Analysis and selection of pan-sharpening assessment measures

Aliaksei Makarau,Gintautas Palubinskas, and Peter Reinartz

German Aerospace Center (DLR), Remote Sensing Technology Insitute, Oberpfaffenhofen, Germany

ABSTRACT

Pan-sharpening of remote sensing multispectral imagery directly influences the accuracy of interpretation, classification, and other data mining methods. Different tasks of multispectral image analysis and processing require specific properties of input pan-sharpened multispectral data such as spectral and spatial consistency, complexity of the pan-sharpening method, and other properties. The quality of a pan-sharpened image is assessed using quantitative measures.

Generally, the quantitative measures for pan-sharpening assessment are taken from other topics of image processing (e.g. image similarity indexes) but the applicability basis of these measures (i.e. whether a measure provides correct and undistorted assessment of pan-sharpened imagery) is not checked and proven. For example, should (or should not) a quantitative measure be used for pan-sharpening assessment is still an open research topic. Also there is a chance that some measures can provide distorted results of the quality assessment and the suitability of these quantitative measures as well as the application for pan-sharpened imagery assessment is under the question.

The aim of this paper is to perform statistical analysis of widely employed measures for remote sensing imagery pan-sharpening assessment and to show which of the measures are the most suitable for use. To find and prove which measures are the most suitable, sets of multispectral images are processed by the General Fusion Framework method (GFF) with varying parameters. The GFF is a kind of the General Image Fusion method. Variation of the method parameter set values allows to produce imagery data with predefined quality (i.e. spatial and spectral consistency) for further statistical analysis of the assessment measures. The use of several main multispectral sensors (Landsat 7 ETM+, IKONOS, and WorldView-2) imagery allows to assess and compare available quality assessment measures and illustrate which of them are most suitable for each satellite. Experimental analysis illustrates adequate assessment decisions produced by the selected measures for the results of representative pan-sharpening methods.

Keywords: P an-sharpening, quality assessment, multispectral image, multiresolution, spectral consistency, spatial consistency, phase congruency

1. INTRODUCTION

Pan-sharpened remote sensing data have many areas of application, therefore different requirements are posed on the pan-sharpened data quality. The requirements can be on spectral consistency, spatial consistency or on both together. Spectral consistency assumes that the pan-sharpened image has increased spatial resolution with spectral properties of the original image. Spatial consistency assumes that "A high spatial quality merged image is that which incorporates the spatial detail features present in the panchromatic image and missing in the initial multispectral one".[1] The ideal case of pan-sharpening is the highest spectral and spatial consistency together.

To meet the requirements on the fused image quality, pan-sharpened image is assessed using numerical assessment measures, mostly taken from other areas of image and signal processing. The employment of such measures is wide but the applicability and comparison of the measures, as well as recommendations on the use in the sense of pan-sharpened image assessment is not given.

In most papers several assessment measures are employed and the decision on the pan-sharpening quality is made by taking into account all the numerical scores calculated by the measures. In many cases several measures can give contradictory results (see e.g.[2]) and the decision on the pan-sharpening quality is difficult to make. Such contradictory results may be caused by the fact that the measures are inappropriate for such use. Therefore the question on the applicability of the measures should be made clear, and the motivation of this research is

analysis and comparison of widely employed pan-sharpening assessment measures and to show which of them are most suitable for the use.

Inspired by the work of Avcibas et. al.[3] on comparison of image quality measures (sensitivity to different types of distortions in images), in this paper a comparison of pan-sharpening assessment measures for remote sensing is carried out on a specially generated test set of images. The test set is composed of remote sensing pan-sharpened images, produced with different quality (spectral and spatial consistency). Analysis of variance and pairwise comparison statistical methods are performed on the assessment measures. The difference to[3] is that the pan-sharpening assessment measure is required to be sensitive to the pan-sharpened imagery quality change (i.e. able to separate imagery with different quality) as well as sensitive to the increase or decrease of the image quality (i.e. provides the increase or decrease of the assessment score). The measures that are most sensitive to the quality change (according to statistical assessment results) are recommended for use.

2. INFLUENCE OF PAN-SHARPENING ON MULTISPECTRAL DATA ANALYSIS

Influence of pan-sharpening on analysis and processing of multispectral remotely sensed data is widely discussed by Yun Zhang in.[4] In this section we refer to this review work. Several important and widely used methods such as classification,[5] change detection,[6] feature extraction,[7,8] urban area growth,[9] land cover mapping[10] illustrate advantages of pan-sharpened data use. Such methods are evaluated on pan-sharpened multispectral data and Table 1 presents influence results of pan-sharpening method on multispectral data analysis.

Most of the works report an increase of interpretation, analysis, or exploration accuracy during pan-sharpened multispectral data use comparing to the use of the original multispectral data. Therefore, pan-sharpening becomes an important preprocessing step for multispectral data especially for applications dealing with high and very high resolution imagery.

Table 1. Influence of different pan-sharpening methods on the accuracy of multispectral data use (information taken from[4])

Use of multi-spectral data	Pan-sharpening Method	Results
Classification[5]	Adaptive Image Fusion	Maximum Likelihood classification achieved the best classification accuracy with an average overall accuracy of 80% for most fusion methods, whereas the average overall accuracies of object-based classification and support vector machines were 75%
Change detection[6]	Synthetic Variable Ratio (SVR), Principal Component Analysis (PCA), High Pass Filtering (HPF)	PCA, SVR and HPF achieved a higher overall accuracy than that of the reference image (original multispectral Ikonos). The maximum overall accuracy obtained from PCA fused Ikonos image reached 94%
Landslide monitoring[11]	PCI Pansharp	Maximum Likelihood classification to identify large scale landslides. PCI Pansharp was judged to be the most satisfactory, reaching the quality comparable to an aerial ortho-photograph of 1:10,000 scale. It was found that PCI Pansharp enabled detailed interpretation of landslides and associated environmental features
Road extraction[12]	PCI Pansharp	Unsupervised fuzzy K-means clustering method was used to classify the pan-sharpened QuickBird image obtaining a binary road image. The completeness value reached 0.94, and the correctness value was 0.98. The results demonstrated that the proposed method achieved significantly higher accuracy than multispectral classification or multispectral and Pan integrated classification
Urban sprawl monitoring[13]	PCI Pansharp	Pan-sharpened ETM+ images enabled the discrimination of finer change detail than the original ETM+ multispectral images
Supervised segmentation[14]	PCI Pansharp	Pan-sharpening was used to improve the segmentation detail. A noticeable accuracy improvement was observed and the segmentation efficiency was significantly increased (from hours or days to minutes or tens of minutes)
Urban land cover mapping[9]	Discrete Wavelet Transform	Object-oriented classification of eCognition and Maximum Likelihood classification were carried out to classify both the original Ikonos multispectral image and the wavelet fused Ikonos MS image. Accuracy assessment showed a general classification accuracy increase with the fused data, compared to that of the original data
Crop classification[15]	Unspecified	Pan-sharpened QuickBird 0.61m image provided the highest overall accuracy 85.2%

3. NUMERICAL MEASURES FOR PAN-SHARPENING ASSESSMENT

A pan-sharpening method may provide perfect spectral consistency together with a poor spatial consistency and vice versa. Therefore, to make a proper assessment of fusion results, assessment of both spectral and spatial consistency should be performed. The most known and used measures are designed for spectral consistency assessment, while there are not many for the spatial consistency assessment.

3.1 Spectral consistency

Spectral consistency assessment is usually performed using Wald's protocol in order to have a multispectral image of high resolution. There is a variety of developed and well-known similarity measures used for spectral consistency assessment. The most known and popular are: Spectral Angle Mapper, SAM (calculated as the angle between two vectors, which are composed using the pixel values of the compared multispectral images); Structural SIMilarity SSIM[16] or extended SSIM - Q4,[17] (correlation, contrast, and luminance similarity between two images are used to calculate one similarity value); ERGAS[18] (similarity measure for multispectral images, based on the mean squared error estimator); Zero mean normalized cross-correlation, ZNCC or usually named as CORR.[19]

3.2 Spatial consistency

Spatial consistency is another aspect of fused imagery assessment. Up to now not many papers deal with spatial consistency assessment. Almost all the methods use a single scale edge detector (Gradient, Laplacian, Sobel edge detector) and an evaluation metric to calculate the distance between the edge maps (usually correlation coefficient).[20–22] For example, the High Pass Correlation Coefficient (HPCC) employs Laplacian and normalized correlation. Here the comparison is made between the fused bands and the corresponding panchromatic image. Another approach calculates the percentage of true and false edges introduced into the fused band using the Sobel edge detector.[22] Several works on fusion use the SSIM and ERGAS measures for spatial consistency assessment[1, 23] (panchromatic image is used as the reference instead of a spectral band, the measures are labeled as as SSIM_PAN and ERGAS_PAN).

An additional measure for spatial consistency assessment is recently proposed for use in.[24] This measure uses phase congruency (PC)[25] for feature extraction from pan-sharpened image. Invariance to intensity and contrast change as well as multiscale nature of this measure allows to obtain more confident assessment comparing to single-scale edge detectors.

4. ASSESSMENT OF MEASURES

4.1 General assumptions

An assessment measure (spectral or spatial consistency) should calculate a score according to the pan-sharpened image quality. For example, pan-sharpened images of different quality can be produced for the same scene. Similar assessment scores calculated by the measure for all these pan-sharpened images mean that the measure is not suitable and can provide distorted results. A measure providing good separation (in the meaning of the numerical score) of pan-sharpened imagery according to the quality is preferred.

The assessment measure should be sensitive to change (monotonous increase or decrease) of pan-sharpened image quality. The higher the quality, the higher the calculated score of the measure and vice versa. Increase (or decrease) of the image quality should lead to increase up to the ideal value (or decrease) of the assessment score.

A numerical measure can be assessed using test data, i.e. a test set of images. Variation of image quality in the test set allows to analyze the sensitivity of the measure using statistical methods. Pan-sharpened images produced with specified quality for one scene can be used as the test images to show which assessment measures are more sensitive to the quality change.

We have used the General Fusion Framework (GFF) pan-sharpening method proposed in[26] for the pan-sharpened image generation. The GFF method similarly as the General Image Fusion (GIF) method[27] shows that many pan-sharpening methods are quite similar and can be described as special cases of more general fusion

methods. GFF pan-sharpening method is selected for this study as it can precisely control the amount of high frequency data (extracted from the panchromatic image) added to the interpolated low resolution spectral image. Therefore the quality of pan-sharpened images can be controlled by varying the amount of added high frequency data. Variations of this amount allows to create a test set of images with specific quality.

4.2 Multispectral data

Medium and high resolution (Landsat 7 ETM+, IKONOS, and WorldView-2) spaceborne imagery is used for generation of the test sets (one test set is produced for each sensor). The images were obtained in different parts of the Earth and have different land cover classes, such as urban, rural, agricultural areas, forest and water regions to represent a high variety. Two scenes for each satellite are chosen. Landsat 7 ETM+ images (8-bit) were acquired at 7-th July 1999, and at 13-th September 1999 for the areas of San Jose city (USA) and Plattling town (Germany). IKONOS images (11-bit) were acquired at 15-th July 2005, 10:28 GMT, and at 24-th July 2004, 09:25 GMT for the areas of Munich city (Germany) and Athens city (Greece). WorldView-2 (11-bit) images were acquired at 12-th July 2010, 10:30 GMT, and at 10-th December 2009, 10:30 GMT for the areas of north of Munich city (Germany) and Rome city (Italy).

Ten nonoverlapping tiles are taken from the acquired images (multispectral and panchromatic) for each sensor. The size is 2000×2000 for panchromatic and 500×500 for multispectral IKONOS, WorldView-2, or 1000×1000 for multispectral Landsat 7 ETM+. Figures 1, 2, and 3 present the employed tiles taken from acquired Landsat-7 ETM+, IKONOS, and WorldView-2 images, respectively. The selected tiles contain desired varying land cover types depicting typical real world content of remotely sensed images.

4.3 Pan-sharpening method

A lot of existing multi-resolution methods can be expressed as an implementation of the General Fusion Framework (GFF) proposed in.[26] The GFF consists of the following steps:

1. Low resolution spectral image interpolation: $msi = I(ms)$,

2. Fusion: $msf = F(msi, pan)$,

3. Histogram matching: $msf = M(msf, ms)$,

where the ms is a low resolution spectral image and pan is the panchromatic (high resolution) image. The first and third steps can be included in the fusion step depending on the method. Usually, the I is a bilinear or cubic convolution interpolation and the F is a linear function of images (e.g. General Image Fusion method (GIF) in[27]).

In order to preserve spectral properties of a low resolution image ms one should add only high frequency information extracted from high resolution image pan. The general way is to perform such calculations in Fourier domain (signal processing view). First, both images are transformed into Fourier domain $MS = FFT(ms)$ and $PAN = FFT(pan)$. Then, high frequencies are extracted from the PAN and added to zero padded spectrum of the MS. The formula is written as

$$MSF = ZP(W \cdot MS) + PAN \cdot HPF, \tag{1}$$

where the ZP stands for zero padding, the W is the Hamming window (used to avoid aliasing and ringing), and the HPF is a high pass filter (e.g. Butterworth[19]). Cut-off frequency of the HPF filter (parameter hf) allows us to control the amount of high frequency data added to the low resolution image. We can rewrite formula (1) in signal domain as:

$$msf = msi + pan * hpf, \tag{2}$$

where $*$ stands for convolution and $hpf = FFT^{-1}(HPF)$. High frequency addition or high pass filtering method (e.g.[27]) is described by the same equation (2) as the GFF.
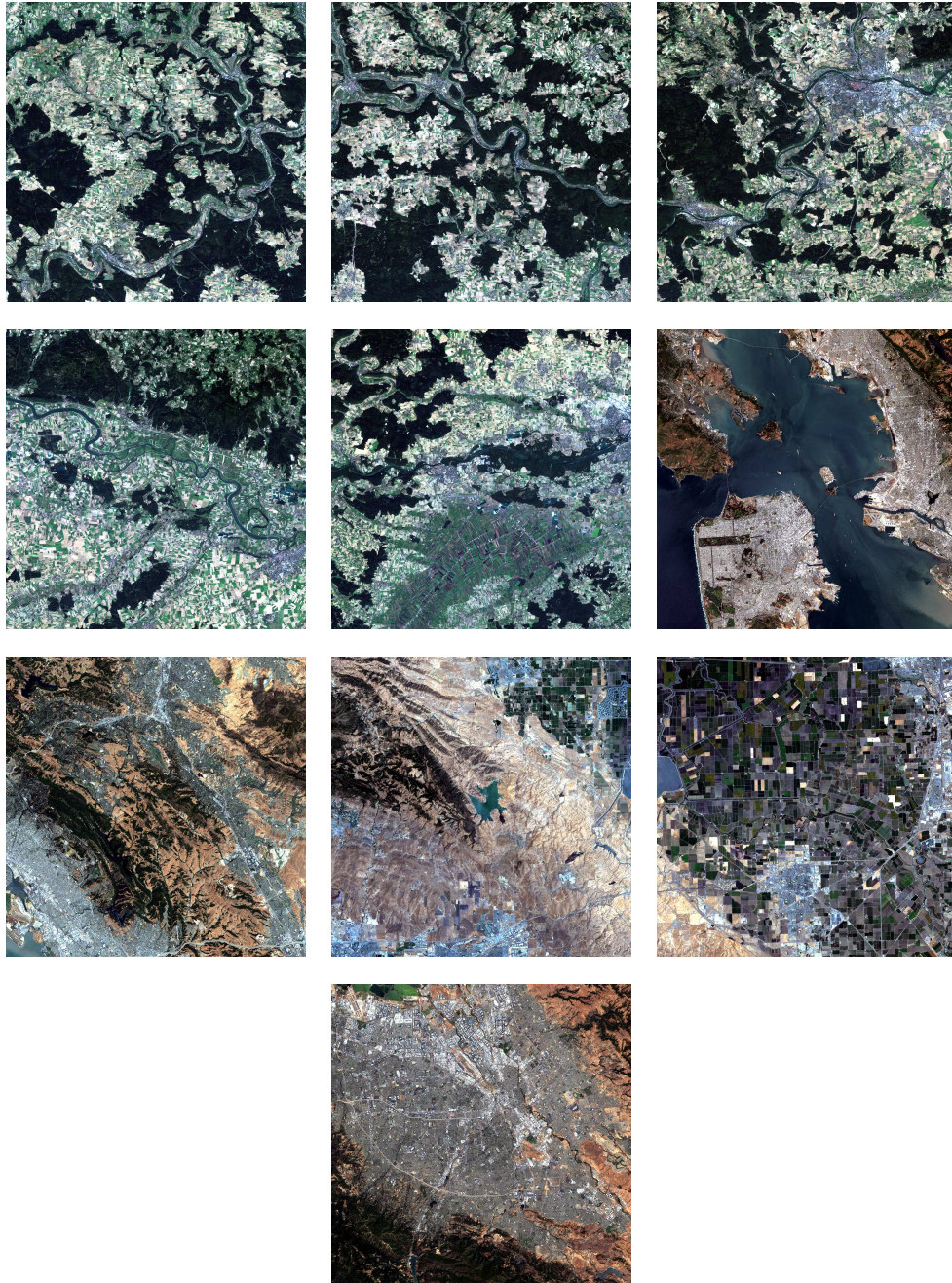
Figure 1. Ten non-overlapping tiles (visible range bands) taken from two Landsat-7 ETM+ scenes. The scenes are obtained in different parts of the Earth and present varying land cover classes.

Figure 2. Ten non-overlapping tiles (visible range bands) taken from two IKONOS scenes. The scenes are obtained in different parts of the Earth and present varying land cover classes.
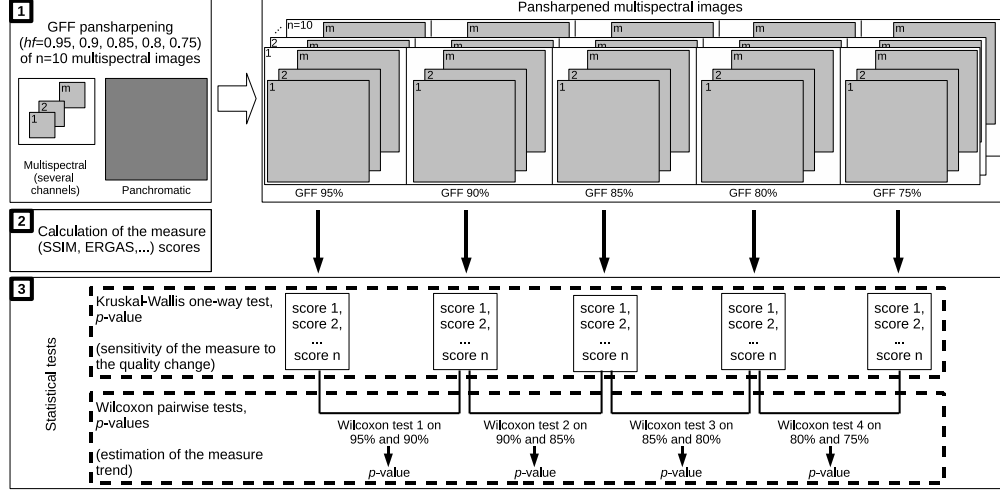
Figure 3. Ten non-overlapping tiles (visible range bands) taken from two WorldView-2 scenes. The scenes are obtained in different parts of the Earth and present varying land cover classes.

Figure 4. Diagram of statistical assessment of pan-sharpening assessment measures. First, the multispectral images ($n = 10$ multispectral images with m bands) are pan-sharpened by the GFF method five times with different parameter hf. Second, the numerical scores are calculated by the assessment measures (five groups of numerical scores, each group consists of $n = 10$ scores) are produced. Third, the statistical tests are performed: Kruskal-Wallis one-way test on the five groups of numerical scores, pairwise Wilcoxon test is performed on the pairs of the groups (95% and 90%, 90% and 85%, and so forth).

The GFF has three important advantages comparing to the methods run in signal domain. First, instead of interpolation of the msi, the GFF employs zero padding. Second, high pass filtering using box filters in signal domain make difficult precise design of a filter with required characteristics. Finally, a linear regression is used instead of histogram matching.

The GFF pan-sharpening method is used for the test data set generation due to its generality and high quality of produced fusion results. This method allows us to control the quality of produced pan-sharpened image by varying the parameter hf. The hf is in the range $[0, 1]$ and controls the proportionality (0%-100%) of high-frequency panchromatic image data to be added to low-resolution spectral image. The higher the value, the larger the high-pass filter width and more high-frequency data is added. Variation of this parameter allows to create fused images with desired quality: the higher the hf value, the more high-frequency data is added, and the higher spatial (lower spectral) consistency, and vice versa.

4.4 Test data generation

The nonoverlapping tiles are pan-sharpened by the GFF method with five values for the parameter hf (hf=0.95, 0.90, 0.85, 0.80, and 0.75, i.e. 95%, 90%, 85%, 80%, and 75% of high frequency panchromatic image data is added). Five groups of fused images (each group consists of ten pan-sharpened tiles) are generated for each test set (Figure 4, Step 1). To show that the GFF performs pan-sharpening with a competitive quality, the same image tiles are pan-sharpened by the ARSIS fusion method[28] (the ARSIS method is used for a comparison).

Figure 5 illustrates an example of WorldView-2 image pan-sharpening by the GFF method. In this figure a single band (green, 510-580 nm) is presented for visual comparison purpose. The GFF is run five times with different parameter value (hf=0.95, 0.90, 0.85, 0.80, and 0.75). The spatial consistency of the fused image decreases from the 95% of added high frequency data (Figure 5(a)) to 75% (Figure 5(e))), while the spectral consistency increases. Figure 5(f) contains the ARSIS fusion. The assessment scores (ERGAS, CORR, and HPCC) illustrate that the ARSIS fusion is comparable to the GFF fusion with 90% and 85% of added high frequency data (see Table 2). This shows an example that the GFF produces fusion with a comparable quality (spectral and spatial consistency).

(a) GFF 95%   (b) GFF 90%   (c) GFF 85%   (d) GFF 80%   (e) GFF 75%

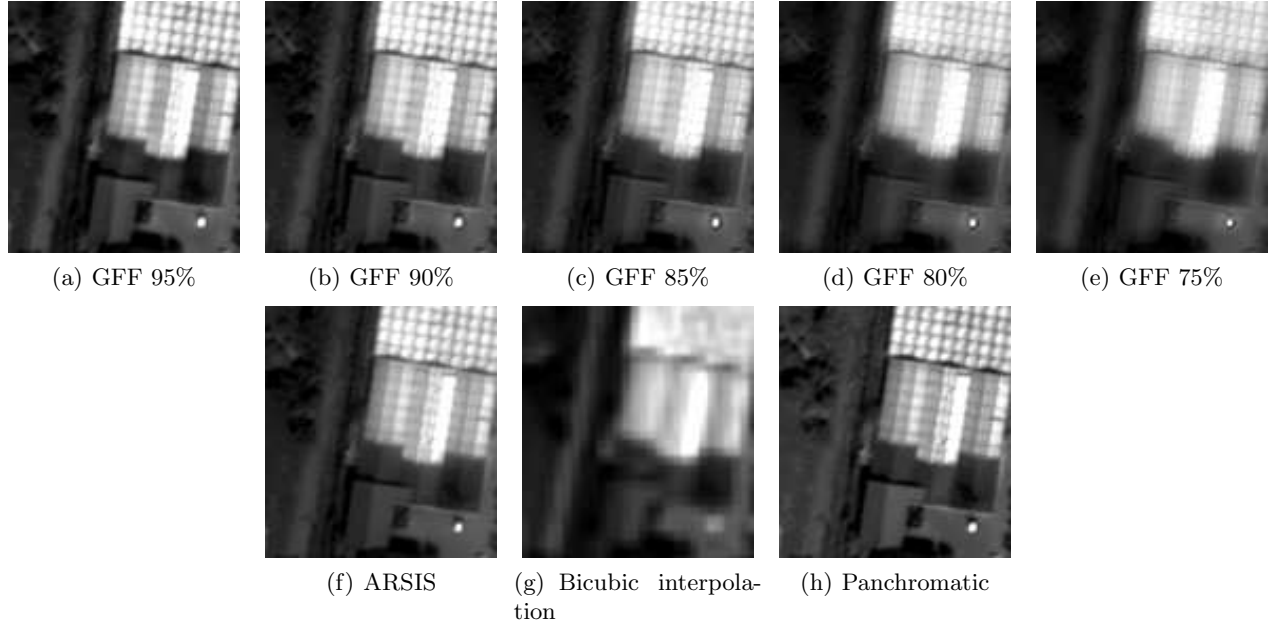(f) ARSIS   (g) Bicubic interpola-   (h) Panchromatic
tion

Figure 5. GFF pan-sharpening of WorldView-2 image is shown (green band is used for easier visual comparison of the fusion results). The GFF is run with varying hf parameter (hf=0.95, 0.90, 0.85, 0.80, and 0.75). The ERGAS, CORR (spectral consistency), and HPCC (spatial consistency) are given in Table 2 for comparison of the pan-sharpening quality with the ARSIS fusion. The quality of the ARSIS fusion (f) is comparable with the GFF 90% (b) (see Table 2).

Table 2. Erreur Relative Globale Adimensionalle de Synthèse (ERGAS), Zero mean normalized cross-correlation (CORR), and High Pass Correlation Coefficient (HPCC) measures calculated for the images shown in Figure 5. All channels of WorldView-2 are used

Figure	pan-sharpening method	ERGAS	CORR	HPCC
5(a)	GFF 95%	2.2535	0.9374	0.9530
5(b)	GFF 90%	2.0774	0.9493	0.9670
5(c)	GFF 85%	1.9103	0.9597	0.9567
5(d)	GFF 80%	1.7976	0.9664	0.9326
5(e)	GFF 75%	1.7309	0.9702	0.9043
5(f)	ARSIS	2.1083	0.9585	0.9900

4.5 Statistical assessment

A numerical measure (except for the SAM, ERGAS, and ERGAS_PAN) is calculated for every channel in the pan-sharpened multispectral images (6 channels in Landsat 7 ETM+, 4 in IKONOS, and 8 in WorldView-2). Measures on spectral consistency are performed using Wald's protocol.[18] To obtain one numeric score for a multispectral image the mean value is taken on the scores calculated for the channels. Five groups of assessment scores (also named as: 95%, 90%, 85%, 80%, and 75%) (Figure 4, Step 2) are calculated and used to assess the sensitivity of the measure to the quality change (the first test) and to estimate and analyze the trend of the measure (increase or decrease, e.g. the measure score change from the 95% to 90%, from the 90% to 85%, and so forth; the second test).

Sensitivity of a measure to the quality variation is assessed using the Kruskal-Wallis one-way analysis of variance[29] on the scores calculated for the 95%, 90%, 85%, 80%, and 75% pan-sharpened images (Figure 4, Step 3). The Kruskal-Wallis test is a non-parametric alternative to the one-way Analysis of variance (ANOVA) and used for testing equality of population medians among groups. The Kruskal-Wallis test does not assume a normal population and can be applied on samples with relatively small size, unlike the analogous one-way ANOVA.

Sensitivity of a measure to the monotonous change of the quality is assessed using pairwise comparisons (pairwise one tailed Wilcoxon rank sum test[30]). The test is performed on each of the following four pairs of the score groups: 1) 0.95% and 0.90%; 2) 0.90% and 0.85%; 3) 0.85% and 0.80%; 4) 0.80% and 0.75% (Figure 4, Step 3). In the case of a quality increase it is expected that the median of the 0.95% group is less than the

median of the 0.90% group, and so forth for the measure with minimal score equal to zero and the ideal value equal to some value (in most cases 1).

A post-hoc analysis (i.e. a special correction) should be carried out to determine if the statistical tests are significant. Such correction for tests with multiple comparisons is used when several statistical tests (dependent or independent) are being performed simultaneously (to reduce the Type I error rejecting the null hypothesis unappropriately). Bonferroni correction[31] is the most known and used way for the results adjustment. In order to get a higher accuracy, a novel method proposed by Benjamini and Yekutieli[32] is employed for the correction of Kruskal-Wallis and the pairwise tests.

In the ideal case the statistical tests should be carried out at some significance level (e.g. $\alpha = 0.05$). Significance level α specifies the probability level to accept that an event did not appear by chance. The lower the significance level, the stronger the evidence required. Usual values for significance level α are 0.05, 0.1, or 0.15. $\alpha = 0.05$ requires stronger evidence than $\alpha = 0.1$, $\alpha = 0.1$ requires stronger evidence than $\alpha = 0.15$, and so forth. The p-value less than 0.05 corresponds to a 5% chance of rejecting the null hypothesis when it is true (Type I error). A p-value less than predefined α level indicates that the null hypothesis should be rejected (i.e. there is a change of the calculated measure scores and the measure is sensitive to the quality variation or to the monotonous change). A p-value more than α level indicates that the null hypothesis should be accepted (i.e. there is no change of the calculated measure scores and the measure is not sensitive). A low p-value confirms higher significance of the test statistic and more support to reject the null hypothesis. Since it is not guaranteed that resultant p-values will be less than $\alpha = 0.05$ significance level, the other levels ($\alpha = 0.1$ or $\alpha = 0.15$) can be used.

5. RESULTS AND DISCUSSION

The spectral consistency of the produced groups of pan-sharpened images (95%, 90%, 85%, 80%, and 75%) is monotonously increasing, while the spatial consistency is monotonously decreasing (since less the high frequency data is added). Therefore the numerical scores of the measures SSIM, CORR (spectral consistency), and ERGAS_PAN (spatial consistency) are expected to increase, while the ERGAS and SAM scores are expected to decrease. The SSIM_PAN, CORR_PAN, HPCC, and PC (spatial consistency) scores are also expected to decrease since less high frequency data is added.

Tables 3-5 show results of the statistical tests for Landsat 7 ETM+, IKONOS, and WorldView-2, respectively. p-values of the Kruskal-Wallis test and pairwise tests are given for all the measures under comparison. Statistical tests illustrate that some measures cannot separate the pan-sharpened imagery according to the quality (e.g. high p-values of the tests for the HPCC for Landsat 7 ETM+ (Table 3), ERGAS_PAN for IKONOS (Table 4) and WorldView-2 (Table 5), the SAM for IKONOS (Table 4), and the CORR_PAN for WorldView-2 (Table 5)). Comparing and ranging the results of the tests by the p-values (in the ideal case a low value is expected), the measures can be ranged from the best one to the worst one. A low p-value of the Kruskal-Wallis test does not guarantee that the measure has required properties, since a pair of different (e.g. 95% and 75%) score sets can contribute to the low p-value. To select the best assessment measure a low p-value of the Kruskal-Wallis test together with low p-values of the pairwise tests are expected.

For Landsat 7 ETM+ pan-sharpening assessment (Table 3) the ERGAS, CORR (spectral consistency), and the PC, SSIM_PAN (spatial consistency) measures are preferable according to separability of the pan-sharpened imagery with different quality and estimated trends. These measures illustrated better estimation of the quality trend (less overlapped boxplots in Figures 6(b) (ERGAS score decrease), 6(d) (CORR score increase), 6(i) (PC score decrease), and 6(e) (SSIM_PAN score decrease)). A boxplot is a convenient way of a graphical description of groups of numerical data in the five-number summaries: 1. the sample minimum; 2. the first quartile (Q_1, lowest 25% of data, 25th percentile); 3. the median (Q_2, 50th percentile); 4. the third quartile (Q_3, highest 25% of data, 75th percentile); 5. the sample maximum. Observations considered as outliers are plotted as small circles.

For IKONOS (Table 4), the CORR (spectral consistency), and the PC (spatial consistency) measures are preferable. The measures show regular estimation of the quality trend in Figures 7(d) (CORR score increase) and 7(i) (PC score decrease).

For WorldView-2 (Table 5), the CORR (spectral consistency), and the SSIM_PAN, and PC (spatial consistency) measures are preferable (Figure 8) because of the good estimation of the regular trend in Figures 8(d) (CORR score increase), 8(e) (SSIM_PAN score decrease), and 8(i) (PC score decrease).

Comparing to other measures the SAM provides stable assessment scores on the imagery with varying spectral consistency. This measure is not sensitive to the quality change (high p-values in Tables 3-5, similar median values in the boxplots, Figures 6(c), 7(c), and 8(c)). This behaviour of the SAM measure can be explained as follows. During pan-sharpening, the values in the channels are modified by addition of one value (the high frequency data is added to the interpolated multispectral image, see the work[27]). Therefore, there is a change of the vector length (constructed from the pixel values of the pan-sharpened multispectral image) but there is no change of the angle between the compared vectors. The SAM allows robust assessment of the spectral consistency irrespectively to the change of spatial consistency.

The HPCC does not show high accuracy of spatial consistency assessment of the pan-sharpened imagery (the score trend is wrong during comparison of the 95% and 90% score groups, Figures 6(h), 7(h), and 8(h)). This may be explained that the edge detector in the HPCC measure can not extract enough spatial detail information. In comparison to the HPCC the PC measure provides more regular trends on the test sets of images. The multiscale nature of phase congruency as well as invariance to intensity and contrast changes allows more thorough analysis of fused data, comparing to single-scale edge detection methods.

The PC is run with default values of the parameter set on the test imagery of medium (Landsat 7 ETM+, 15m) and high (IKONOS, 1m and WorldView-2, 0.5m) resolution. Nevertheless, this measure provides required characteristics (image separability and expected trend of quality change) on the test data with different spatial resolution. These results again illustrate adaptability of the measure to spatial resolution allowing multiscale analysis of spatial consistency comparing to single-scale edge detection methods.

The boxplots illustrating an irregular trend (e.g. Figures 6(h), 7(a), 7(h), 8(a), and 8(h)) show that the measures do not estimate the quality change properly (wrong trend is estimated).

Table 6 contains the list of the measures more suitable for the Landsat 7 ETM+, IKONOS, and WorldView-2 pan-sharpened imagery assessment according to the obtained results.

Table 3. Kruskal-Wallis two-way test p-value and p-values of the pairwise tests (95% and 90%, 90% and 85%, 85% and 80%, and 80% and 75%) on Landsat 7 ETM+ pan-sharpened imagery. (*** - 0.05 significance level, ** - 0.10 significance level, * - 0.15 significance level)

Consistency	Measure	Kruskal-Wallis test, p-value	Pairwise tests, p-values			
			95% and 90%	90% and 85%	85% and 80%	80% and 75%
Spectral	SSIM*	0.1121	1	0.3440	0.3029	0.2565
	ERGAS†	0.1676	0.1801	0.1318*	0.1710	0.1710
	SAM‡	1	0.9417	0.9417	0.9417	1
	CORR§	0.0009***	0.1866	0.3029	0.3186	0.3186
Spatial	SSIM_PAN¶	0.8286	1	0.7688	0.6663	0.6663
	ERGAS_PAN‖	1	0.9233	0.9233	0.9233	0.9233
	CORR_PAN**	1	0.9233	0.9233	0.9233	0.9233
	HPCC††	1	1	1	1	0.2559
	PC‡‡	0.0914**	0.5630	0.5125	0.5123	0.3981

Table 4. Kruskal-Wallis two-way test p-value and p-values of the pairwise tests (95% and 90%, 90% and 85%, 85% and 80%, and 80% and 75%) on IKONOS pan-sharpened imagery. (*** - 0.05 significance level, ** - 0.10 significance level, * - 0.15 significance level)

Consistency	Measure	Kruskal-Wallis test, p-value	Pairwise tests, p-values			
			95% and 90%	90% and 85%	85% and 80%	80% and 75%
Spectral	SSIM*	0.8865	1	1	0.6967	0.5131
	ERGAS†	1	1	0.7092	0.7092	0.7092
	SAM‡	1	1	1	1	1
	CORR§	0.1465	1	0.6951	0.4730	0.4233
Spatial	SSIM_PAN¶	1	0.7744	0.2692	0.2692	0.2692
	ERGAS_PAN‖	1	1	1	1	1
	CORR_PAN**	1	0.7745	0.7745	0.7745	0.7745
	HPCC††	0.2367	1	1	0.3542	0.0717**
	PC‡‡	0.0009***	0.3186	0.2620	0.2620	0.3096

Table 5. Kruskal-Wallis two-way test p-value and p-values of the pairwise tests (95% and 90%, 90% and 85%, 85% and 80%, and 80% and 75%) on WorldView-2 pan-sharpened imagery. (*** - 0.05 significance level, ** - 0.10 significance level, * - 0.15 significance level)

Consistency	Measure	Kruskal-Wallis test, p-value	Pairwise tests, p-values			
			95% and 90%	90% and 85%	85% and 80%	80% and 75%
Spectral	SSIM*	0.0015***	1	0.1624*	0.1624*	0.1624*
	ERGAS†	1	0.7618	0.7035	0.7035	0.7618
	SAM‡	1	0.5622	0.5622	0.6917	0.6817
	CORR§	0.003***	0.1968	0.1968	0.1968	0.2052
Spatial	SSIM_PAN¶	0***	0.0549***	0.0005***	0.0005***	0.0070***
	ERGAS_PAN‖	1	0.9919	0.9919	0.9919	0.9919
	CORR_PAN**	1	0.8646	0.8646	0.9122	0.9122
	HPCC††	0***	1	0.0040***	0.0005***	0.0005***
	PC‡‡	0.0570***	0.3257	0.3257	0.3257	0.3257

Table 6. Recommended measures for pan-sharpened imagery assessment (spectral and spatial consistency)

Sensor	Bandwidths of spectral channels, μm	Spectral consistency	Spatial resolution: spectral and panchromatic, m	Spatial consistency
Landsat 7 ETM+	1: 0.450-0.515 2: 0.525-0.605 3: 0.630-0.690 4: 0.750-0.900 5: 1.550-1.750 6: 2.080-2.350	1) ERGAS, 2) CORR	30 and 15	1) PC, 2) SSIM_PAN
IKONOS	1: 0.445-0.516 2: 0.506-0.595 3: 0.632-0.698 4: 0.757-0.853	1) CORR	4 and 1	1) PC
WorldView-2	1: 0.400-0.450 2: 0.450-0.510 3: 0.510-0.580 4: 0.585-0.625 5: 0.630-0.690 6: 0.705-0.745 7: 0.770-0.895 8: 0.860-1.040	1) CORR	2 and 0.5	1) SSIM_PAN, 2) PC

The same statistical assessment experiment is carried out on the pan-sharpened data produced by Ehlers fusion[33] to illustrate the applicability of the selected measures for the data produced by other known methods. Ehlers method (in comparison to IHS, PCA, Brovey fusion) allows to control the amount of high frequency information extracted from a panchromatic image injected into an interpolated spectral image. In this experiment WorldView-2 data is pan-sharpened and the recommended measures (spectral consistency: CORR, spatial consistency: SSIM_PAN, PC) are employed. The assessment results (see Table 7 and Figure 9) illustrate an adequate trend (Figure 9(d) for the CORR, Figures 9(e) and 9(i) for the SSIM_PAN and PC). Correct trends on the measures assessment results illustrate that the selected measures can be applied for the evaluation of the pan-sharpened data produced by the other well known and representative pan-sharpening methods.

6. PAN-SHARPENED DATA ASSESSMENT USING SELECTED MEASURES

In this section the selected measures are employed for an assessment of pan-sharpened images produced by widely employed methods. The following representative methods are selected: Brovey transform, Intensity-Hue-Saturation fusion (IHS), Principal Component Analysis based fusion (PCA), Gram-Schmidt fusion (GS), and

---

*Structural SIMilarity (correlation, contrast, and luminance similarity between two images are used to calculate one similarity value)[16]

†Similarity measure for multispectral images, based on the mean squared error estimator (ERGAS)[18]

‡Spectral Angle Mapper (calculated as the angle between two vectors, which are composed using the pixel values of the compared multispectral images)

§Zero mean normalized cross-correlation[19]

¶Structural SIMilarity (SSIM) measure for spatial consistency assessment (panchromatic image is used as the reference)

‖ERGAS measure for spatial consistency assessment (panchromatic image is used as the reference)

**Zero mean normalized cross-correlation[19] (panchromatic image is used as the reference)

††High Pass Correlation Coefficient (employs Laplacian and normalized correlation)
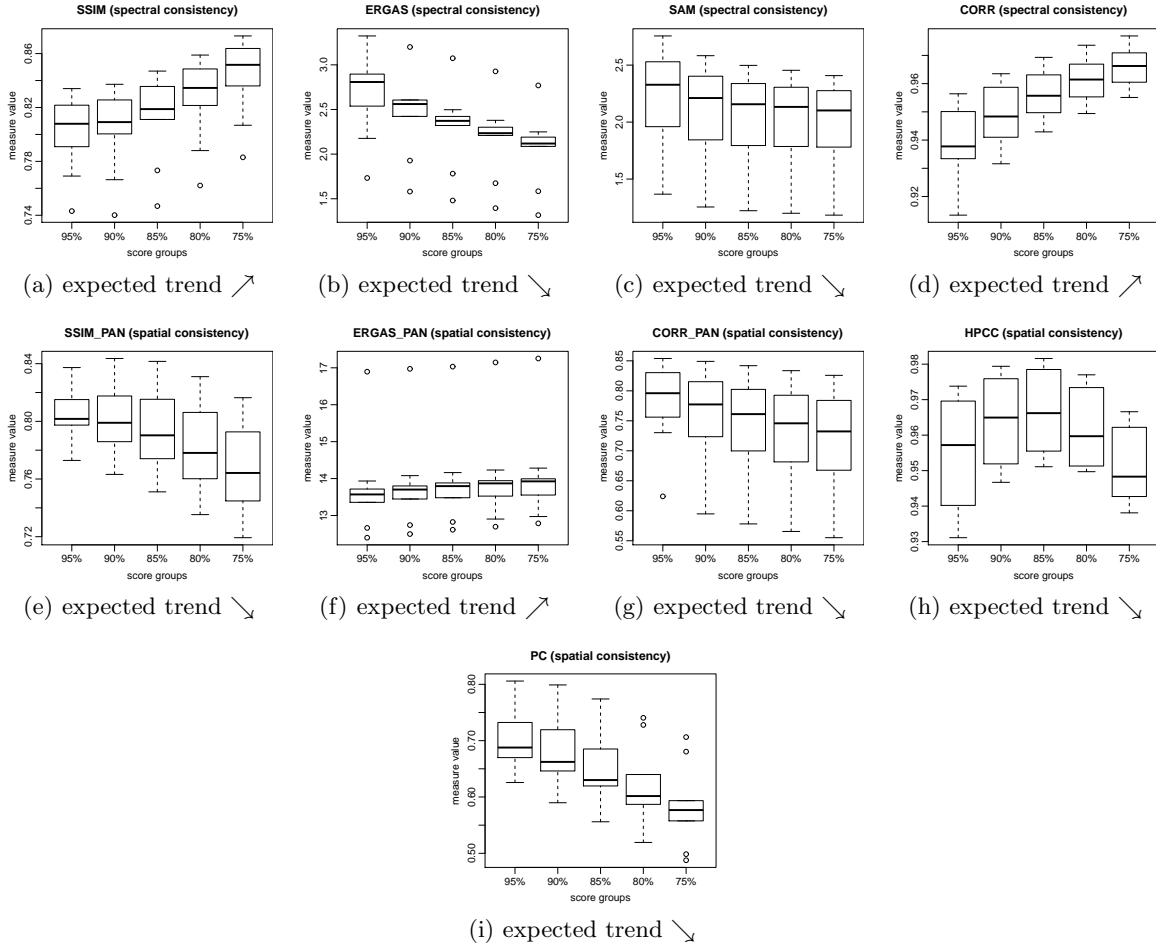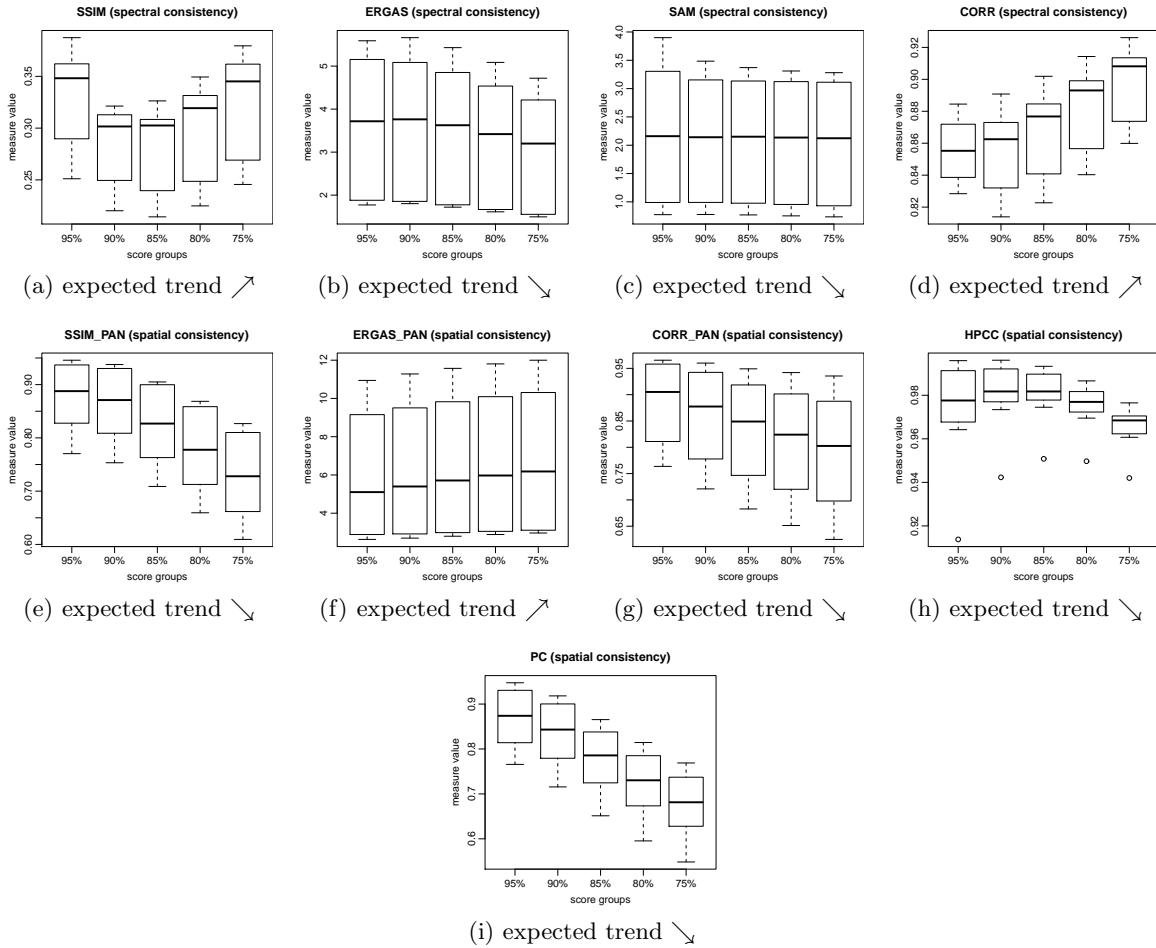
‡‡Phase congruency[25]

Figure 6. Boxplots of the measures scores carried out on pan-sharpened Landsat 7 ETM+ imagery (95%, 90%, 85%, 80%, and 75%): (a) SSIM, (b) ERGAS, (c) SAM, (d) CORR, (e) SSIM_PAN, (f) ERGAS_PAN, (g) CORR_PAN, (h) HPCC, (i) PC.

Table 7. Kruskal-Wallis two-way test p-value and p-values of the pairwise tests (95% and 90%, 90% and 85%, 85% and 80%, and 80% and 75%) on IKONOS imagery pan-sharpened by Ehlers method. (*** - 0.05 significance level, ** - 0.10 significance level, * - 0.15 significance level)

Consistency	Measure	Kruskal-Wallis test, p-value	Pairwise tests, p-values			
			95% and 90%	90% and 85%	85% and 80%	80% and 75%
Spectral	SSIM*	0***	0.006***	0.0125***	0.0145***	0.215
	ERGAS†	0.1974	0.7048	0.7048	0.7048	0.7048
	SAM‡	0.3993	0.9417	0.9417	0.9417	0.9417
	CORR§	0***	0.0486***	0.1153*	0.171	0.2787
Spatial	SSIM_PAN¶	0.0011***	0.2423	0.2423	0.2423	0.2423
	ERGAS_PAN$^{\parallel}$	0.9838	1	1	1	1
	CORR_PAN**	0.1299*	0.6830	0.7195	0.7744	0.7744
	HPCC††	0.0010***	0.3186	0.3186	0.3186	0.3186
	PC‡‡	0***	0.0240***	0.0527**	0.0577**	0.0767**

Figure 7. Boxplots of the measures scores carried out on pan-sharpened IKONOS imagery (95%, 90%, 85%, 80%, and 75%): (a) SSIM, (b) ERGAS, (c) SAM, (d) CORR, (e) SSIM_PAN, (f) ERGAS_PAN, (g) CORR_PAN, (h) HPCC, (i) PC.
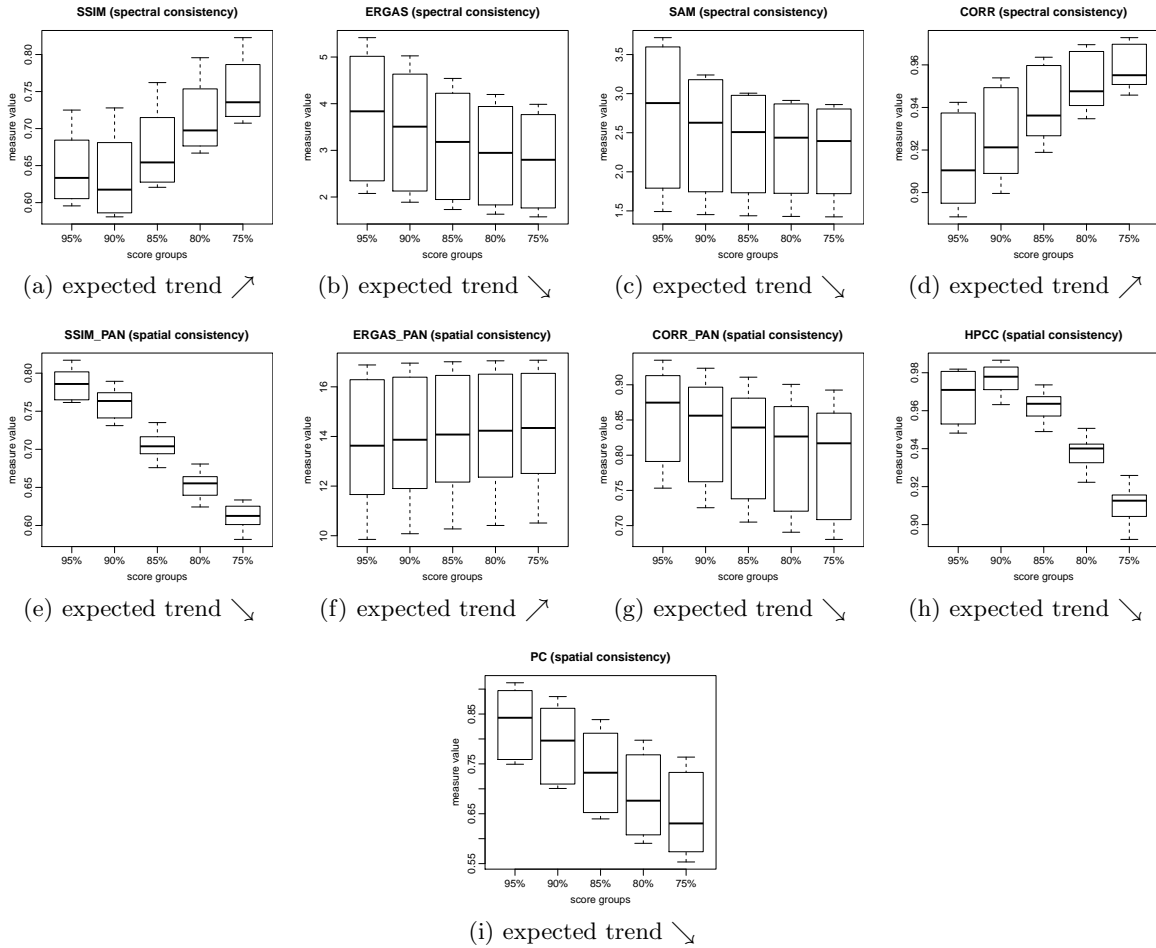
Figure 8. Boxplots of the measures scores carried out on pan-sharpened WorldView-2 imagery (95%, 90%, 85%, 80%, and 75%): (a) SSIM, (b) ERGAS, (c) SAM, (d) CORR, (e) SSIM_PAN, (f) ERGAS_PAN, (g) CORR_PAN, (h) HPCC, (i) PC.
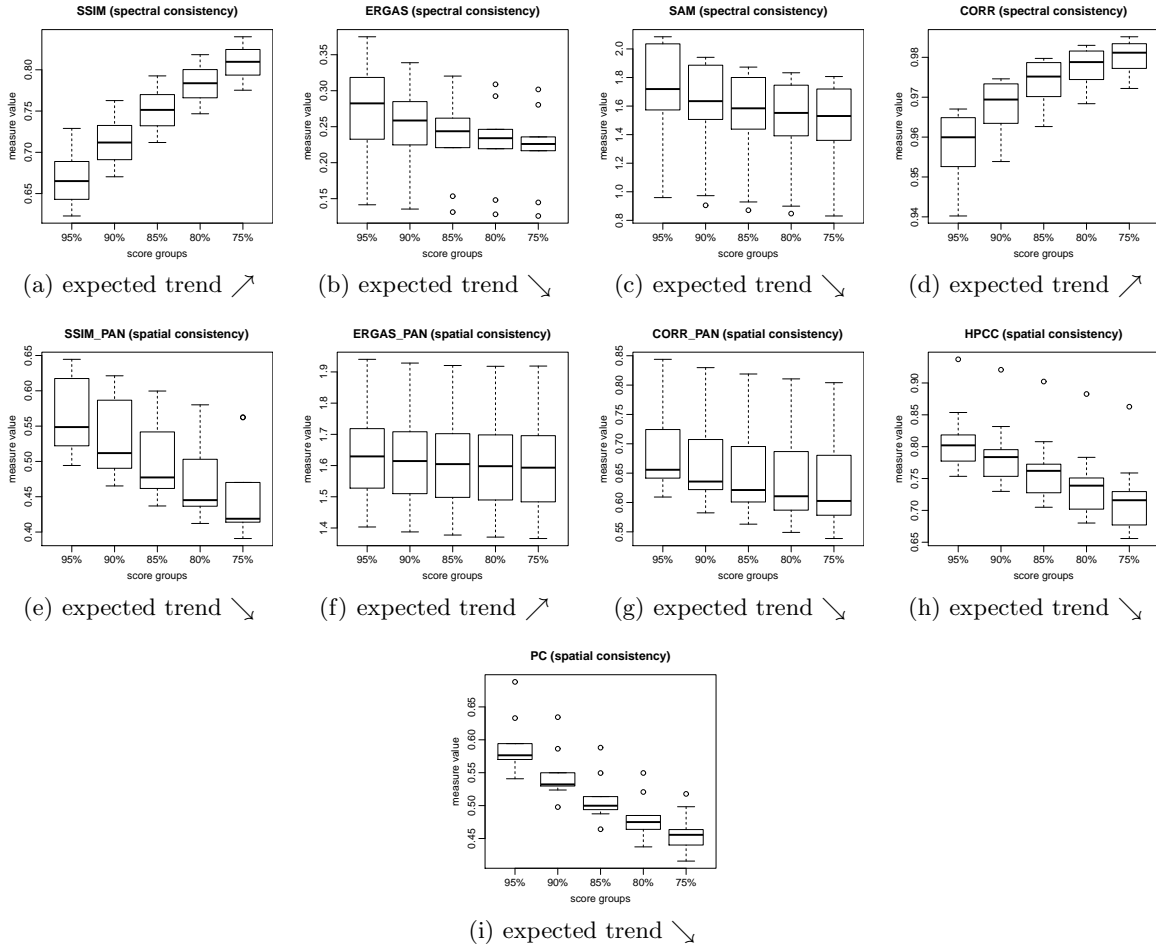
Figure 9. Boxplots of the measures scores carried out on WV2 imagery pan-sharpened by Ehlers fusion (95%, 90%, 85%, 80%, and 75%): (a) SSIM, (b) ERGAS, (c) SAM, (d) CORR, (e) SSIM_PAN, (f) ERGAS_PAN, (g) CORR_PAN, (h) HPCC, (i) PC.

ARSIS fusion (from French name "Amèlioration de la Rèsolution Spatiale par Injection de Structures"). The employed methods are thoroughly investigated in the literature and possess known properties and produce pan-sharpened imagery with an expected quality (spectral and spatial consistency). Knowing the properties of the employed method (e.g. IHS fusion produces a high spatial consistency with a loss of spectral properties or ARSIS fusion allows to obtain a high spectral and spatial consistency), a measure assessment result obtained for the methods can be compared and a conclusion on the measure sensitivity can be made.

In this experiment only the bands in RGB range are used since Brovey and the IHS fusion methods can be run only for RGB data. Landsat 7 ETM+, IKONOS, and WorldView-2 data is used. Calculating the CORR, PC, and SSIM_PAN measures for separate bands a mean value is calculated and used for comparison. Table 8 presents the numerical assessment results for the selected measures on the pan-sharpened images produced by the representative methods. The highest spectral accuracy assessment result has been given to ARSIS fusion by ERGAS and CORR, while these measures result in low spectral consistency for Brovey fusion, IHS, PCA, and GS. The PC and SSIM_PAN are showing highest spatial consistencies in comparison to Brovey fusion, IHS, PCA and GS, while ARSIS receives a lower spatial consistency assessment result. The experimental results agree with the expectations that the selected measures calculate adequate assessment results. It can be concluded that the selected measures in most cases represent the properties of the pan-sharpened imagery (with some small variations).

Table 8. Numerical results of the selected measures on pan-sharpened images produced by representative and widely employed methods. RGB bands are employed for the measures calculations (mean values over RGB bands are calculated for CORR, PC, and SSIM_PAN measures)

		Measure result on fusion method						
Sensor	Measure (consistency)	Brovey	IHS	PCA	GS	ARSIS	GFF	EHLERS
Landsat 7 ETM+	ERGAS (spectral)	18.7656	8.7119	6.4571	6.4312	2.5552	0.1289	0.1727
	CORR (spectral)	0.5502	0.6453	0.5573	0.5599	0.9261	0.9202	0.8619
	PC (spatial)	0.8991	0.8529	0.9529	0.9571	0.6054	0.5807	0.5904
	SSIM_PAN (spatial)	0.5037	0.9183	0.9324	0.9313	0.8489	0.8478	0.8348
IKONOS	CORR (spectral)	0.5628	0.6411	0.4748	0.5931	0.8897	0.9078	0.9102
	PC (spatial)	0.9617	0.9408	0.9713	0.6149	0.8603	0.7232	0.6072
WorldView-2	CORR (spectral)	0.8574	0.8953	0.9559	0.8805	0.9397	0.9539	0.9687
	SSIM_PAN (spatial)	0.6321	0.9544	0.9215	0.9238	0.8705	0.7917	0.7854
	PC (spatial)	0.9586	0.9081	0.9787	0.9812	0.9242	0.8297	0.8158

7. CONCLUSIONS

Statistical analysis reveals that not all the widely employed measures calculate accurate regular results of pan-sharpened imagery assessment and distortion of calculated scores can appear. A proper assessment of pan-sharpened results requires selection of appropriate assessment measures and the measures are important for both spectral and spatial consistency. Use of the selected assessment measures is expected to decrease errors during pan-sharpened imagery assessment.

The following recommendations can be given according to the experimental results. Spectral consistency assessment should employ ERGAS and Normalized Correlation measures (ERGAS and Normalized Correlation for Landsat 7 ETM+, Normalized Correlation for IKONOS and WorldView-2). Spatial consistency assessment should employ Phase Congruency and SSIM measures (Phase Congruency and SSIM for Landsat 7 ETM+ and WorldView-2, Phase Congruency for IKONOS). SAM provides stable assessment scores of the imagery irrespectively to varying spectral and spatial consistency, therefore it should be used with caution. The SSIM measure illustrates suitability for spatial consistency assessment, but unfortunately this measure is not widely used. Phase Congruency shows a good separability of pan-sharpened imagery in the sense of spatial consistency, and sensitivity to the trend of the quality change (for all used sensors data). Therefore it can be recommended for spatial consistency assessment.

Assessment of pan-sharpened imagery acquired by other satellites can be performed by the measures recommended for the medium, or high resolution imagery, respectively.

ACKNOWLEDGMENTS

REFERENCES

[1] González-Audícana, M., Otazu, X., Fors, O., and Seco, A., "Comparison between Mallat's and the á trous discrete wavelet transform based algorithms for the fusion of multispectral and panchromatic images," *International Journal of Remote Sensing* **26**(3), 595–614 (2005).

[2] Klonus, S. and Ehlers, M., "Performance of evaluation methods in image fusion," in [*12th International Conference on Information Fusion*], 1409–1416 (6-9 July 2009).

[3] Avcibas, I., Sankur, B., and Sayood, K., "Statistical evaluation of image quality measures," *Journal of Electronic Imaging* **11**(2), 206–223 (2002).

[4] Zhang, Y., [*Advances in Photogrammetry, Remote Sensing and Spatial Information Sciences: 2008 ISPRS Congress Book*], ch. Chapter 14: Pan-Sharpening for Improved Information Extraction, 185–202, Taylor Francis Group, London (2008).

[5] Colditz, R., Wehrmann, T., Bachmann, M., Steinnocher, K., Schmidt, M., Strunz, G., and Dech, S., "Influence of image fusion approaches on classification accuracy: A case study," *International Journal of Remote Sensing* **28**(27), 3311–3335 (2006).

[6] Shah, C. and Quackenbush, L., "Analyzing multisensor data fusion techniques: a multi-temporal change detection approach," in [*ASPRS 2007 Annual Conference*], (7-11 May 2007).

[7] Ehlers, M., "Multisensor image fusion techniques in remote sensing," *ISPRS Journal of Photogrammetry and Remote Sensing* **46**(1), 19–30 (1991).

[8] Jahjah, M. and Ulivieri, C., "Automatic archaeological feature extraction from satellite vhr images," *Acta Astronautica* **66**(9-10), 1302–1310 (2010).

[9] Boloorani, A., Erasmi, S., and Kappas, M., "Urban land cover mapping using object/pixel-based data fusion and ikonos images," in [*ISPRS VII Midterm Symposium*], (8-11 May 2006).

[10] Bachmann, M. and Habermeyer, M., "Evaluation of image fusion techniques for large-scale mapping of non-green vegetation," in [*3rd EARSeL Workshop on Imaging Spectroscopy*], (13-16 May 2003).

[11] Nichol, J. and Wong, M., "Satellite remote sensing for detailed landslide inventories using change detection and image fusion," *International Journal of Remote Sensing* **26**(9), 1913–1926 (2005).

[12] Zhang, Y., Wang, R., and Munroe, K., "Multi-resolution and multi-spectral image fusion for urban object extraction," in [*XXth ISPRS Congress, Commission III*], 960–966 (1224 July 2004 2004).

[13] Forsythe, K., "Monitoring rapid urban expansion: A case study of calgary, alberta, canada," in [*3rd International Symposium Remote Sensing and Data Fusion Over Urban Areas (URBAN 2005)*], (14-16 March 2005).

[14] Maxwell, T. and Zhang, Y., "A fuzzy logic approach to optimization of segmentation of object-oriented classification," in [*SPIE 50th Annual Meeting Optics Photonics*], (31 July-4 August 2005).

[15] Ozdarici, A. and Turker, M., "Field-based classification of agricultural crops using multi-scale images," in [*First International Conference on Object-based Image Analysis (OBIA 2006)*], (4-5 July 2006).

[16] Wang, Z., Bovik, A. C., and Sheikh, H. R., "Image quality assessment: From error visibility to structural similarity," *IEEE Transaction on Image Processing* **13**(4), 600–612 (2004).

[17] Alparone, L., Baronti, S., Garzelli, A., and Nencini, F., "A global quality measurement of pan-sharpened multispectral imagery," *IEEE Geoscience and Remote Sensing Letters* **1**(4), 313–317 (2003).

[18] Wald, L., Ranchin, T., and Mangolini, M., "Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images," *Photogrammetric Engineering & Remote Sensing* **63**(6), 691–699 (1997).

[19] Gonzales, R. and Woods, R., [*Digital Image Processing, 3rd edition*], Prentice Hall, New Jersey (2008).

[20] Shi, W., Zhu, C., Zhu, C., and Yang, X., "Multi-band wavelet for fusing SPOT panchromatic and multi-spectral images," *Photogrammetric Engineering & Remote Sensing* **69**(5), 513–520 (2003).

[21] Zhou, J., Civco, D., and Silander, J., "A wavelet transform method to merge Landsat TM and SPOT panchromatic data," *International Journal of Remote Sensing* **19**(4), 743–757 (1998).

[22] Pradhan, P., King, R., Younan, N., and Holcomb, D., "Estimation of the number of decomposition levels for a wavelet-based multiresolution multisensor image fusion," *IEEE Transactions on Geoscience and Remote Sensing* **44**(12), 3674–3686 (2006).

[23] Lillo-Saavedra, M., Gonzalo, C., Arquero, C., and Martinez, E., "Fusion of multispectral and panchromatic satellite sensor imagery based on tailored filtering in the Fourier domain," *International Journal of Remote Sensing* **26**(6), 1263–1268 (2005).

[24] Makarau, A., Palubinskas, G., and Reinartz, P., "Multiresolution image fusion: phase congruency for spatial consistency assessment," in [*ISPRS Technical Commision VII Symposium - 100 Years ISPRS*], Wagner, W. and Szekely, B., eds., **XXXVIII**, 226–239 (5-7 July 2010).

[25] Kovesi, P., "Image features from phase congruency," *Videre: A Journal of Computer Vision Research* **1**(3), 2–26 (1999).

[26] Palubinskas, G. and Reinartz, P., "Multi-resolution, multi-sensor image fusion: general fusion framework," in [*Joint Urban Remote Sensing Event (JURSE), 2011*], 313–316 (april 2011).

[27] Wang, Z., Ziou, D., Armenakis, C., Li, D., and Li, Q., "A comparative analysis of image fusion methods," *IEEE Transactions on Geoscience and Remote Sensing* **43**(6), 1391–1402 (2005).

[28] Ranchin, T. and Wald, L., "Fusion of high spatial and spectral resolution images: The ARSIS concept and its implementation," *Photogrammetric Engineering & Remote Sensing* **66**(1), 49–61 (2000).

[29] Sheskin, D. J., [*Handbook of Parametric and Nonparametric Statistical Procedures, Second Edition*], Chapman & Hall/CRC (2000).

[30] Mann, H. B. and Whitney, D. R., "On a test of whether one of two random variables is stochastically larger than the other," *The Annals of Mathematical Statistics* **18**(1), 50–60 (1947).

[31] Abdi, H., "Bonferroni and Sidak corrections for multiple comparisons," in [*In Encyclopedia of Measurement and Statistics. Edited by: Salkind NJ. Thousand Oaks, CA: Sage*], (2007).

[32] Benjamini, Y. and Yekutieli, D., "The control of the false discovery rate in multiple testing under dependency," *Annals of Statistics* **29**(4), 1165–1188 (2001).

[33] Ehlers, M., "Spectral characteristics preserving image fusion based on fourier domain filtering," in [*Proceedings of SPIE - Remote Sensing for Environmental Monitoring, GIS Applications, and Geology IV*], M. Ehlers, H. K. and (Eds.), U. M., eds., 1–13 (2004).

Aliaksei Makarau received the Diploma (Dipl.-Ing.) in computer science in 2003 from Belarusian State University of Informatics and Radioelectronics and Ph.D. in technical sciences from the United Institute of Informatics Problems, Minsk, Belarus in 2008. His dissertation was on fast methods for multispectral image fusion and processing. He is a postdoctoral fellow at the department "Photogrammetry and Image Analysis" at the German Aerospace Centre (DLR), Remote Sensing Technology Institute (IMF). His research interests include multimodal data fusion, pattern recognition, and classification.

Gintautas Palubinskas received the M.S. and Ph.D degrees in mathematics from Vilnius University, Lithuania, in 1981 and Institute of Mathematics and Informatics (IMI), Vilnius, in 1991. His doctoral

dissertation was on spatial image recognition.

He was a Research Scientist at the IMI from 1981 to 1997. Since 1997, he has been a Research Scientist at Remote Sensing Technology Institute, German Aerospace Center (DLR), Oberpfaffenhofen, Germany. From 1993 to 1997, he was a Visiting Research Scientist at German Remote Sensing Data Center, DLR; the Department of Geography, Swansea University, Wales, U.K.; Institute of Navigation, Stuttgart University, Germany; Max-Planck-Institute of Cognitive Neuroscience, Leipzig, Germany. His current interests are in image processing, classification and change detection, data fusion for optical and SAR remote sensing applications.

Peter Reinartz received the Diploma (Dipl.-Phys.) in theoretical physics in 1983 from the University of Munich and his Ph.D. (Dr.-Ing) in civil engineering from the University of Hannover, in 1989. His dissertation was on statistical optimization of classification methods for multispectral image data. He is head of the department "Photogrammetry and Image Analysis" at the German Aerospace Centre (DLR), Remote Sensing Technology Institute (IMF) and holds a professorship on Geoinformatics at the University of Osnabrueck. He has more then 20 years of experience in image processing and remote sensing and over 150 publications in these fields. His main interests are in direct georeferencing, stereo photogrammetry and data fusion, generation of digital elevation models and interpretation of VHR image data from sensors like WorldView, GeoEye a.o. He is also engaged in using remote sensing data for disaster management and using high frequency time series of airborne image data for real time application in case of disasters as well as for traffic monitoring.

List of Figures