13th FIG Symposium on Deformation Measurement and Analysis
4th IAG Symposium on Geodesy for Geotechnical and Structural Engineering

LNEC, LISBON 2008 May 12-15

AN EVALUATION OF METHODS FOR THE IDENTIFICATION OF VARIANCE CHANGES IN DEFORMATION ANALYSIS

Hans NEUNER

Geodetic Institute, Leibniz University of Hanover, Germany

Abstract: The present paper deals with methods for the identification of variance changes in recorded time series. The first method uses a statistical test with a test value based on the cumulative sums of squares. In the second approach the configuration of the change-points is estimated. In order to avoid influences due to changes of the mean and to separate the different spectral components, the time series are first decomposed using a Wavelet-Transform.

1. INTRODUCTION

In modern deformation analysis the deformation process is described in a system theoretical approach. The most comprehensive description of a system is given by the dynamic model. Therein the reaction of the monitored object is expressed as a function of time and acting loads. If the functional relationships between these two components that represent the input and output to the system are set regardless of its physical structure, the resulting model belongs to the class of behaviour models. If the model doesn't take the acting loads into account and describes the system's state only as a function of time, then it is of kinematic type. The content of this paper refers to these two modelling strategies.

In most cases, when one of the above mentioned models is applied, the data is available as time series. The model parameters result from an analysis in time or frequency domain. Regardless of the processing strategy, it is usual to apply for a global model which includes the entire available observations and treats them undifferentiated. Thus, the statistical homogeneity of the analysed data is presumed implicitly. In terms of time series analysis it is required that the observed processes are stationary up to the 2^{nd} order.

Local effects in processes with varying statistical properties - such as changes of variance - cannot be described by the global approach. Moreover their occurrence leads to biases in the estimated parameters and disturbances in the residuals. Yet it is precisely the expectance of such changes that often motivates the monitoring activity. Therefore, it is a natural way to proceed by trying to identify these kinds of changes and by modelling them appropriately.

This paper deals with two methods of automatic identification of variance changes in time series. It is assumed that these changes are performed quickly compared to the length of the time series and are followed by a new state of statistical equilibrium. This assumption conforms to the treatment in dynamic analogy models and to practical reality. Due to their prior identification one can account for the changes and maintain a reduced order of the model at

the same time by introducing a new set of suitable parameters for each interval of homogeneous variance. Furthermore, as a result of a proper transformation of observations to data with unitary variance, one can obtain system parameters like the reaction delay which cannot be accurately calculated in the global approach.

In order to assess the variance homogeneity of the different components that overlay in the time series (especially the one of the deformation) these are decomposed in spectral components by means of the Wavelet Transform. The identification methods are applied in a second stage at the level of the wavelet coefficients.

The paper is organised as follows: the next section introduces basic concepts of the Wavelet Transform. The theoretical backgrounds of the identification methods are given in the third section. Results obtained by applying them to synthetic and real signals are presented in the last section.

2. THE WAVELET-ANALYSIS OF TIME SERIES

The Wavelet Transform offers the possibility to extract and study local characteristics of the signal at subsequent resolution levels. This can be done in different ways according to the purpose of the data decomposition (Percival and Walden, 2002). If further processing of the transformed data is needed the Discrete Wavelet Transform (DWT) is appropriate.

The DWT is formulated in terms of an orthogonal filter bank and consists in passing the low frequent signal component u, separated in a previous step j, through a quadrature mirror filter pair and decimating by 2 the components by retaining every other value:

$$u_{j+1,n} = \sum_k h_{k-2n} \cdot u_{j,k} \qquad v_{j+1,n} = \sum_k g_{k-2n} \cdot u_{j,k}. \tag{1}$$

The high- and low-pass filters of the bank, h and g, are called wavelet and scaling filter. The corresponding outputs, v and u, are the wavelet and scaling coefficients.

To interpret the physical meaning of the coefficients that result in the scale $s_j = 2^j$, the recursive filtering can be replaced by single equivalent filters. Their transfer functions are derived from the transfer functions of each applied scaling and wavelet filter. The pass-band of these equivalent filters is given by $0 - 1/2^{j+1}$ for the low-pass filtering cascade and $1/2^{j+1} - 1/2^j$ for the sequential filtering that includes the high-pass filter at the level j (Percival and Walden, 2002). Thus the coefficients of the j^{th} decomposition level represent the spectral components of the signal contained in the pass-band of the equivalent filters.

Because the used filters are of FIR-type signal components with frequencies in the transition band may occur in the spectrum of the wavelet coefficients of adjacent scales. The separation between the spectral components is stricter if the characteristic equation H(f) = 0 has a root f = 0.5 of increased order. H denotes herein the transfer function of the low-pass filter. The higher the root's order the better the frequency localisation is. This increase also leads to an increase of the filter length which on the other hand has a negative effect on the aimed localisation in time of the variance changes. These opposite effects make the selection of a filter structure necessary that represents a good balance between the localising properties in time and frequency domain, according to the scope of the analysis. In the present paper the DAUBECHIES-filter of the 4^{th} order (Percival and Walden, 2002) seemed to be a good trade-off.

The DWT is due to its orthogonality an isometry. This means that the energy of the analysed signal equals the total energy of the coefficients resulting from the transformation:

$$\|X\|^2 = \sum_{j=1}^{J} \|U_j\|^2 + \|V_J\|^2 \tag{2}$$

where U_j and V_j stand for the vector of scaling and wavelet coefficients at scale 2^j and J for the maximal decomposition level.

This property is similar to the FOURIER-transform and enables the separate analysis of the contributions of each scale to the total variance. Thus, it is possible to follow the build-up of the variance in each scale and hence to study the change of variance at a scale-by-scale basis. The change-points of the variance in each scale can be identified with one of the approaches presented in the next section.

3. METHODS FOR THE ASSESSMENT OF VARIANCE HOMOGENEITY

3.1. The variance homogeneity test

There are various statistical approaches available that aim the assessment of variance homogeneity. However a great amount of these are designed for the identification of a single variance change or suffer from the lack of localising properties. A test procedure that avoids these disadvantages was introduced by INCLAN and TIAO (1994). They investigated the detection of multiple variance changes in a sequence of N uncorrelated normally-distributed variables x with mean 0 and variance σ_k^2, with k = 1, 2, .., N, starting from the formulation of the following null and alternative hypothesis:

$$H_0 : \sigma_1^2 = \sigma_2^2 = K = \sigma_N^2 \tag{3}$$

$$H_A : \sigma_1^2 = \sigma_2^2 = K = \sigma_k^2 \neq \sigma_{k+1}^2 = K = \sigma_N^2 \tag{4}$$

The test value, used to distinguish between these two hypothesis, is based on the centred cumulative sums of squares (CCSS):

$$D_k = \frac{C_k}{C_N} - \frac{k}{N} \text{ where } C_k = \sum_{i=1}^{k} \left(\frac{x_i^2}{\sigma_i^2} \right) \text{ and } k = 1, 2, K\ , N \tag{5}$$

The ratio between the sums of squares follows a beta-distribution if the null hypothesis is valid. The expected value of this ratio can be calculated straightforward from the moment generating function of the beta-distribution and fits the term subtracted in (5) from the ratio of the sums of squares. Thus, the measure D_k can be interpreted as the deviation of the ratio of cumulative sums of squares from their expected value.

It was proven by INCLAN and TIAO (1994) that $\sqrt{N/2} \cdot D_k$ resembles in the asymptotic case $N \to \infty$ a BROWNian bridge process. For this kind of process useful probability measures have been obtained by BILLINGSLEY (1999). It can be noticed, that a change of variance at a certain location k* leads to an increased value of $|D_k|$ for $k = k^*$ but also for values of k in the neighbourhood of k*. Therefore it is reasonable to choose $\max_k|D_k|$ as an appropriate test

value for the assessment of the variance homogeneity of the time series. Based on the probability measures given in BILLINGSLEY (1999) one obtains:

$$P\left[\sqrt{\frac{N}{2}}\max_k|D_k|\le b\right]=1+2\sum_{i=1}^{+\infty}(-1)^i e^{-2i^2b^2},\ b>0 \tag{6}$$

For the confidence level of 95% adopted in this paper the resulting quantile is b = 1.358.

Performed simulation studies have revealed that this value is usable in practice for time series with more than 5,000 terms. For shorter lengths of the data series a lower value was obtained. However, the derivation by simulations of proper quantiles that conforms to the length of the analysed data is of high computational burden. Therefore the following closed relation is proposed for the calculation of the quantiles in case of time series with less than 5,000 terms:

$$P\left(\max_k|D_k|\ge\frac{a}{N'+1}\right)=2(a+1)\cdot \sum_{j=[a]+1}^{N'}\left\{\left[\frac{N'}{2\pi j(N'-j)}\right]^{\frac{1}{2}}\frac{1}{(a+N'+1-j)}\left(\frac{N'-j}{j}\cdot\frac{j-a}{a+N'+1-j}\right)^j\left[\prod_{i=1}^{l}\left(\frac{N'}{N'-j}\cdot\frac{a+N'+1-j}{N'+1}\right)^{\frac{N'}{l}}\right]\right\} \tag{7}$$

In (7) [a] is the next lower integer to a, and N' = N/2. This probability relation was obtained on the basis of the results in DURBIN (1968) for the k^{th} order statistic of N independent variables of the (0, 1) uniform distribution. This statistic is beta-distributed. Thus, the probability relations derived in DURBIN (1968) are also valid for D_k. The original relation of DURBIN was adapted using STIRLING's formula for the approximation of faculty. The approximation series was interrupted after the first order term. Therefore, the relation (7) is only valid for $N \le 5{,}000$. Nevertheless it is a good complement to the relation (6). The results obtained for the 95% quantiles agree up to 10^{-3} with the results of the simulations.

The structures of the test value (5) and the one used in the KOLMOGOROV-SMIRNOV goodness-of-fit test are similar. The agreement between the empirical distribution function and an assumed theoretical distribution function is assessed in the latter by means of the maximum absolute value of their difference. It was shown that this difference resembles the BROWNian bridge process, in case of compliance. This analogy opens the possibility to build alternative measures for the assessment of the variance homogeneity by using test values of other goodness-of-fit tests that evaluate the difference between the distribution functions. One of these is the CRAMÉR - von MISES test. It's based on the sum of the squared differences between the distribution functions. Applying for the mentioned analogy the resulting measure is:

$$\frac{1}{2}\sum_{k=1}^{N}D_k^2=\frac{1}{2}\sum_{k=1}^{N}\left(\frac{C_k}{C_N}-\frac{k}{N}\right)^2 \tag{8}$$

This measure has a global character and doesn't possess localising properties. Thus, it can only signalise the heterogeneous variance of the analysed time series if its estimation exceeds the quantile of the CRAMÉR – von MISES test. One aim of the simulations presented in the following section is to compare the power of the two presented measures.

If the test value in (6) or (7) exceeds the chosen 95 % quantile, then k* corresponding to $\max|D_k|$ is marked as a possible change-point. To check for further variance change-points

the samples are divided into subseries $[x_1..x_{k^*}]$ and $[x_{k^*+1}..x_N]$ and the procedure is applied again for each of them. This process is iterated until the test value falls below the quantile for every subset of variables $[x_{t1}..x_{t2}]$. In a further step the previously obtained candidates and their location are validated in an iterative algorithm (Inclan and Tiao, 1994).

One drawback of the measure (5) is its lack of robustness. This follows from figure 1 which contains the sensitivity curves SC(x) calculated for D_k at the points k = N/4, N/2 and 3N/4:

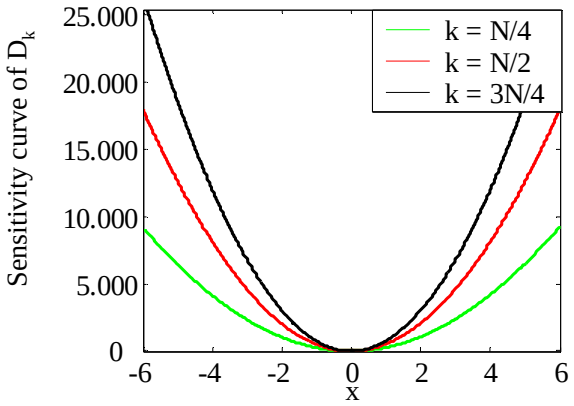


Figure 1 - Sensitivity curves SC(x) of the measure D_k, calculated for k=N/4, N/2 and 3N/4

As can be observed, the measure D_k becomes more sensitive to outliers as k approaches the end of the series. Outliers have a direct impact on the test results as they cause local maxima of $|D_k|$ at the time of their occurrence. Therefore, they lead to misinterpretation by feigning variance change. In order to circumvent this false conclusion we propose to use the above testing procedure based of the following modified version of D_k:

$$D'_k = \frac{C'_k}{C'_N} - \frac{k}{N} \text{ where } C'_k = \sum_{i=1}^{k} \left(p_i^2 x_i^2 \right) \text{ and } k = 1, 2, K, N \tag{9}$$

The introduced weights p_i account for the "distance" between x_i and the adjacent data situated in a neighbourhood of a certain extension. They are defined as follows:

$$p_i = \begin{cases} \dfrac{1}{\sigma_i} & \text{if } \left| x_i - \text{med}(x) \right| < c \cdot \sigma_i \\ \dfrac{c}{\left(\left| x_i - \text{med}(x) \right| \right)} & \text{if } \left| x_i - \text{med}(x) \right| \geq c \cdot \sigma_i \end{cases} \tag{10}$$

In (10) med(x) denotes the median of the series x, σ_i denotes the value of the standard deviation estimated on the basis of the median absolute deviation (MAD) in the neighbourhood of x_i, and c represents a constant between 2.5 and 4. The introduced weighting scheme is similar to the one proposed by HUBER in the robust adjustment of GAUSS-MARKOV linear models but uses the more robust median for the estimation of the statistic parameters.

Strictly, the probability relations (6) and (7) are no longer valid for the modified CCSS in the presence of outliers. However, for small ratios of outliers to unbiased data the available quantiles proved to be useful further on. This can be explained by the much higher benefit ob-

tained from the mitigation of the influence of outliers on D_k compared to the deviation of the distribution of the modified measure D_k' from the assumed beta-distribution.

3.2. The identification method based on contrast functions

The approach of this method is different from the strategy presented in 3.1 because it estimates the entire configuration of change-points at a time. It was developed by LAVIELLE (1998) and is based on the likelihood function. For the outline of the method's theoretical background we consider a set of N variables x_i, with i =1, 2,.., N, which have a density function f depending on the unknown parameters θ. At least one of these parameters is changing K times during the observation period. At first we consider a known number of changes K and define as primary goal the estimation of the locations of these change-points and of the parameters θ according to a minimising criterion. The minimised function is called contrast function γ and the obtained estimators are named minimum contrast estimators. By expressing the contrast function γ with respect to the empirical data, one obtains the empirical contrast γ_n. The minimum contrast estimator $\hat{\theta}$ is minimising the empirical contrast.

For density estimation problems the log-likelihood function l is a proper choice of a contrast function provided that the density function is continuous. This condition is satisfied by the normal density function which is also appropriate if the changing parameters θ refer to the mean μ and/or the variance σ^2 of the observations. The changing parameters are calculated as the ml-estimates from the observations x_k belonging to each homogeneous segment k between the times $t_{k-1}+1$ and t_k. The other parameters are calculated using the entire set x. The configuration τ of the changes is then obtained by minimising the empirical contrast:

$$\gamma_N(\tau) = \frac{1}{N}\sum_{k=1}^{K} -l\left(x_{t_{k-1}+1}, x_{t_{k-1}+2}, ..., x_{t_k}; \hat{\theta}_k\right) \rightarrow \min. \tag{11}$$

When the changes affect the variance the log-likelihood function in (10) has the following structure:

$$l\left(x_{t_{k-1}+1}, x_{t_{k-1}+2}, ..., x_{t_k}; \hat{\sigma}_k^2\right) = \frac{N_k}{2}\left[1+\ln(2\pi)\right] + \frac{N_k}{2}\ln\left[\frac{\sum_{j=1}^{N_k}\left(x_{t_{k-1}+j}-\mu\right)^2}{N_k}\right] \tag{12}$$

where N_k is the number of samples in the k^{th} homogeneous set. Similar relations can be obtained if the changes affect either the mean or both statistical parameters. The minimising of (11) can be solved only using a combinatorial approach because τ is not explicitly contained in the functional model.

In the more practical case when the number K of changes is unknown the minimisation function (11) has to be extended with a penalisation factor that includes this additional unknown:

$$\gamma_N(\tau) = \frac{1}{N}\sum_{k=1}^{K} -l\left(x_{t_{k-1}+1}, x_{t_{k-1}+2}, ..., x_{t_k}; \hat{\theta}_k\right) + \text{pen}(K) \rightarrow \min. \tag{13}$$

Solving this minimisation problem requires some prior knowledge about the maximum number of change-points K_{max}. Thus, with a proper choice of the penalisation factor one may apply the same computation procedure as in the case of a known number of change-points with

$K = K_{max}$. The estimation problem is now enhanced by an aspect of model selection. It is known from similar model selection problems that an optimal penalisation factor will compensate for the decrease of the empirical contrast (11) after attaining a plausible number of change-points. MASSART (2003) derives two structures for the penalisation term depending on the selection procedure for the set of change-points. In case of an ordered selection, performed by adding successively one element to the set of already selected change-points, a linear penalisation of the form $\kappa \cdot D/N$ is suggested. Herein is D the number of assumed change-points and κ a scale factor. In case of a complete selection, performed by analysing anew all possible locations of an assumed number of change-points, a linear-logarithmic penalisation of the type $\kappa \cdot D\left[1+\log\left(2N/D\right)\right]/N$ is proposed. Both penalisation forms were implemented for the practical applications discussed in the 4th section.

4. THE EVALUATION OF THE TWO METHODS FOR THE IDENTIFICATION OF VARIANCE CHANGES

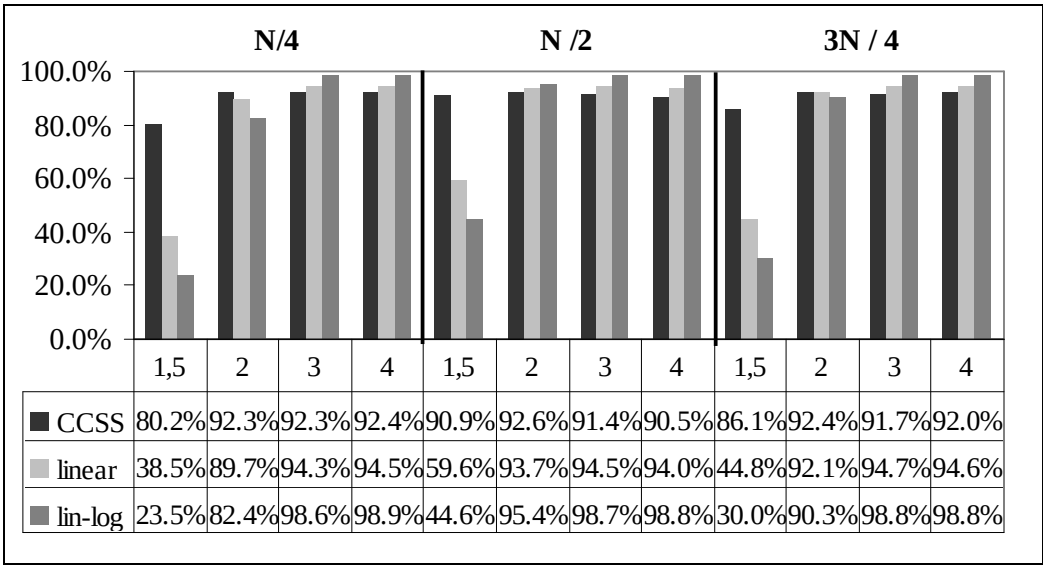
4.1. The evaluation based on synthetic signals

For the performance assessment of the described identification methods numerous analysis were done using synthetic signals. Within the simulations the length of the time series, the location and the magnitude of the variance change has been varied. For every combination of these parameters 10.000 simulations were performed.

At first, the quantiles of the probability relations (6) and (7) as well as for the CRAMÉR-von MISES test could be confirmed by means of series with homogeneous variance. A single change-point was introduced by increasing the variance of all variables that follow the chosen location. The increase was of the order of 1.5, 2, 3 and 4 and its location was varied in quarters of the total length. For every combination of the parameters only the correct identification of just one variance change was marked as successful. The identification with the variance homogeneity test described in 3.1 was performed with both test values derived from the KOLMOGOROV-SMIRNOV and the CRAMÉR – von MISES goodness-of-fit test. Although in all tests the successful identification rates were close the KOLMOGOROV-SMIRNOV based statistic led systematically to better results. Therefore, the following results refer exclusively to the test performed with this measure.

The expected dependency of the successful identification rates on the length of the time series and on the magnitude of the change could be confirmed for both introduced methods. Figure 2 shows exemplary the successful identification rates obtained with the two methods for time series of length N=1,000 subjected to the location and the magnitude of the change.

One observes that for changes of small magnitude the CCSS-based method seem to perform better while for greater changes the method based on contrast function with the combined linear-logarithmic penalisation is superior. Thus the two methods complement each other very well. The comparison of the two penalisation forms available for the contrast-based method reveals that the linear penalisation leads to better results exactly in the situations where the variance homogeneity test leads to the very best results. Therefore, the method based on contrast functions with linear penalisation gets no further attention.



	1,5	2	3	4	1,5	2	3	4	1,5	2	3	4
CCSS	80.2%	92.3%	92.3%	92.4%	90.9%	92.6%	91.4%	90.5%	86.1%	92.4%	91.7%	92.0%
linear	38.5%	89.7%	94.3%	94.5%	59.6%	93.7%	94.5%	94.0%	44.8%	92.1%	94.7%	94.6%
lin-log	23.5%	82.4%	98.6%	98.9%	44.6%	95.4%	98.7%	98.8%	30.0%	90.3%	98.8%	98.8%

Figure 2 - Successful identification rates for series with N = 1.000 samples

Although the main conclusions of the simulation study were drawn here from the results presented in figure 2 they are also supported by the results obtained for time series with N = 500 and 10,000 samples.

Another study performed on synthetic signals aimed to assess the robustness of the modified test value (9). Therefore, the identification procedure presented in section 3.1 was applied with both test values (5) and (9). In cases when no error was introduced to the data the measures led to identical results. In the presence of outliers the measure (9) performed significantly better. Whereas the localisation of the variance change improves, the number of rejections of the null hypothesis (3) remains comparable. For example, in time series with N = 500 samples that contains a variance increase of 1.5 in the middle and a gross error at the location k* = 375 the test was rejected in all simulations when using the measure (5) and in 89 % of cases when using the measure (9). However, while the initial test locates the change correctly only in 65 % of the simulations the modified test leads to 81 % successful identifications. In cases of strongly biased data it is therefore advisable to apply for the modified test value (9) when the variance homogeneity test is used.

The gained cognitions about the performance of the introduced methods are further used to assess the identification results of real recorded data and to understand possible differences between the identification results.

4.2. An evaluation based on recorded time series

The analysed time series represents deformations of the tower of a wind energy turbine occurring due to operating and wind loads. For reasons mentioned in the 2nd section the signal was first transformed using the DWT. In the transformation up to the 4th decomposition level all spectral components of the time series were projected onto the wavelet coefficients. A main purpose of this monitoring was the modelling of the components corresponding to the first eigenfrequency of the tower. Therefore, the following results are focused on the coefficients of the 3rd and 4th decomposition level which contains the eigenfrequency. This spectral com-

ponent overlaps the rotation frequency of the turbine in the coefficient series of the 4th level and the blade frequency in the coefficients of the 3rd level.

The identification results obtained for the wavelet coefficients of the 4th decomposition level with the two discussed methods are presented in the following figure:

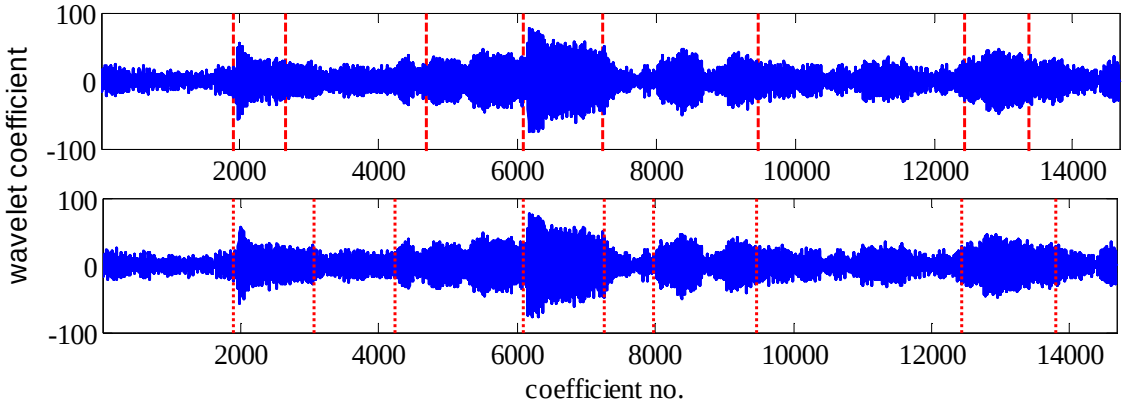


Figure 3 - Variance changes identified in the series of 4th level wavelet coefficients. Top: results of the variance homogeneity test. Bottom: results of the contrast-based method.

As can be observed the number of changes and their locations agree very well. Solely one further change-point situated nearly 8000 was identified with the contrast-based method. Its existence is justified because a similar change was identified in series of the coefficients of the 3rd level. The lower sensitivity of the test in this situation is caused by the masking effect due to the high energy segment situated just before the unidentified location. This segment is longer than the lower energy interval. The same effect could be observed in the study of synthetic signals. One notice from figure 2 that the identification rates of a change occurring at the location 3N/4 are systematically higher than the rates obtained when the change is located at N/4.

The good agreement between the results obtained with the two identification methods could be confirmed for the wavelet coefficients of the 3rd level only for the section starting at 5400 (s. figure 4). Up to this location the number of identified change-points and their corresponding locations are quite different. The main cause for this disagreement is the increased value of the coefficients around the locations 2000. As discussed in the 3rd section the CCSS-based test is very sensitive to increased values of single samples (s. figure 1). These have only a logarithmic contribution to the contrast function as can be noticed from (12). In a straightforward approach one can use the variance homogeneity test with the modified test value (9). The obtained results are marked in figure 4 with green lines. The identical change-points are marked with half-lines. Thus, an equal number of change-points could be obtained. Nevertheless the difference between the locations of the first change-point is still significant and its explanation needs further research.

Some of the identified segments with homogeneous variance correspond to the structure's reaction to changes of the operating states. Thus, the first, the fourth and the eighth change-point occur due to changes of the azimuth of the nacelle or to changes of the pitch. The change-points following the abovementioned ones mark the end of the transition between two equilibrium states. The remaining change-points are caused by variations of the wind load.

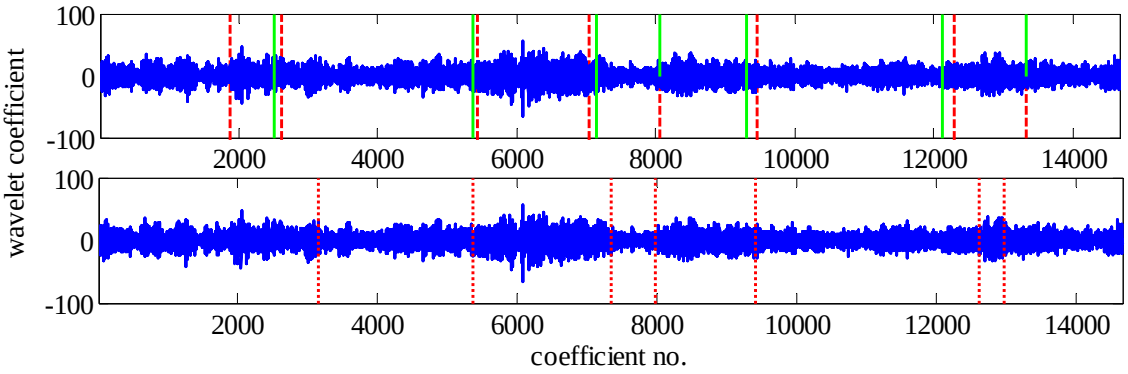


Figure 4 - Variance changes identified in the series of 3rd level wavelet coefficients. Top: results of the variance homogeneity test (red) and modified variance homogeneity test (green). Bottom: results of the contrast-based method

Relating the sudden changes of the acting loads to the deformations from the segments with homogeneous variance corresponding to the transitions enables the derivation of specific parameters that characterise i.e. the structure's stiffness. Therefore, the proposed approach is consistent with the objectives of modern deformation analysis discussed in the 1st section. Moreover, the improvement obtained by accounting for the identified variance changes in a kinematic modelling of the deformation signal encourages further research on the presented topic. Future activities will deal especially with the assessment and improvement of the robustness of the presented identification methods.

References

Billingsley, P. (1999): *Convergence of Probability Measures*. 2nd Edition, Wiley, New York.

Durbin, J. (1968): *The Probability That the Sample Distribution Function Lies Between Two Parallel Straight Lines*. The Annals of Mathematical Statistics, Vol. 39, No. 2, p. 398 – 411, Waverly Press, Baltimore, USA.

Inclán, C., Tiao, G. (1994): *Use of Cumulative Sums of Squares for Retrospective Detection of Changes of Variance.* Journal of the American Statistical Association, Theory and Methods, Vol. 89, No. 427, p. 913 – 923.

Lavielle, M. (1998): *Optimal Segmentation of Random Processes*. IEEE Transactions on Signal Processing, Vol. 46, No. 5, p. 1365 - 1373.

Massart, P. (2003): *Concentration Inequalities and Model Selection*. Berlin: Springer.

Percival, D. B., Walden, A. T. (2002): *Wavelet Methods for Time Series Analysis.* Cambridge University Press, Cambridge.

Corresponding author contacts

Hans NEUNER
neuner@gih.uni-hannover.de
Geodetic Institute, Leibniz University of Hanover
Germany