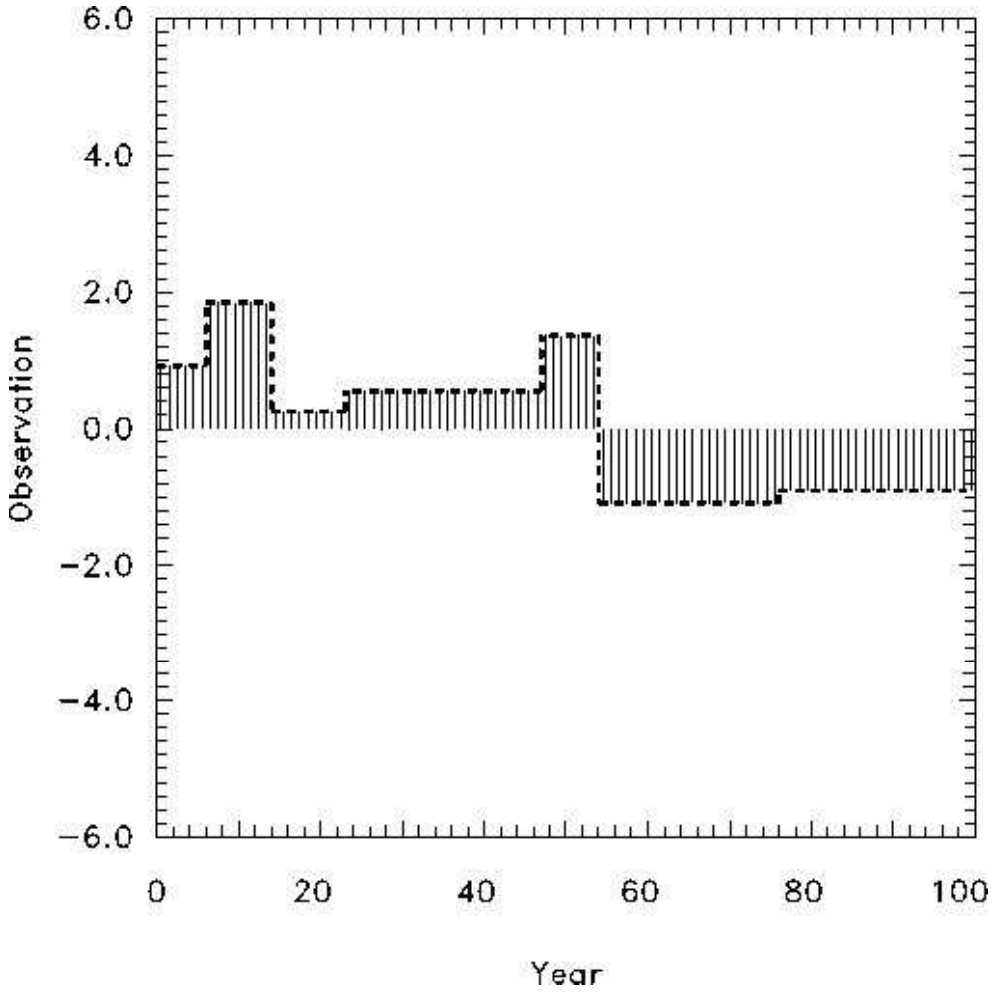
Part I

True skill (signal RMS) and explained variance are only weakly correlated

Meteorologisches Institut
Universität Bonn

RMS² as skill measure

Consider the time series of the inhomogeneities as a signal that we want to detect.

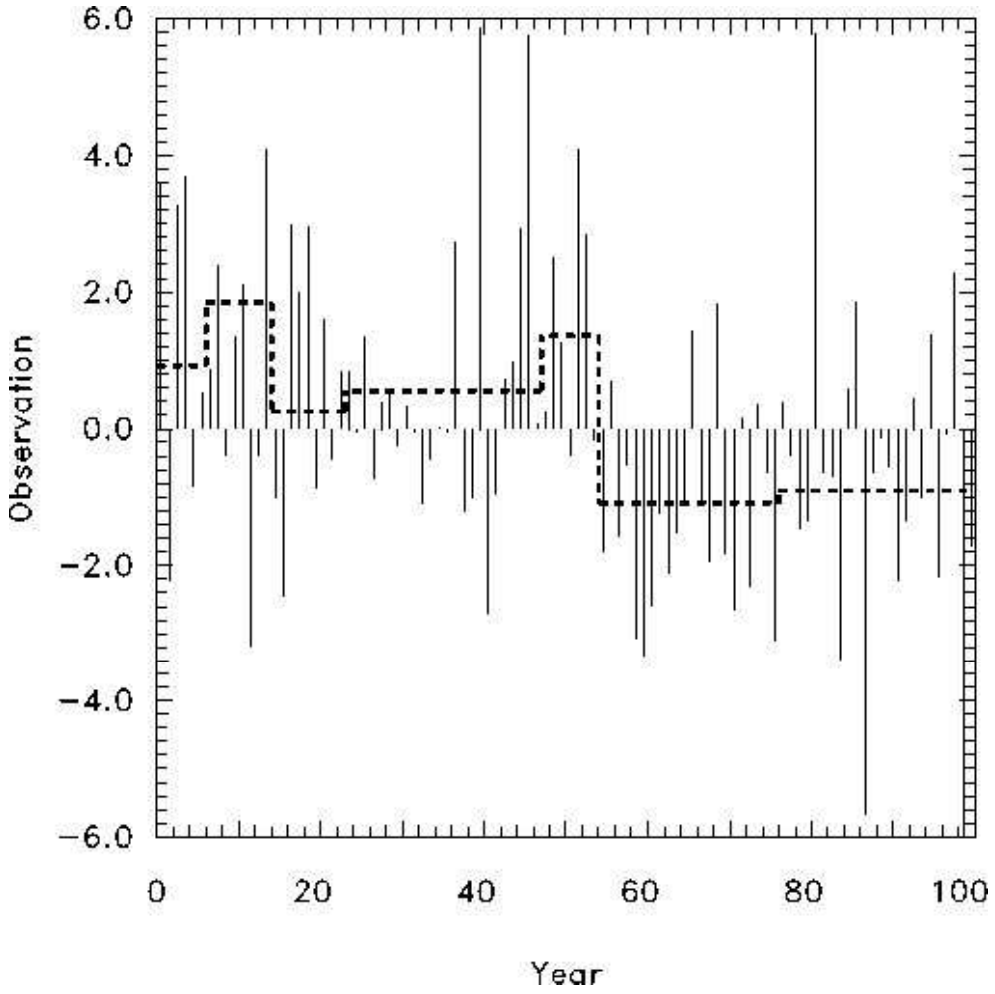


Meteorologisches Institut
Universität Bonn

RMS² as skill measure

Consider the time series of the inhomogeneities as a signal that we want to detect.

This is hampered by superimposed noise.



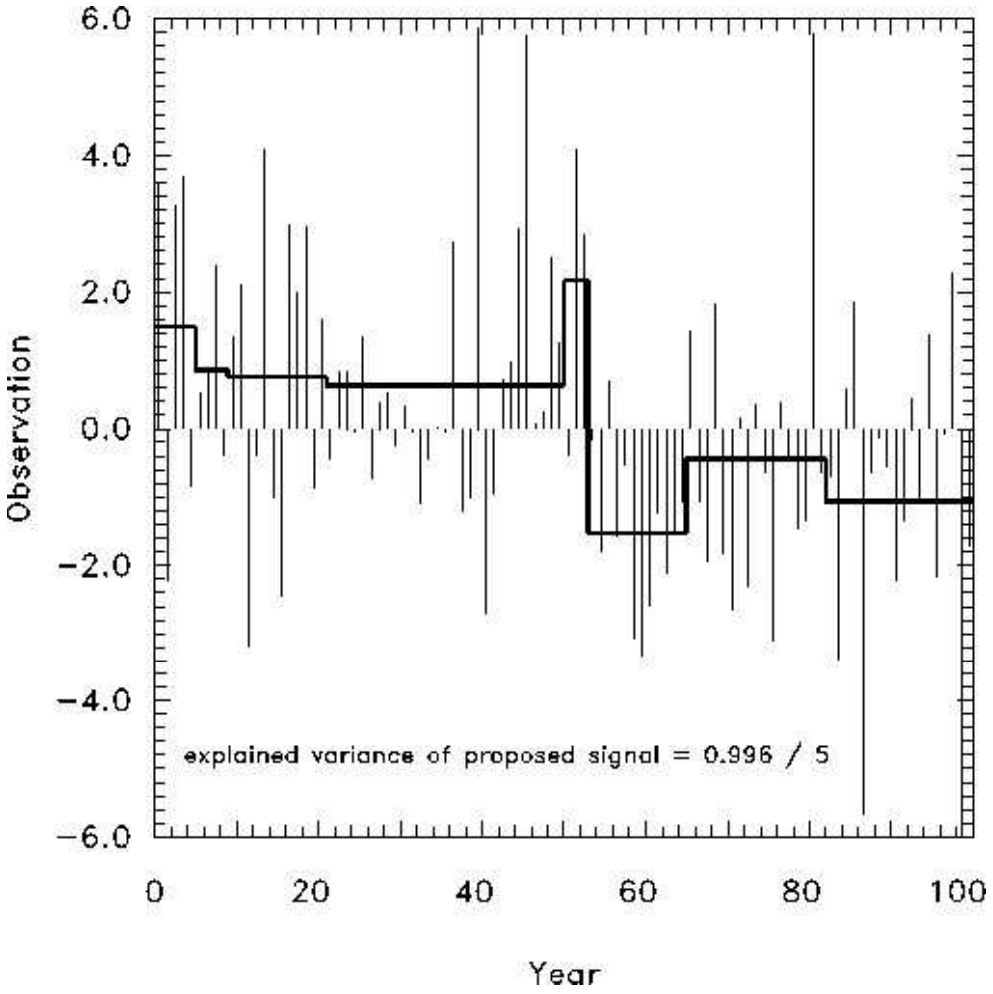
Meteorologisches Institut
Universität Bonn

RMS² as skill measure

Consider the time series of the inhomogeneities as a signal that we want to detect.

This is hampered by superimposed noise.

Homogenization algorithms search for the maximum external variance of the noisy data. This is the proposed signal.



Meteorologisches Institut
Universität Bonn

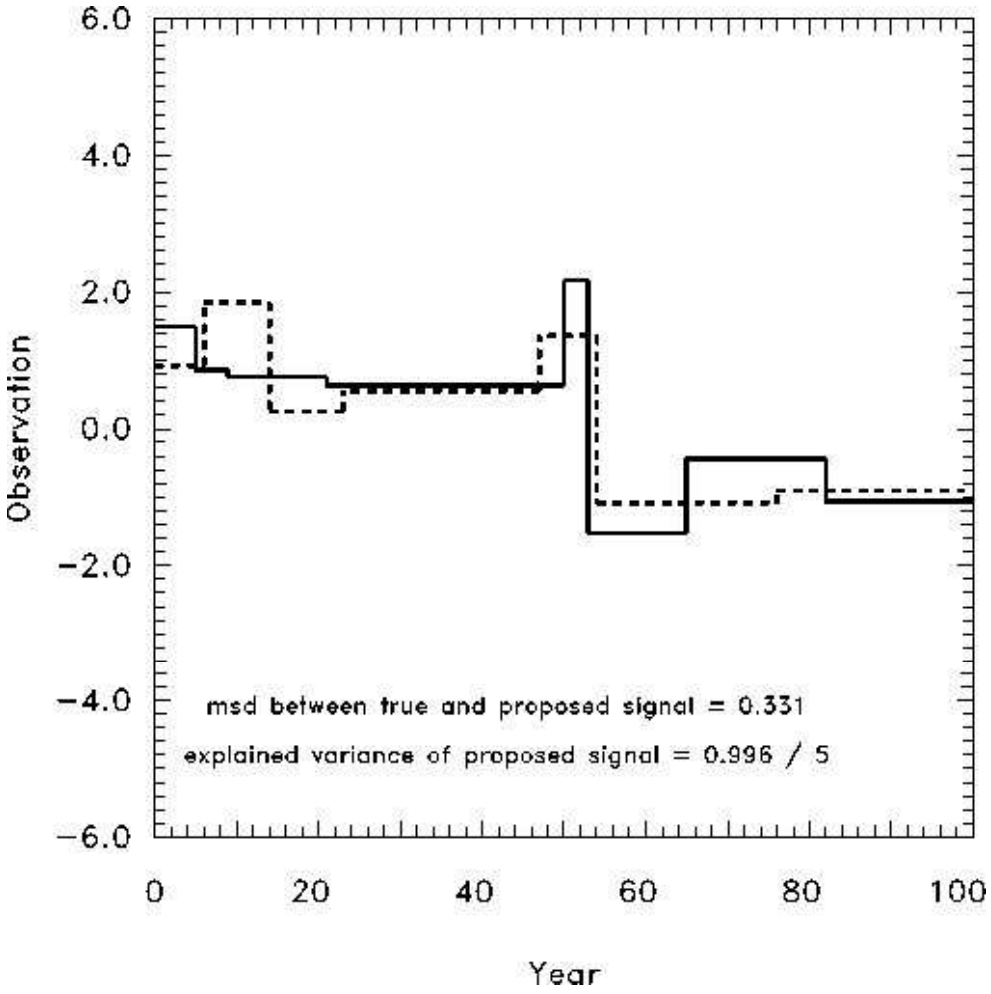
RMS² as skill measure

Consider the time series of the inhomogeneities as a signal that we want to detect.

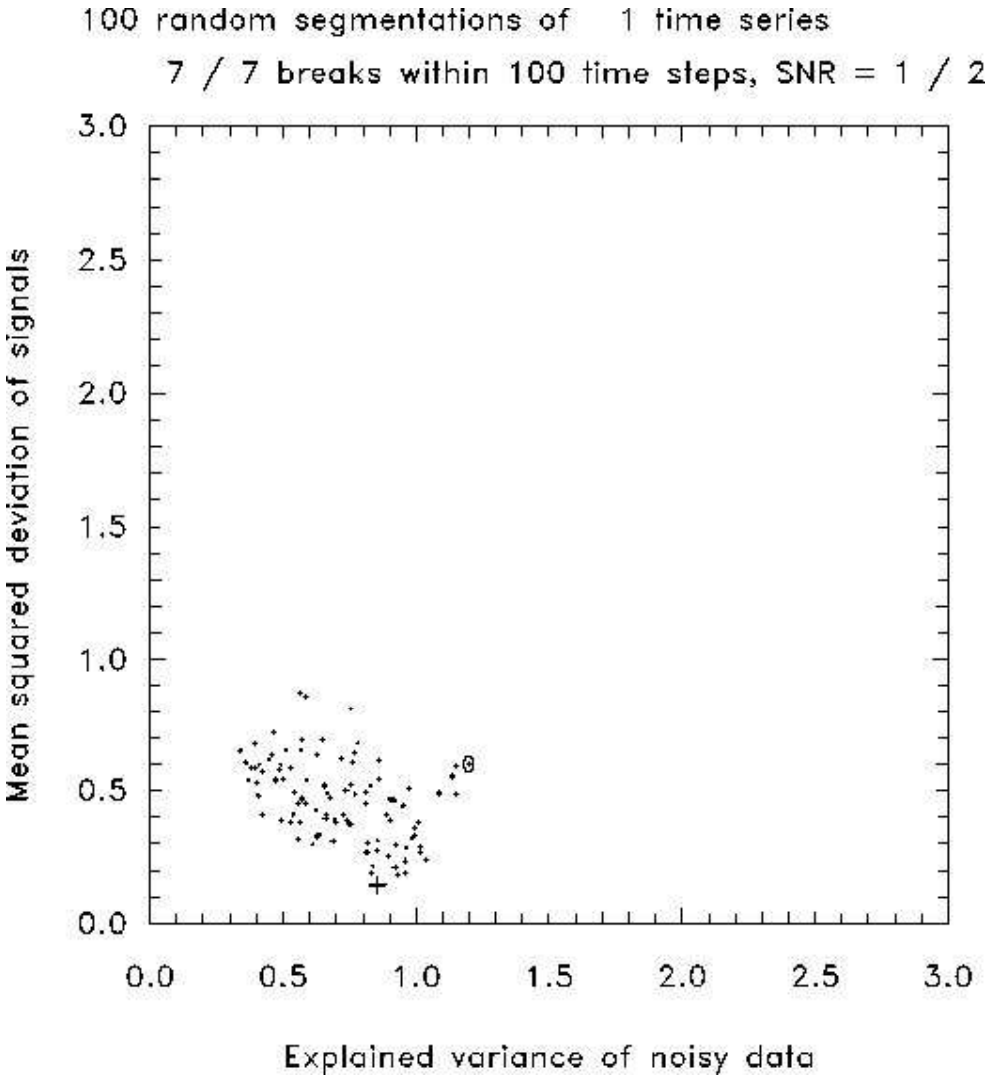
This is hampered by superimposed noise.

Homogenization algorithms search for the maximum external variance of the noisy data. This is the proposed signal.

The Mean Squared Difference (RMS^2) between proposed and true signal is an appropriate skill measure.



Explained Variance versus Signal RMS



We start from very simplistic (random) segmentations to see the full variety of solutions and their correlation.

We have two measures:

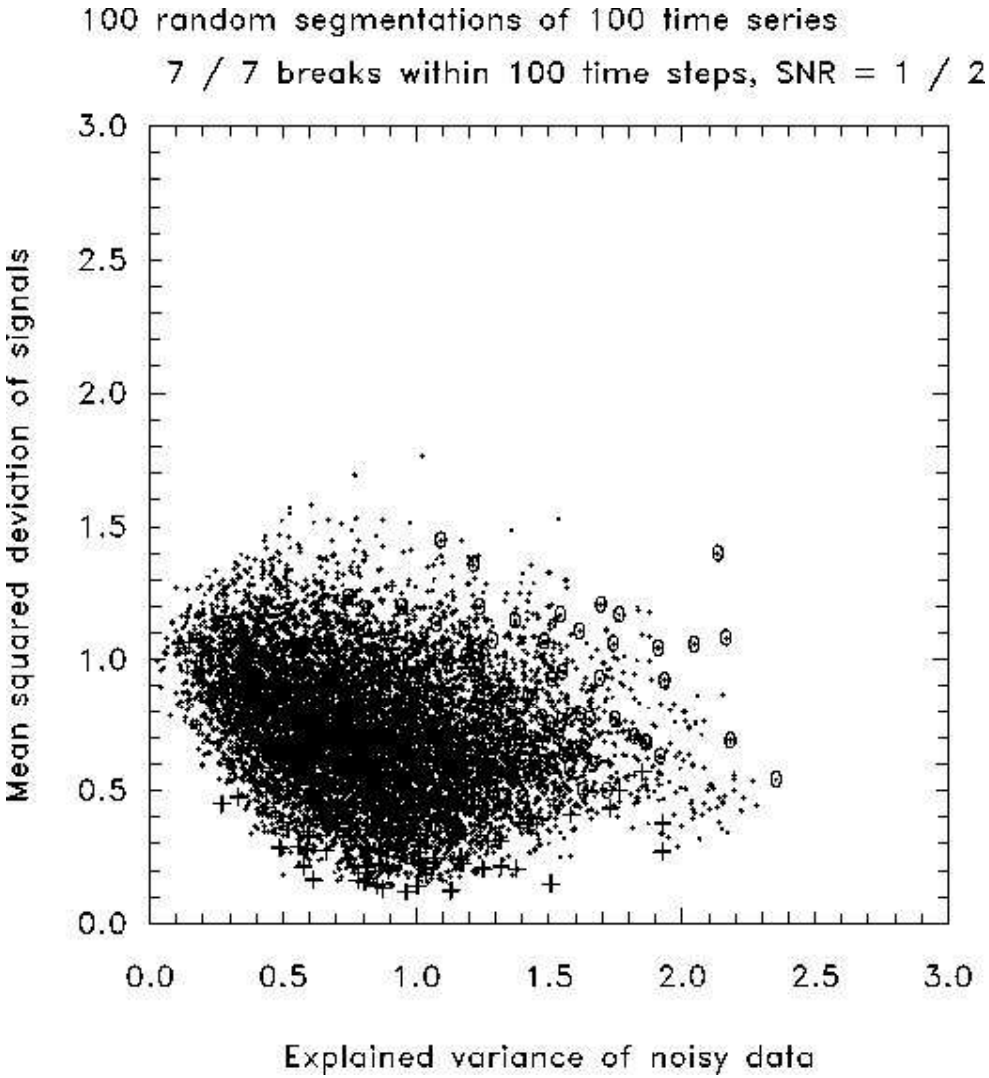
1. **The variance explained by the tested breaks in the noisy data.**
2. **The Mean Squared Deviation between proposed and true signal.**

For real cases, 1 is the only available measure as the true signal is not known.

With simulated data we are able to compare 1 and 2.

Meteorologisches Institut
Universität Bonn

100 instead of 1 time series



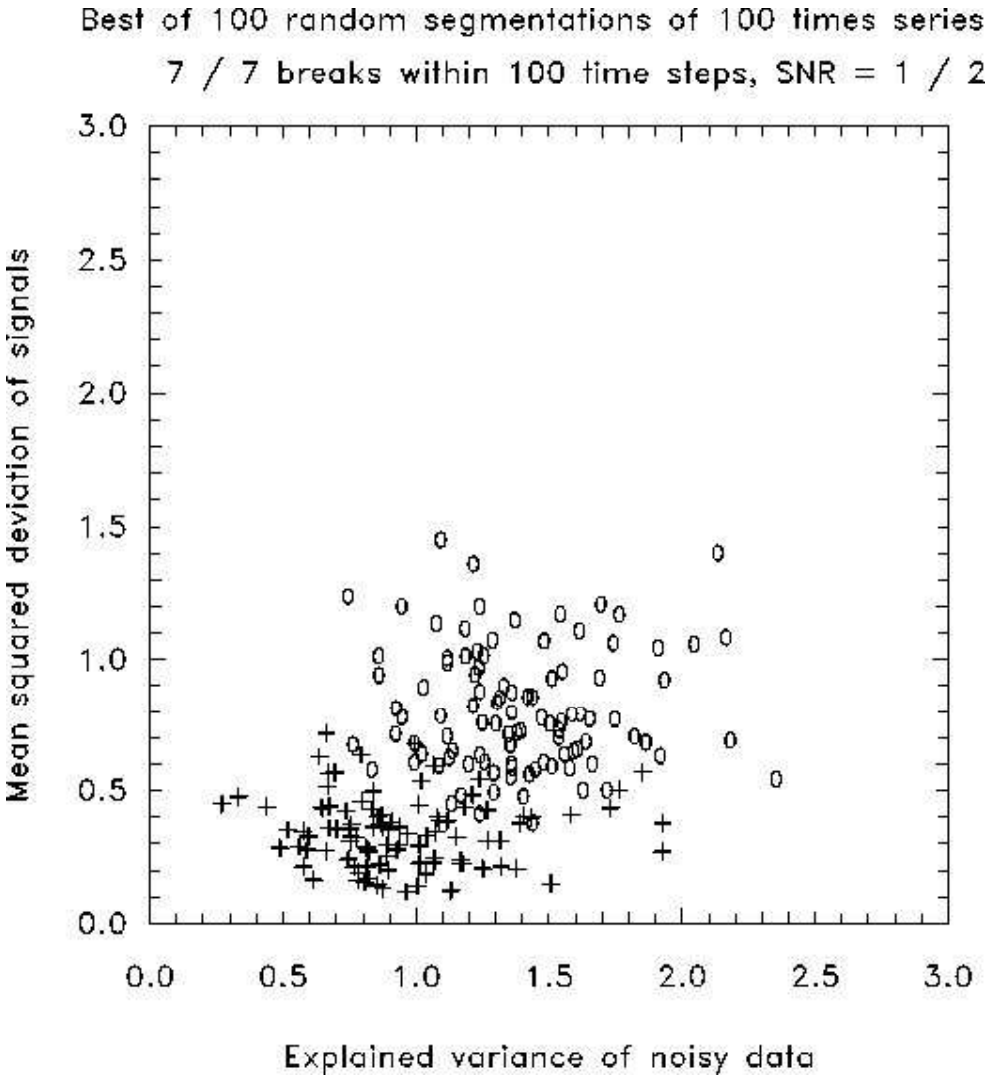
Repeat the exercise for 100 instead of one time series.

The best of 100 random solutions are marked for each of the 100 time series.

For the explained variance with 0

For the really best solution with +

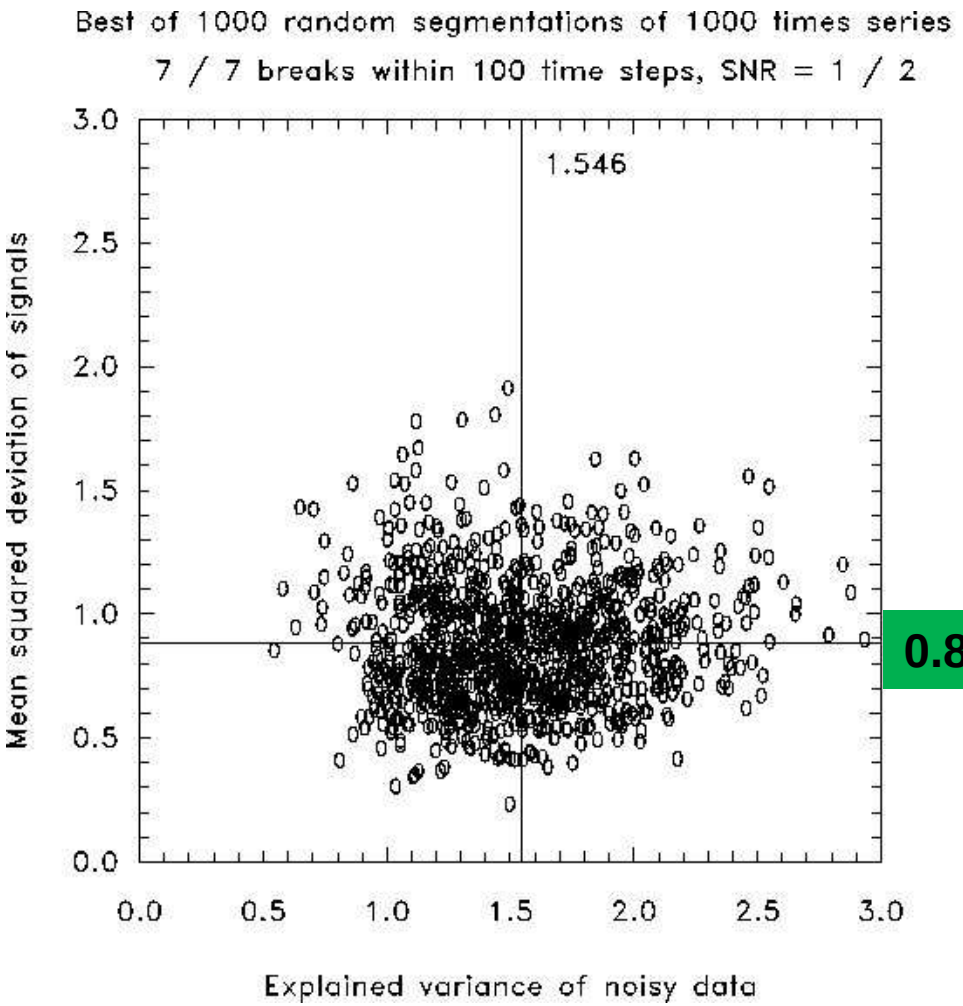
Best solutions



Show only the best solutions (0 and +) for each time series.

Crosses and circles are clearly separated.

From 100 to 1000



Increase numbers from 100 by 100 to 1000 by 1000.

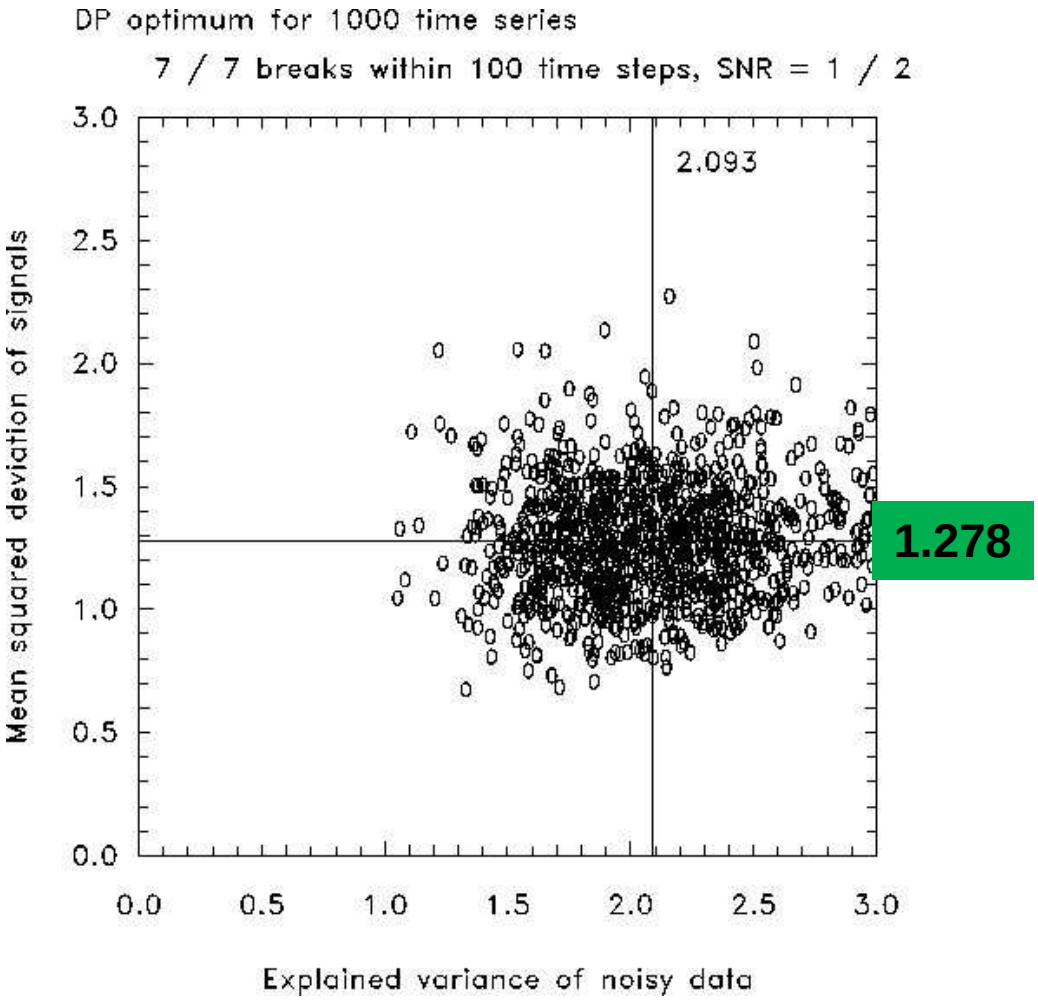
Show only circles, the normally proposed solutions, determined by the maximum explained variance.

Mean explained variance is 1.546.

Mean RMS^2 is 0.881, not far away from 1 (no skill).

Meteorologisches Institut
Universität Bonn

Dynamic Programming



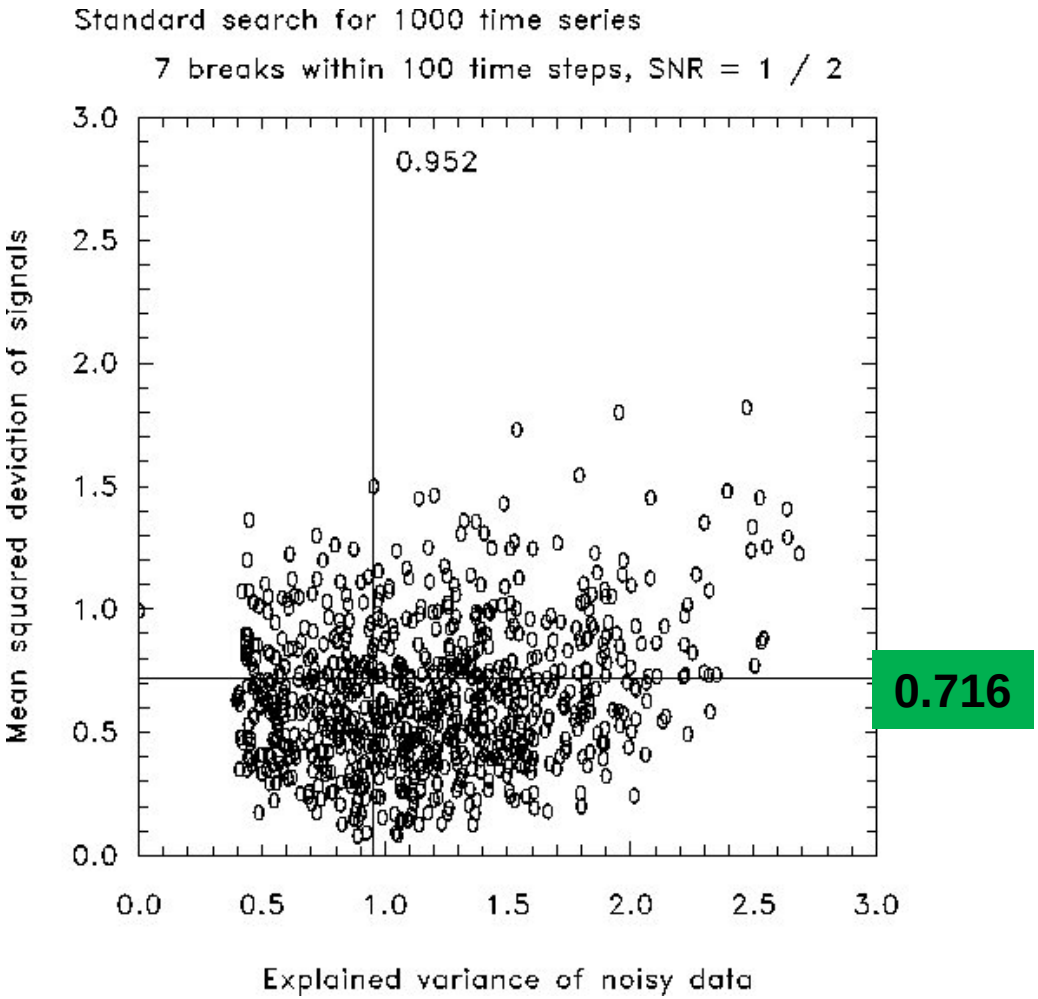
Now use Dynamic Programming to find the optimum in explained variance, instead of choosing just the best of 1000.

Explained variance increases, but also the signal deviation.

With 1.278 it is larger than 1, which is worse than doing nothing.

Continuing the search until the true number of breaks is reached, produces very bad solutions.

Standard search



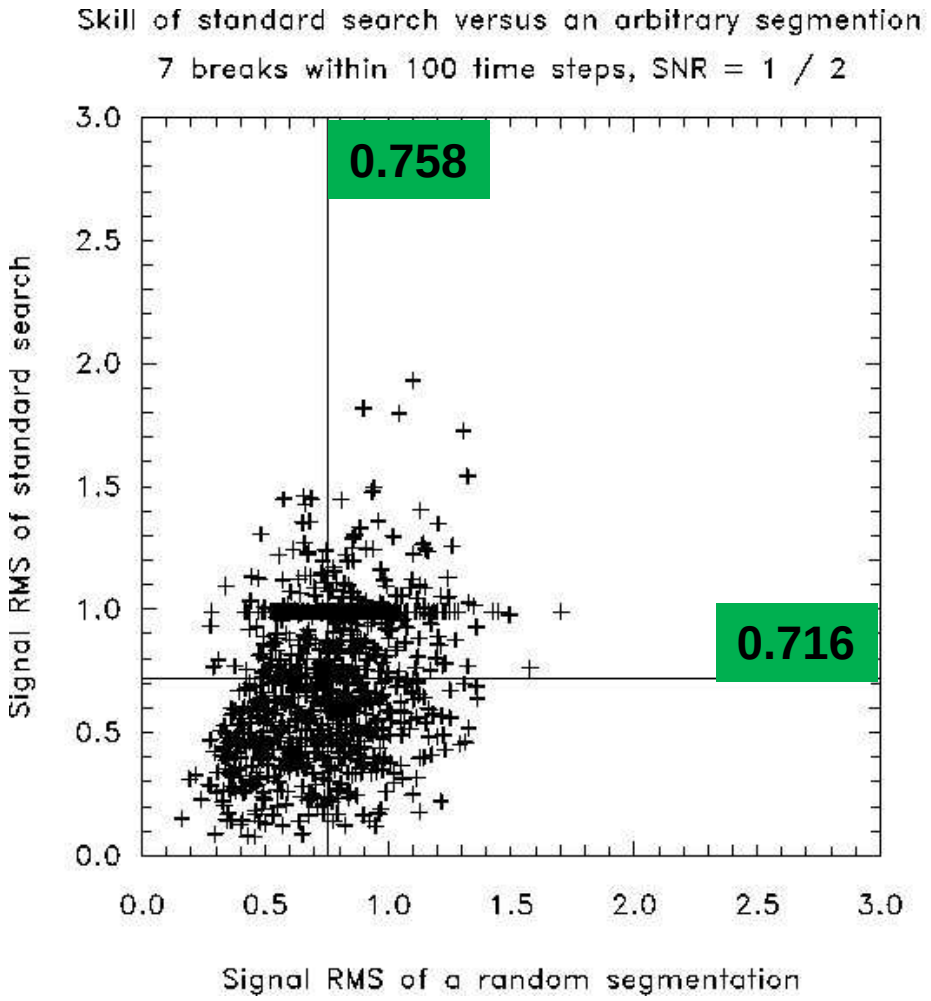
So, finally, the standard stop criterion (CM) is added.

$$ln\,(1-v) + \frac{2\,k\,ln(n)}{n-1} = min$$

This reduces the RMS, but only to 0.716, which is not far below 1. Is that better than random?

Thus, compare Standard Search RMS with random segmentation RMS.

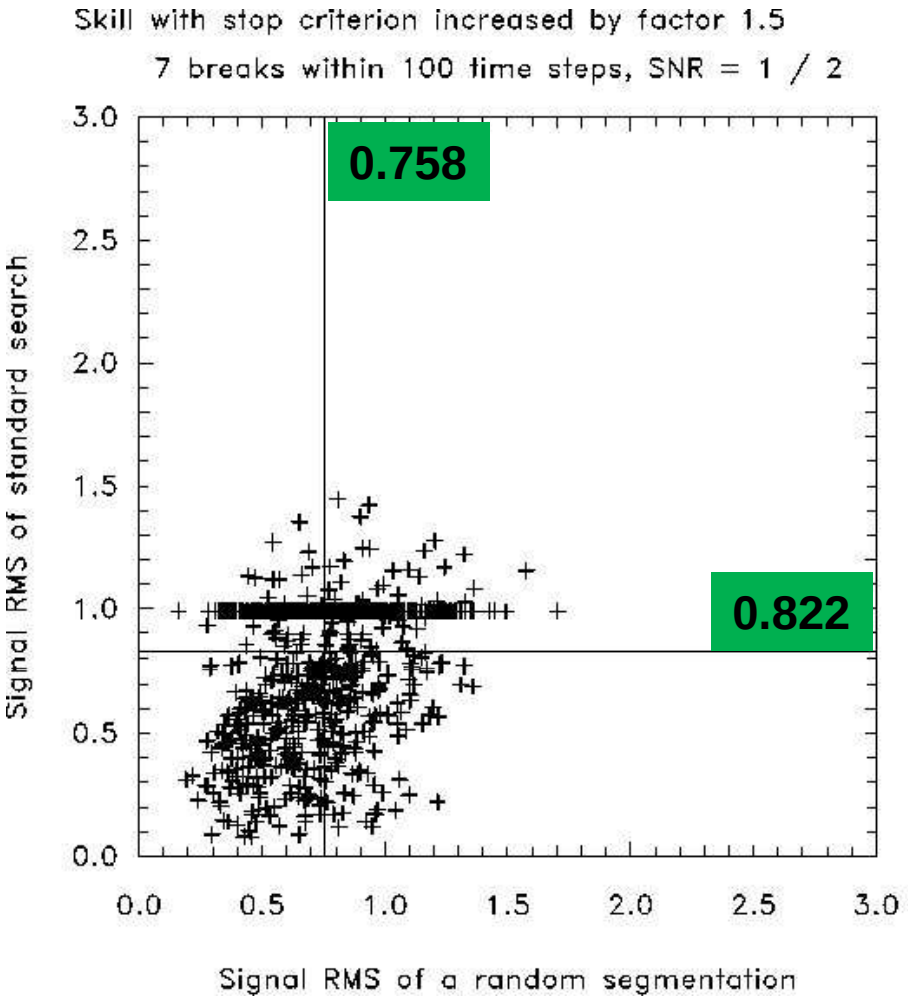
RMS Standard vs. arbitrary



For SNR = ½, the skills of standard search and an arbitrary segmentation are comparable.

Obviously, the standard search is mainly optimizing the noise, producing completely random results.

Increased Stop Criterion



Does a higher stop criterion help?

Increase the stop criterion by a factor of 1.5 (from 2 ln(n) to 3 ln(n)).

The signal deviation even increases from 0.716 to 0.822.

The reason is that more zero solutions are produced (with RMS of 1), which is not compensated by more accurate non-zero solutions.

Which SNR is sufficient?

RMS skill for:

0 **Random segmentation**

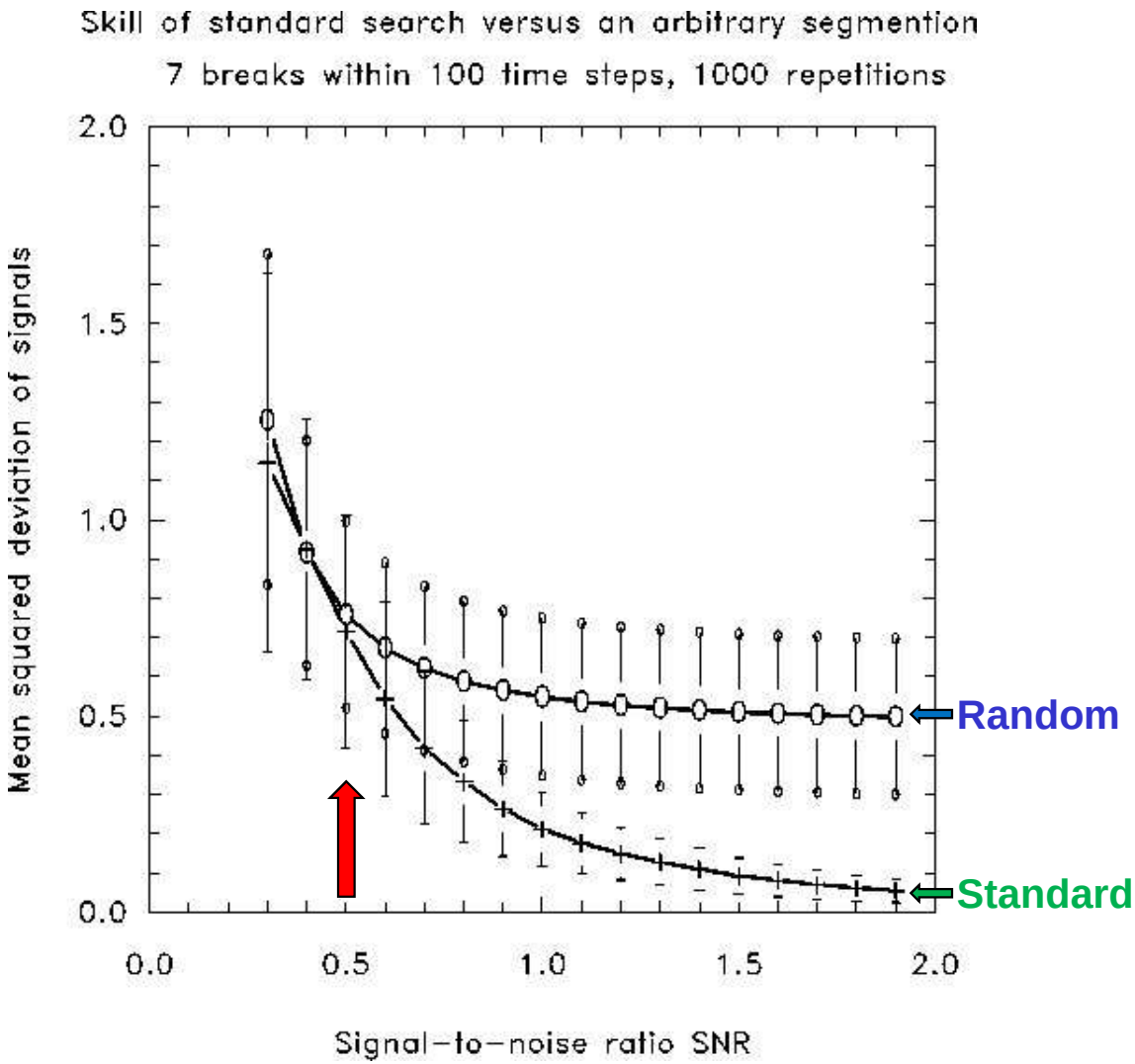
+ **Standard search**

for different SNRs.

So far we considered SNR = ½

Random segmentation and standard search have comparable skills.

Only for SNR > 1, the standard search is significantly better.



Meteorologisches Institut
Universität Bonn

Conclusions Part I

Break search algorithm rely on the explained variance to identify the breakpoints.

For signal to noise ratios of ½, the explained variance does not reflect the true skill.

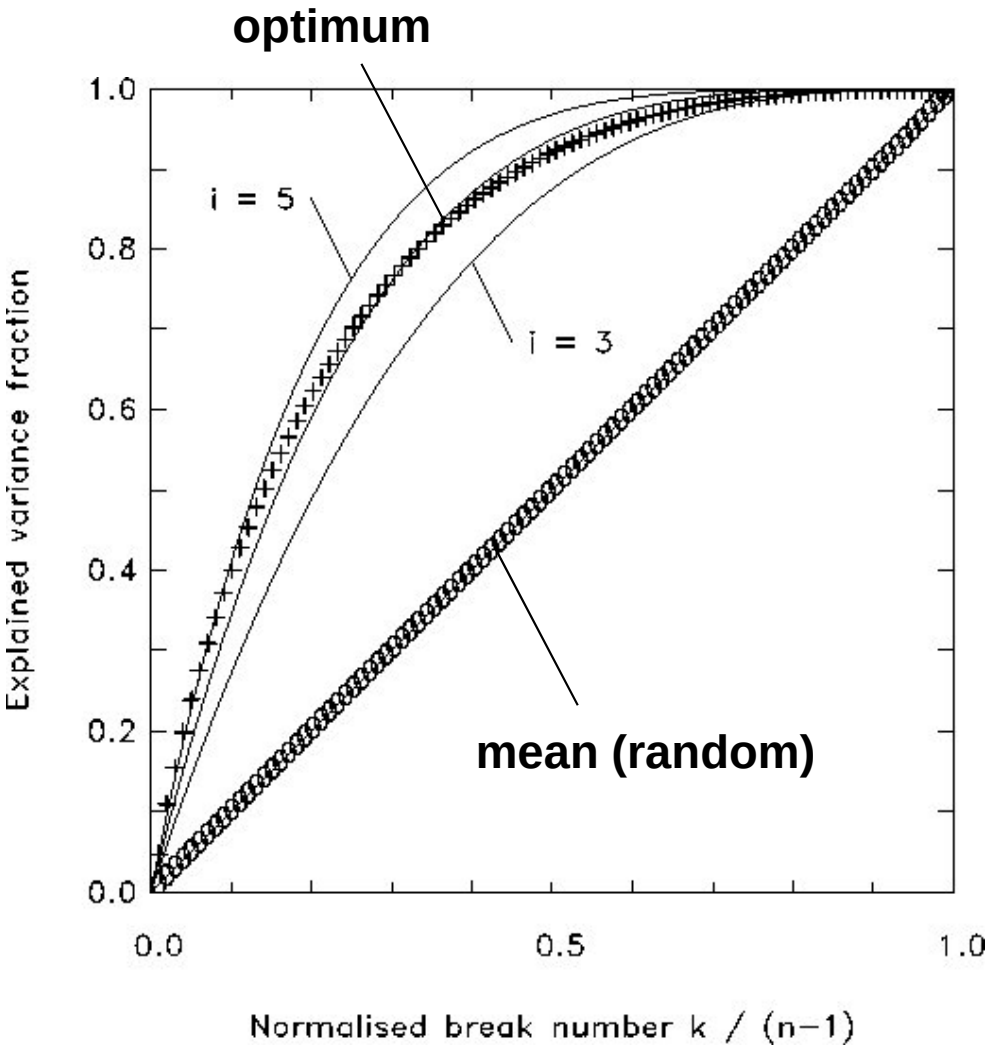
Consequently, the obtained segmentations do not differ significantly from random.

Part II

Theoretical Explanation:
Break and Noise Variance

Meteorologisches Institut
Universität Bonn

Behavior of Noise



External variance as function of tested break number. No breaks, no variance explained. If n-1 breaks were included, the full variance would be explained. The two fat lines show the transition.

Optimum segmentation:

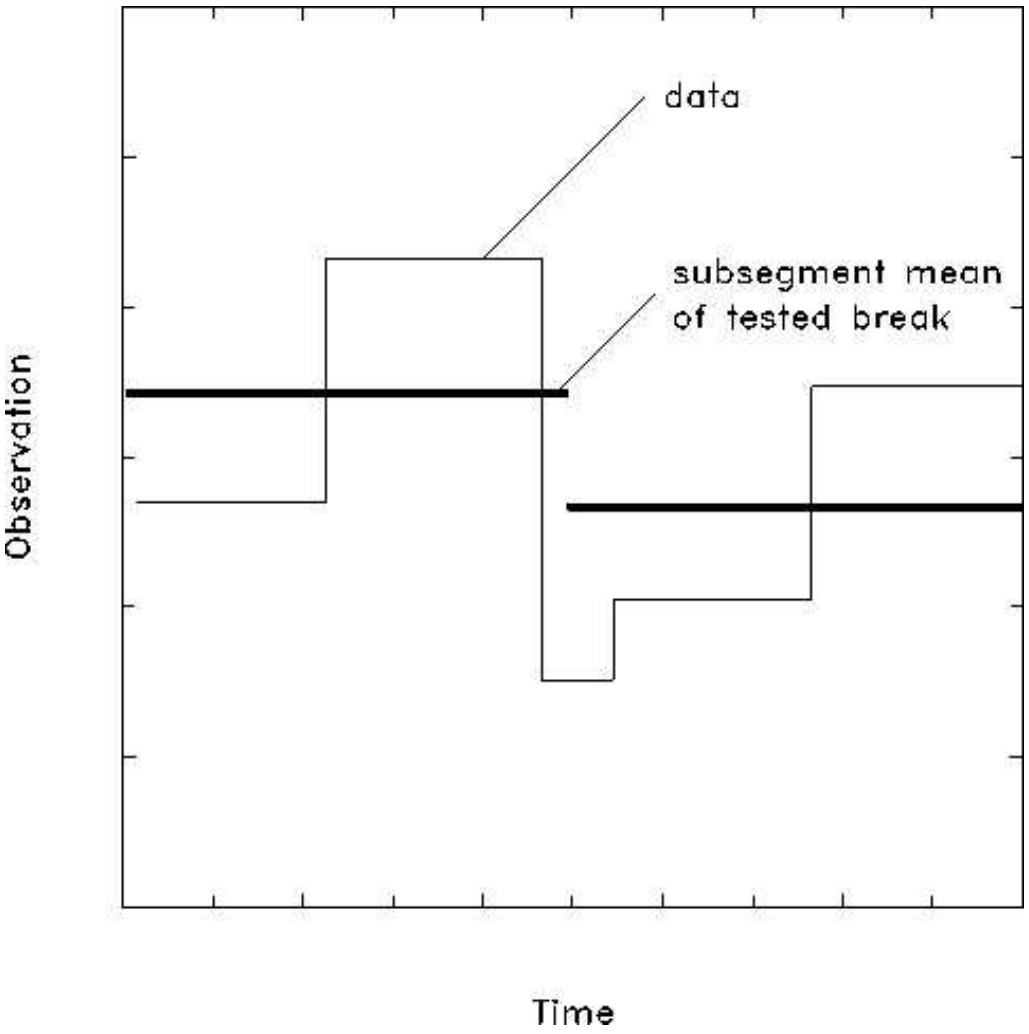
$$1 - v = \left(1 - \frac{k}{n-1}\right)^4$$

Mean (random) segmentation:

$$v = \frac{k}{n-1}$$

Lindau, R. and V. Venema, 2013: On the multiple breakpoint problem and the number of significant breaks in homogenization of climate records, Idöjaras - Quarterly Journal of the Hungarian Meteorological Service, 117, No. 1, 1-34.

True Breaks



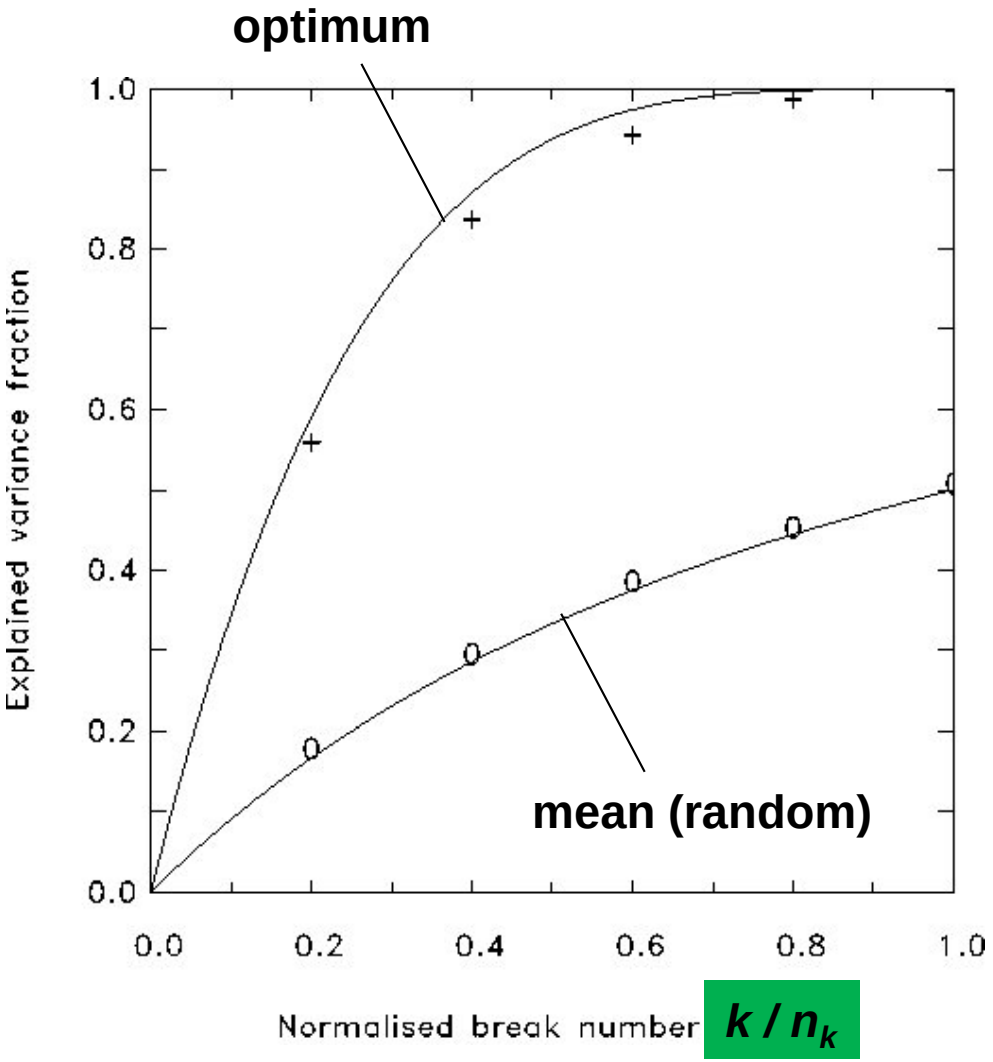
For true breaks, constant periods exist. Tested segment averages are the (weighted) means of such (few) constant periods.

This is quite the same situation as for random scatter, only that less independent data is underlying.

Obviously, the number of breaks n_k plays the same role as the time series length n did before for random scatter.

Consequently, we expect the same mathematical behaviour, but on another scale.

Behavior of breaks



As expected, the best segmentations for pure breaks behave similar to the best segmentation for pure noise.

However, an important difference is that length n (about 100) is replaced by n_k (about 5).

And random breaks behave completely different. The variance do not grow linearly with $v = \frac{k}{n-1}$

but with

$$v = \frac{k}{n_k + k}$$

Why $k/(n_k+k)$?

Short explanation:

Consider a random segmentation trial with k break positions, where k is equal to the correct number of breaks n_k.

$$k = n_k$$

Each test segment spans (in average) over two true segments. This means that always 2 segment means are averaged, which reduces the variance by a factor of 2.

Four formulae

We described four types of external variance growth with break number k.

We distinguished break and noise variance, both for random and optimum segmentations.

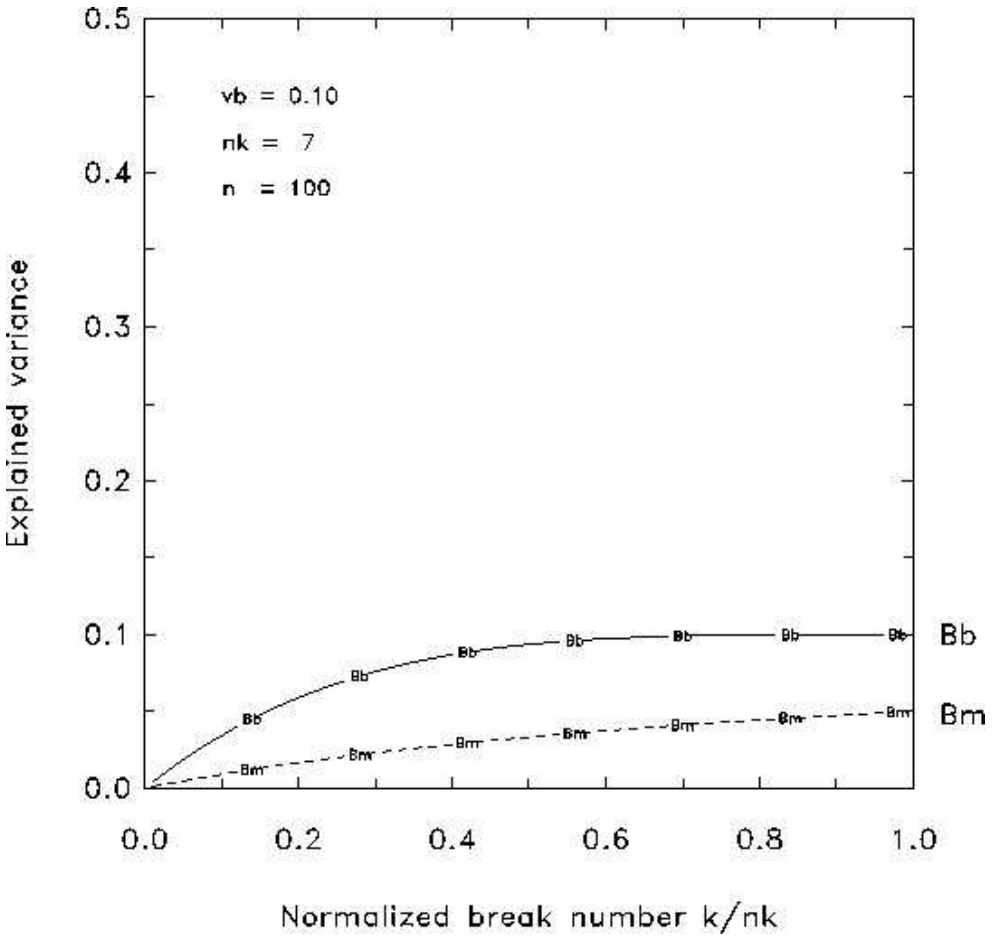
Optimum segmentations of break variance: $$\mathbf{v} = \mathbf{1} - \left(\mathbf{1} - \frac{k}{n_k}\right)^4$$

Optimum segmentations of noise variance: $$\mathbf{v} = \mathbf{1} - \left(\mathbf{1} - \frac{k}{n-1}\right)^4$$

Random segmentations of break variance: $$v = \frac{k}{n_k + k}$$

Random segmentations of noise variance: $$v = \frac{k}{n-1}$$

Best and mean break variance



Signal to noise ratio = 1 / 3.

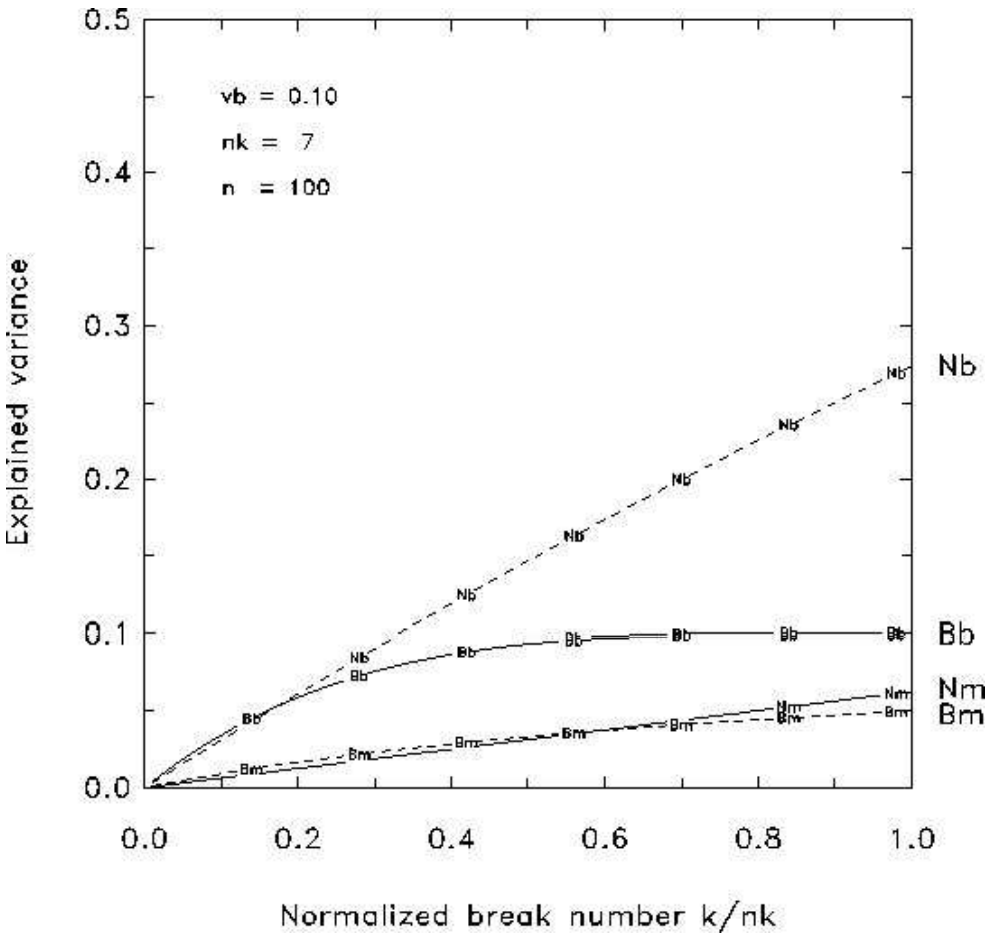
V_{break}= 0.1
V_{noise}= 0.9

Draw known formulas for v(k).

Best break segmentation Bb reaches full break variance early before n_k.

Mean break segmentation Bm reaches half break variance at n_k.

Best and mean noise variance



Best noise segmentation grows with about 4% per break;

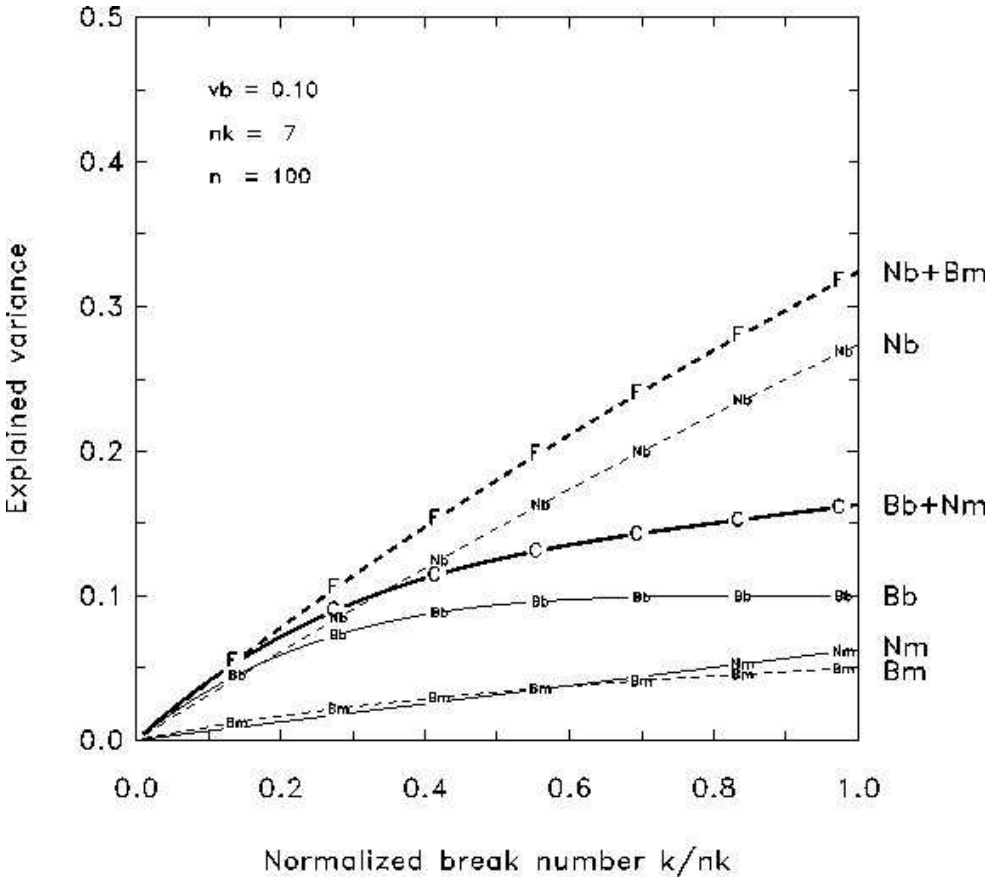
Mean noise segmentation with 1% per break.

The correct segmentation combines the best break Bb with mean noise Nm (solid).

An alternative combination is best noise Nb with mean break Bm (dashed).

Here, only the noise is optimally segmented.

Wrong and right combination

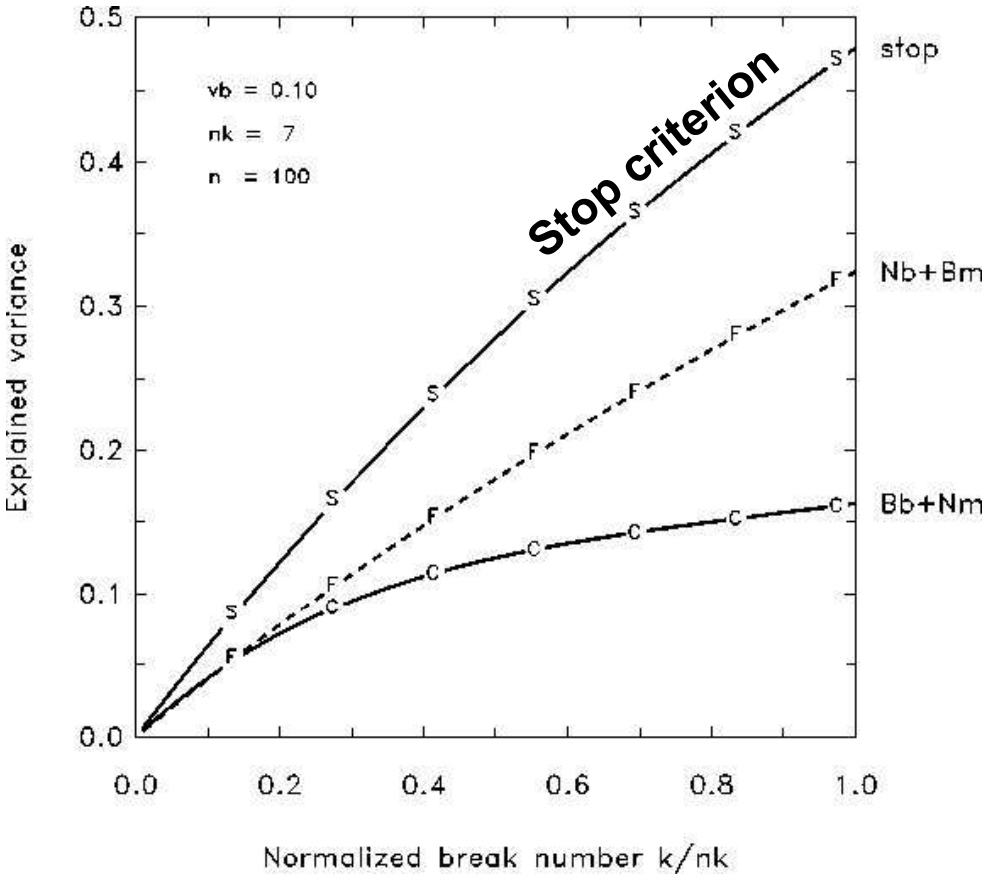


The false (noise) segmentation is always larger than the true (break) segmentation.

However, there is still the stop criterion, which may reject any segmentation at all, preventing in this way these wrong solutions.

Meteorologisches Institut
Universität Bonn

Stop criterion as last help



Only solutions exceeding the stop criterion are accepted.

So it seems that it actually prevents the false combinations.

However, we will see that this is not always the case.

Consider not only the two extremes (completely wrong, completely right), but all transitions in between.

Meteorologisches Institut
Universität Bonn

Break search with simulated data

Create 1000 time series of length 100 with 7 breaks and SNR of 1/3.

Search for the best segmentation and check, which part of the break variance and which part of the noise variance is explained.

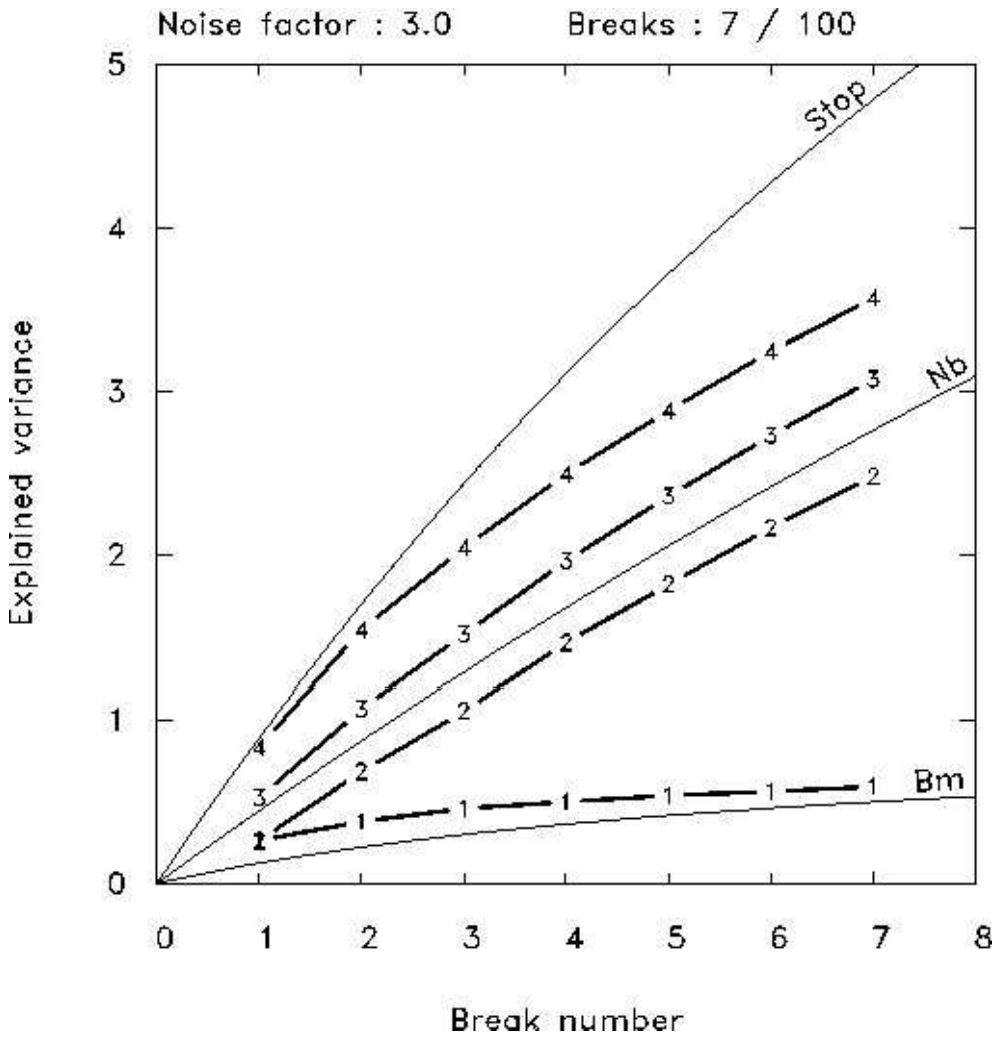
1: Break part

2: Noise part

3: Sum of both

4: Totally explained

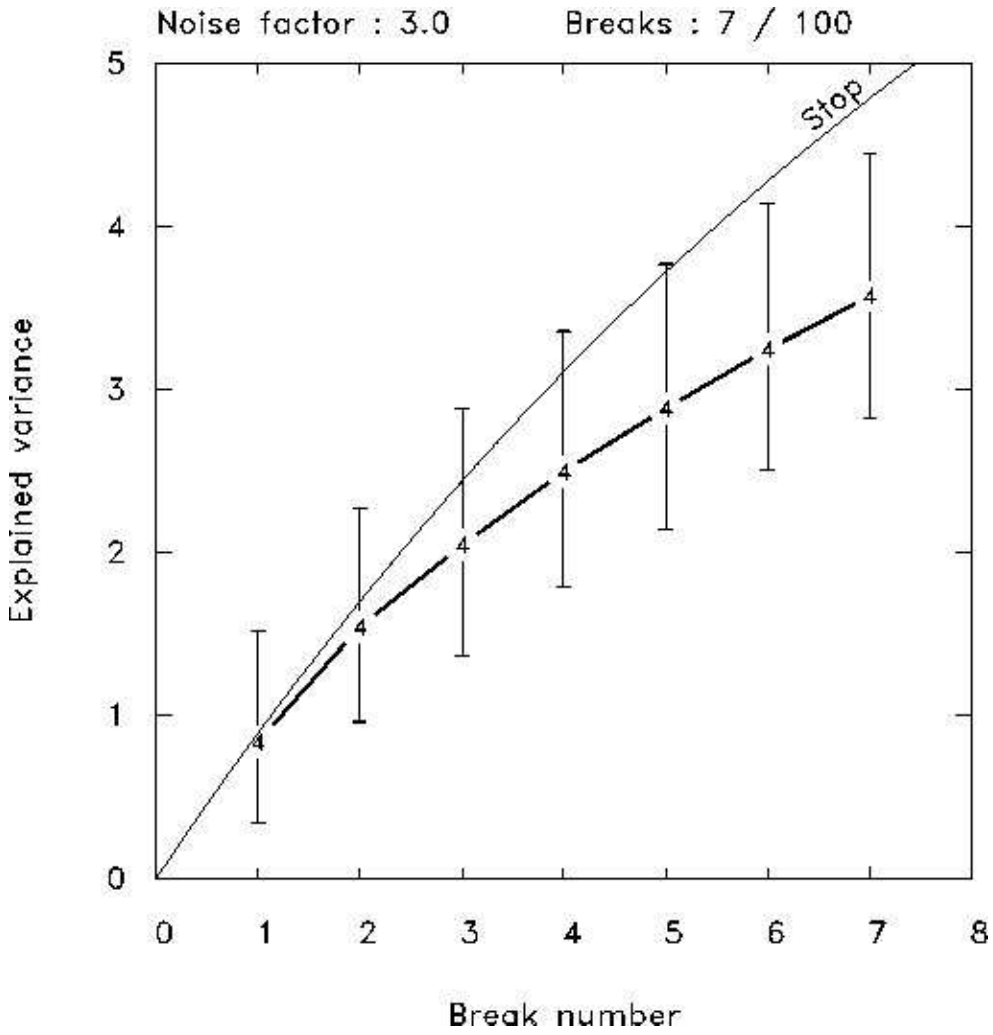
As the best solution is chosen, 1 and 2 are typically correlated, enhancing the total explained variance (4) compared to (3).



Solutions are varying

At first glance, the totally explained variance does not exceed the threshold.

However, up to now we looked at the means over 1000 realisations. But these solutions are varying so that the threshold is often exceeded, at least for low break numbers.



Meteorologisches Institut
Universität Bonn

Conclusions Part II

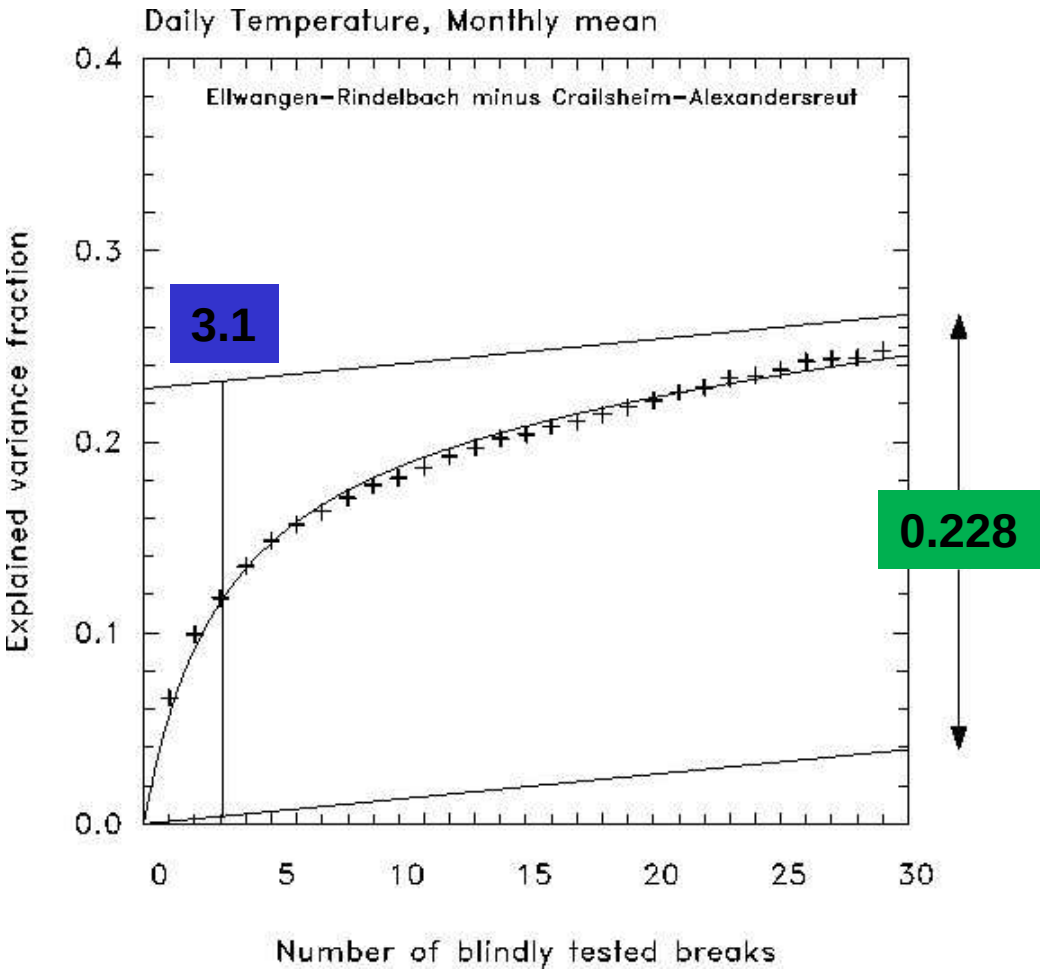
Random segmentations are able to explain a considerable fraction of the break variance. For reasonable break numbers they explain about one half.

Consequently, the breaks are set to positions where a maximum of noise is explained. Hereby, the explained noise part is increased by a factor of four compared to random.

Unfortunately, this is a profitable strategy as the signal part decreases in return only by a factor of 2, compared to the optimum.

Meteorologisches Institut
Universität Bonn

A priori formula



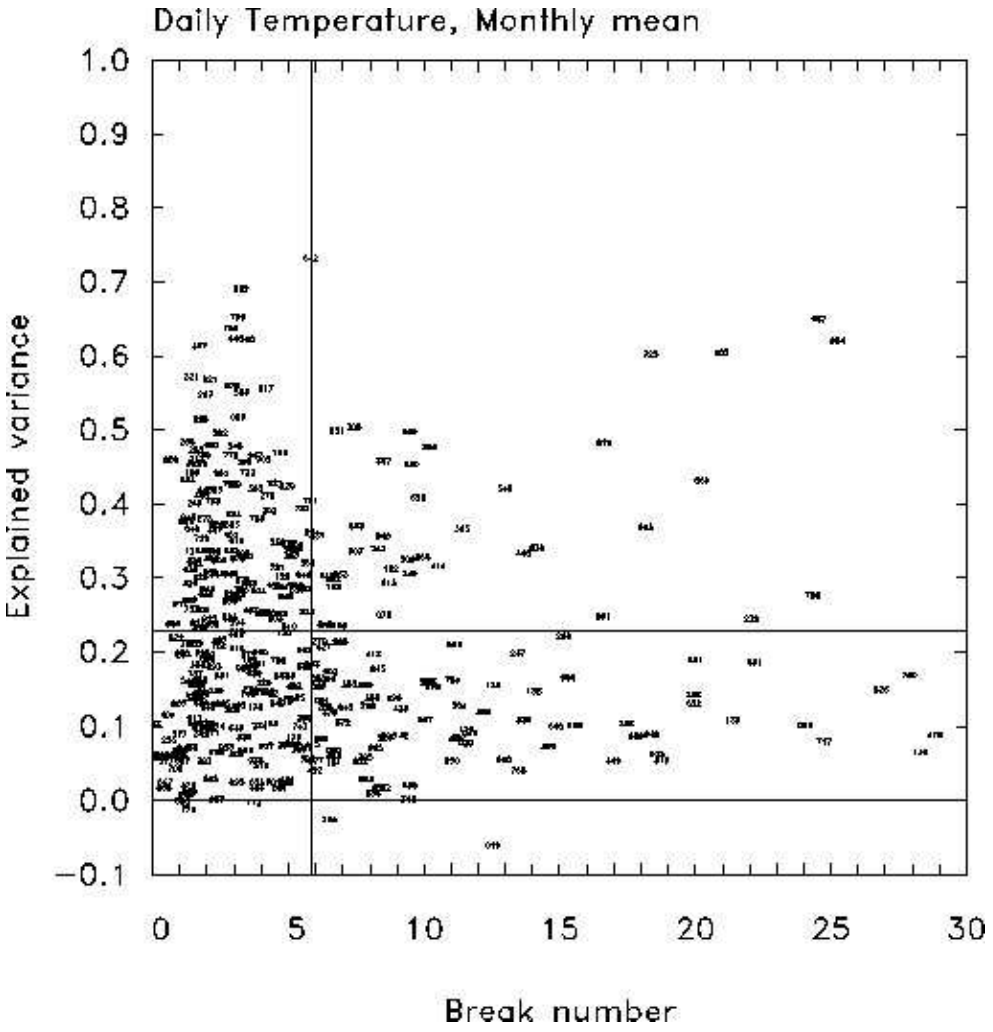
The different reaction of breaks and noise on randomly inserted breaks makes it possible to estimate break variance and break number *a priori*.

If we insert many breaks, almost the entire break variance is explained plus a known fraction of noise.

At $k = n_k$ half of the break variance is reached (22.8% in total).

Meteorologisches Institut
Universität Bonn

Break variance



Repeated for all station pairs we find a mean break variance of about 0.2

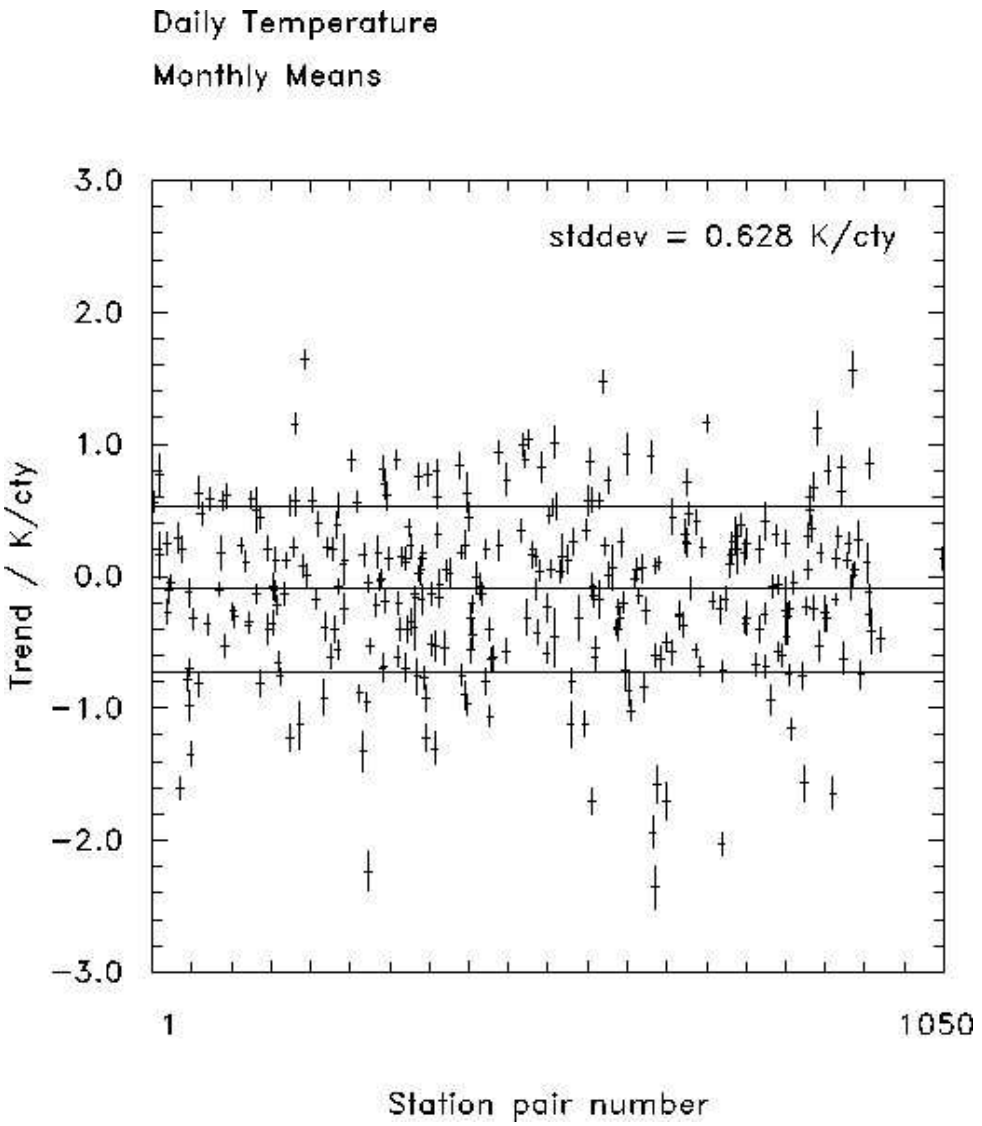
Thus the ratio of break and noise variance is

0.2 / 0.8 = ¼

The signal to noise ratio
SNR = ½

Meteorologisches Institut
Universität Bonn

Trend differences from Data



German climate stations have SNR of 0.5.

Trend differences of neighboring stations reflect the true uncertainty of trends (position of crosses).

Errrors calculated by assuming homogeneous data are much smaller (vertical extend of crosses).

We conclude that the data is strongly influenced by breaks.

Meteorologisches Institut
Universität Bonn

Overall Conclusions

For signal-to-noise ratios of ½ standard break search algorithms are not superior to random segmentations.

This can be understood by considering the theoretical behavior of break and noise variance.

For monthly temperature at German climate stations the SNR can be estimated by an a priori method to ½.

Although the relative break variance might be small, breaks influence the trend estimates strongly.

Interpretation of k/(n-1)

For random segmentation of random scatter the external variance is:

$$v = \frac{k}{n-1}$$

Very short interpretation of k:

more breaks, more external variance

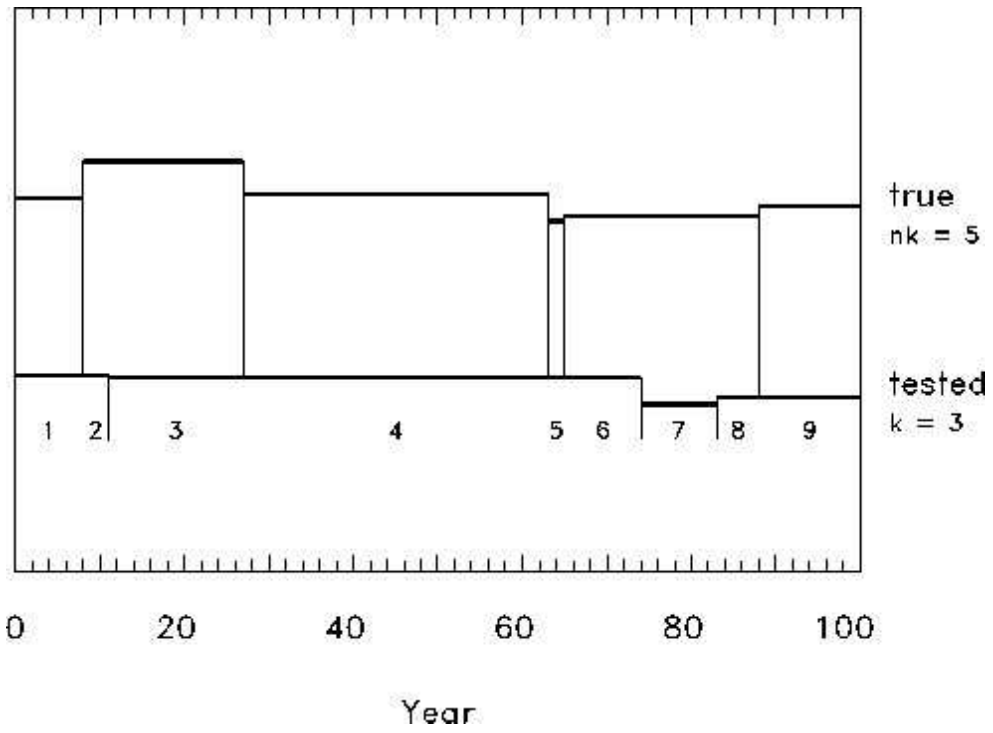
Short interpretation of n:

The external variance is the variance of the segment means.

The more independent values are underlying, the less the means vary and the smaller the external variance is.

Interpret *n* as number of independent values in each segment summed up over all segments.

Why $k/(n_k+k)$?



For a random segmentation of true breaks:

Originally, the time series contains n_k+1 independent values.

Each inserted break *k* cuts a true segment into two pieces, which contribute then to two different tested segments.

The effective number of independents is increased from n_k+1 to n_k+1+*k*.

n*-1 is replaced by n_k+*k

Signal Noise to Ratio = ½

For SNR = ½ , things do not look much better.

The correct combination is slightly better than the false one, but still comparable in magnitude.

In return, the threshold is easier to exceed.