Forest/Non-Forest Classification from TanDEM-X Interferometric Data by means of Multiple Fuzzy Clustering

Michele Martone, Paola Rizzoli, Benjamin Bräutigam, Gerhard Krieger
Microwaves and Radar Institute, German Aerospace Center, Michele.Martone@dlr.de, Germany

Abstract

In this paper we introduce a method to derive forest/non-forest maps from TanDEM-X interferometric synthetic aperture radar (InSAR) data, acquired at global scale in stripmap single polarization (HH) mode. Among the several observables systematically provided by the TanDEM-X system, volume decorrelation, derived from the interferometric coherence, shows to be consistently sensitive to the particular land cover type, and is therefore used as an input data set for applying a classification method based on a fuzzy clustering algorithm. Since the considered InSAR quantity strongly depends on the geometric acquisition configuration, namely the incidence angle and the interferometric baseline, a multi-clustering classification approach is used. Algorithms for the identification of additional information layers such as urban and water areas are discussed as well, and the mosaicking of multiple acquisitions to improve the resulting accuracy is shortly introduced. The preliminary classification results shown in this paper are very promising for the generation of a global land classification map from TanDEM-X interferometric quicklook data as a next step.

1 Introduction

Remote sensing data represent a highly valuable source for land classification purposes. A precise and up-to-date knowledge of the land cover information is of great importance for a wide range of scientific and commercial purposes. In particular, the identification and the monitoring of vegetated areas is critical for a variety of applications, such as agriculture, cartography, geology, forestry, global change research, as well as for regional planning. High-resolution forest classification maps at a global scale have been produced in the last years [1], [2]. In this paper we present a method to generate forest classification maps from TanDEM-X interferometric SAR data. The TanDEM-X mission comprises the two twin satellites TerraSAR-X and TanDEM-X with the main goal of producing a global and consistent Digital Elevation Model (DEM) with an unprecedented accuracy, by exploiting single-pass SAR interferometry [3]. Hence, the present method has the potential to be applied to the global TanDEM-X data set. Several observables are systematically provided by the TanDEM-X system. Among them, the interferometric coherence is sensitive to the type of land cover under illumination, and may therefore be exploited for classification purposes. The interferometric coherence represents the correlation coefficient between master and slave acquisitions and is a key parameter for the evaluation of the InSAR performance. Several contributions may affect the quality of interferometric data. In particular, the coherence loss caused by volume scattering represents the contribution which is predominantly influenced by the presence of vegetation. Since the beginning of the TanDEM-X mission (end of 2010), about half a million of high-resolution bistatic single polarization (HH) scenes covering all the Earth's land masses have been acquired and processed. A single bistatic scene typically extends over an area of about 30 by 50 km in range and azimuth, respectively. From this, quicklook images representing several SAR and InSAR quantities (like backscatter and coherence maps, or the calibrated RawDEM) are generated at a ground pixel spacing of about 50 m $\times$ 50 m by applying a spatial averaging process to the corresponding operational TanDEM-X interferometric data at full resolution. Working with such reduced resolution data makes feasible the exploitation of the TanDEM-X dataset on a global scale, by keeping low the computational load. Global mosaics from TanDEM-X quicklook data have been produced in the last years [4], [5] and are of great interest for a variety of applications. In this paper, we extend the method presented in [6], where a preliminary algorithm for forest/non-forest classification from TanDEM-X interferometric data by means of the volume decorrelation contribution is presented, whose derivation is recalled in the next section. The proposed multi-clustering classification algorithm is based on fuzzy logic and described in Section 3. The output is a metric to determine a sort of probability of vegetation from TanDEM-X InSAR data. Exemplary results from typical forest types such as rainforest, temperate, and boreal forest are shown in Section 4, and the potentials for further possible applications (as e.g., forest change detection), are proposed as well. Here, particular focus is given on the identification of urban settlements, water, as well as areas affected by geometrical distortions, which affect the classification accuracy and need therefore to be identified. The paper is concluded in Section 5 with a summary and an outline of the next steps.

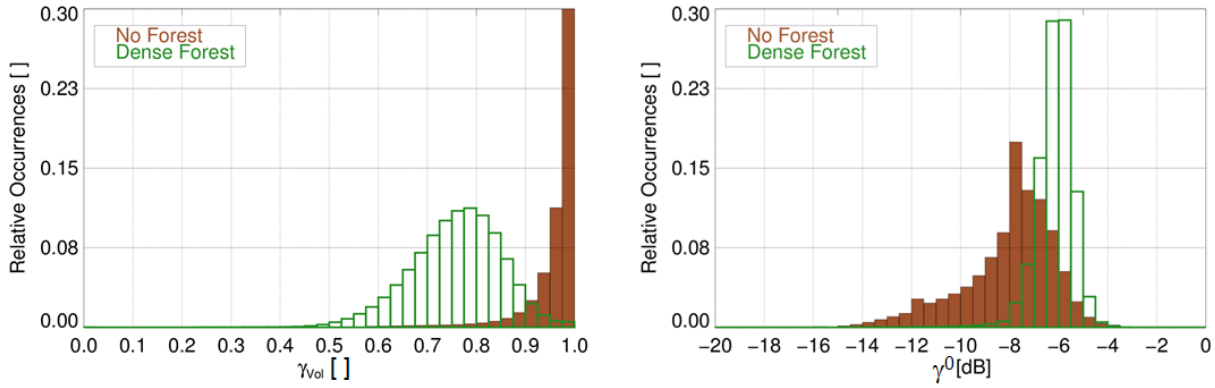

Figure 1: Occurrence distribution for the (left) volume decorrelation contribution estimated according to [6] and (right) the backscatter information for densely forested and non-vegetated areas, depicted in green and brown, respectively, estimated from TanDEM-X stripmap single-pol HH bistatic acquisitions over a large region located in the Amazon rainforest. For data consistency (i.e. similar decorrelation effects) heights of ambiguity between 40 and 50 meters are considered. For land cover identification, the 30-m resolution vegetation continuous fields tree cover data provided by the multispectral sensor Landsat-5 have been used.

2 Volume Decorrelation Derivation

Several contributions cause a coherence loss in TanDEM-X interferometric data [7], which, assuming statistical independence, can be factorized as follows:

$$\gamma = \gamma_{\mathrm{SNR}} \cdot \gamma_{\mathrm{Quant}} \cdot \gamma_{\mathrm{Amb}} \cdot \gamma_{\mathrm{Range}} \cdot \gamma_{\mathrm{Azimuth}} \cdot \gamma_{\mathrm{Temp}} \cdot \gamma_{\mathrm{Vol}}. \quad (1)$$

The terms on the right-hand side describe decorrelation due to: limited signal-to-noise ratio (γ_{SNR}), quantization errors (γ_{Quant}), ambiguities (γ_{Amb}), baseline decorrelation (γ_{Range}), errors due to relative shift of Doppler spectra ($\gamma_{\mathrm{Azimuth}}$), and temporal decorrelation (γ_{Temp}). The last term (γ_{Vol}) describes the coherence loss caused by volume scattering, and represents the contribution which is predominantly affected by the presence of vegetation. Given a coherence estimate γ, it is straightforward to quantify the volume decorrelation contribution as

$$\gamma_{\mathrm{Vol}} = \frac{\gamma}{\gamma_{\mathrm{SNR}} \cdot \gamma_{\mathrm{Quant}} \cdot \gamma_{\mathrm{Amb}} \cdot \gamma_{\mathrm{Range}} \cdot \gamma_{\mathrm{Azimuth}} \cdot \gamma_{\mathrm{Temp}}}. \quad (2)$$

The impact and the estimation procedure of each decorrelation contribution on TanDEM-X data is discussed in detail in [6]. For each quicklook product a local incidence angle map is derived from the orbit parameters and the calibrated RawDEM. This allows to improve the estimation accuracy and, in particular, to precisely compensate for possible geometrical decorrelation due to the presence of topography. The X-band radar signal backscattered by the Earth's surface is also correlated with the particular land cover type. However, due to the influence of soil moisture variation, roughness, and other dielectric and geometric characteristics, the discrimination of the particular land type may become difficult. For a first assessment of the potentials for forest classification of the volume decorrelation contribution γ_{Vol} and of the normalized radar backscatter $\gamma^0 = \sigma^0/\cos(\theta_{\mathrm{inc}})$ (being σ^0 the calibrated backscatter coefficient and θ_{inc} the local incidence angle) we have considered TanDEM-X bistatic acquisitions acquired over a large region located in the Amazon rainforest, Brazil. For the discrimination of the land cover, we have used the 30-m resolution global vegetation continuous fields tree cover data from the multispectral sensor Landsat-5 in 2005 and freely available online [1], which provide the percentage of area covered by vegetation. For this analysis, pixels having a value smaller than 15% and larger than 65% have been selected as representative for non-vegetated and forested areas, respectively. The corresponding occurrence distributions of backscatter and volume decorrelation estimated from TanDEM-X data are depicted in Fig. 1 for densely vegetated areas and for bare surfaces, which clearly shows the coherence information to be a more powerful indicator for classification purposes than the pure backscatter information, for which a higher confusion between the two land cover types is observed. Hence, for the classification method presented in this paper the coherence information only will be exploited. Once the volume contribution has been estimated from (2), it has to be properly associated to a forest/non-forest classification, as detailed in the next section.

3 Fuzzy Clustering for Forest Classification

In a general sense, clustering indicates the task of grouping a set of objects coming from N input observations $\mathbf{Y} = [\mathbf{y}_k]$ $(k = 1, \ldots, N)$, each one characterized by a set of P features, depending on how similar they are to each other. The observations are then divided into c non-empty subsets, called clusters. Fuzzy-clustering has been introduced in order to allow a certain amount of overlap between different clusters [8]. According to it, a so-called

membership function $\hat{\mathbf{U}}$ is defined, which describes the probability of an observation to belong to each cluster ($\hat{\mathbf{U}} = [\hat{u}_{ik}] \in [0,1]$, $i = 1, \ldots, c$). The results are fuzzy c-partitions of the input observation data set, which contain observations characterized by a high intracluster similarity and a low extracluster one. Fuzzy clustering represents a powerful and effective approach widely used in numerous contexts and applications (such as data mining or pattern recognition). Its potentials have been already shown for classifying the Greenland ice sheet snow facies by using TanDEM-X interferometric data [9]. As previously explained, for forest/non-forest classification from TanDEM-X data we exploit the volume decorrelation information only (i.e. $P = 1$). The number of clusters is set to two, to discriminate forest (F) from non-forest (NF) areas. The cluster centers $\{v_F, v_{NF}\}$ are identified by their own feature $\{\gamma_{\mathrm{Vol},F}, \gamma_{\mathrm{Vol},NF}\}$. However, one has to be aware that the intensity of the volume decorrelation contribution γ_{Vol} strongly depends on the specific acquisition geometry as well (beyond the specific type of forest, of course). In [6] it is verified that the coherence loss over forest is notably influenced by the specific incidence angle: a stronger decorrelation is expected for steeper incidence angles, since the radar microwaves are able to penetrate deeply through the canopy, "sensing" it predominantly like a volume, whereas for shallow incidence angles the surface scattering component becomes dominant, resulting in a higher coherence. When implementing a clustering algorithm for forest classification by using X-band interferometric data, such dependencies have to be taken into account. For this reason a partitioning of the original data into S subsets is performed, depending on the specific pair of baseline $B_\perp$ (i.e. height of ambiguity) and incident angle θ_{inc}. According to this, an observation (pixel) k is associated to the i-th subset if

$$B_{\perp k} \in \left[B_{\perp i,\min}, B_{\perp i,\max}\right], \quad (3)$$

$$\theta_{\mathrm{inc}k} \in \left[\theta_{\mathrm{inc}i,\min}, \theta_{\mathrm{inc}i,\max}\right]. \quad (4)$$

An important step of the present algorithm consists in the definition of the cluster centers, each one identified by a P-dimensional tie-point vector v_i (P being the number of features). For their determination TanDEM-X bistatic data takes acquired over a large region located in the Amazon rainforest are "trained" by using the forest density information provided by Landsat [1] (see also Fig. 1). For each subset i the sample expectation of γ_{Vol} is finally taken as cluster center. In Fig. 2 the markers indicate the cluster centers for forest and non-forest areas, for different heights of ambiguity and the local incidence angle range. Near, mid, and far range are identified with angles steeper than 35°, between 35° and 45°, larger than 45°, respectively. From each cluster, a minimum mean square error fitting is applied and a more continuous subcluster distribution is determined, which is indicated by the continuous lines. Then, for each subset a fuzzy clustering algorithm is run independently. This approach allows us to avoid the necessity of correcting the input data by means of a priori information (such as extinction models or assumptions for the type of forest under investigations [6]), which may end up in incorrect classification results. The output of the clustering is the so-called membership function, which is confined between 0 and 1 and describes the probability of the k-th observation to belong to the i-th cluster and is derived as for the fuzzy clustering algorithm [8]

$$\hat{u}_{ik} = \left(\sum_{j=1}^{c} \left(\frac{\|y_k - \hat{v}_i\|}{\|y_k - \hat{v}_j\|} \right)^{\frac{2}{m-1}} \right)^{-1}, 1 \leq k \leq N. \quad (5)$$

In this sense, this metric can be interpreted as a sort of probability of a pixel to be "covered" by vegetation. The next step will be to try to link this quantity to the local vegetation properties, such as its density, which represents an important parameter for several forestry applications such as e.g. biomass estimation.

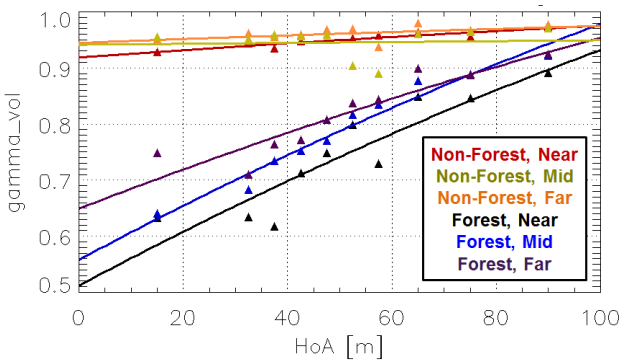

Figure 2: The markers indicate the cluster centers of volume decorrelation for a large region located in the Amazon rainforest (Brazil), and calculated according to (2). The clusters are derived for forest and non-forest areas, and for the corresponding height of ambiguity and the local incidence angle range. From this, a MMSE fitting is applied (curved lines) to get a more dense subcluster distribution.

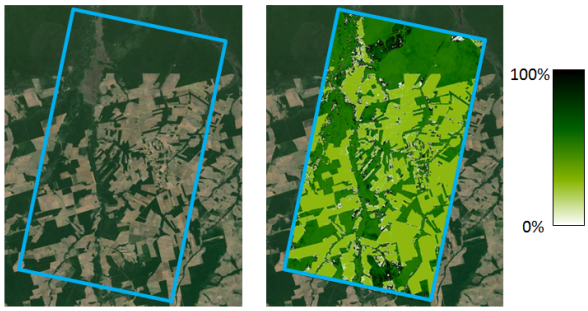

Figure 3: (Right) Forest classification map of an area located in the Amazon rainforest, Brazil, overlaid on a GoogleEarth optical image, which is given on the left-hand side for comparison. The considered area is delimited by the blue rectangles and extends by about 30 km in range and 55 km in azimuth.

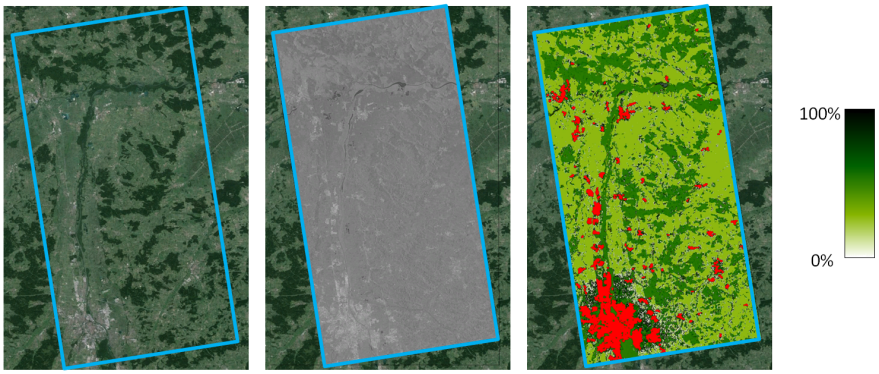

Figure 4: (Right) Forest classification map of a forested area located in the Bavaria region, Germany, overlaid on a GoogleEarth optical image. Urban areas are highlighted in red. For comparison, the γ^0 map (mid) and the optical GoogleEarth image (left) are depicted as well. The considered area is delimited by the blue rectangle, and extends by about 50 km $\times$ 30 km.

4 Additional Layers and First Examples

In this section, the algorithm and method described in Sections 2 and 3 are implemented and first example from TanDEM-X data are shown. The forest classification (membership function in (5)) for a bistatic TanDEM-X acquisition over a vegetated region in Brazil overlaid on a GoogleEarth optical image is shown on the right-hand side of Fig. 3. Dense forest (represented in dark green) and non-forest (light green) areas are clearly distinguishable. On the left-hand side of the figure the optical image of the area is given for comparison. The same classification map and the corresponding optical image are depicted for an area over temperate forest in Bavaria (Germany) in Fig. 4. One can observe that, in general, the different land cover types are correctly discriminated. The values of the membership function concentrate around 0.3 for bare areas and around 0.6-0.7 for the vegetated areas. This effect is consistent with the way the membership is calculated (see (5)) but also strongly depends on the "degree of separability" of the cluster centers, which is related to the imaging geometry employed for the specific acquisition (see Fig. 2). Urban areas are highlighted on the right-hand side of Fig. 4 and for their identification a dedicated algorithm has been implemented as explained in the following. For a better visual comparison, the SAR amplitude map is depicted in the centre of the figure as well.

4.1 Identification of Urban and Invalid Areas from TanDEM-X Quicklook Data

In urban areas InSAR performance is typically degraded due to the presence of additional geometrical distortions, such as layover and multiple reflections. Thus, if exploiting the coherence information only, urban settlements may represent an actual error source for forest classification and need to be identified. Looking at TanDEM-X quicklook images we verified that over urban regions the local distribution of the backscatter information can be suitably approximated by a uniform probability density function (pdf), whereas forested and bare (flat) areas can be better modeled with a Rayleigh distribution (for this analysis we have used local windows of 11 by 11 pixels). The reason for this is, on the one hand, the high dynamic range of backscatter due to the dominant presence of man-made structures and, on the other hand, the inherent average process for the generation of quicklook products (from about 3 to 50 m resolution, resulting in more than 250 looks), which strongly smooths the possible presence of local amplitude peaks. According to this, for the identification of urban settlements we implemented a method which is based on a Kolmogorov-Smirnov test of the local normalized backscatter γ^0 distribution by using a uniform pdf as reference. The Kolmogorov-Smirnov (K-S) statistic quantifies a distance between the empirical distribution function of the samples and the pdf function of the reference distribution. To further improve the classification capability we exploit the local backscatter variation as well (namely, the product of the local mean and the standard deviation of γ^0, being typically both larger than those observed over natural uniformly distributed targets). An additional check on the local DEM-derived slope (smaller than 10°) has been also applied to avoid mixing up areas affected by rugged terrain, which may exhibit similar local statistics due to the geometry-induced distortions. Region growing on the local K-S test and backscatter variation is finally performed to further look for neighboring settled areas. We did not validate the present method yet, but first experimental results show to be promising. As a next step, one could think to exploit full resolution TanDEM-X data with about 3 m resolution only on those areas which already indicate the presence of urban areas at quicklook level (hence keeping the computational load limited), to improve the classification accuracy.

In addition to urban settlements, regions characterized by rugged terrain are strongly influenced by geometric (baseline) decorrelation and the accuracy of the present classification method may be consequently affected. The compensation of the local slope in the computation of the

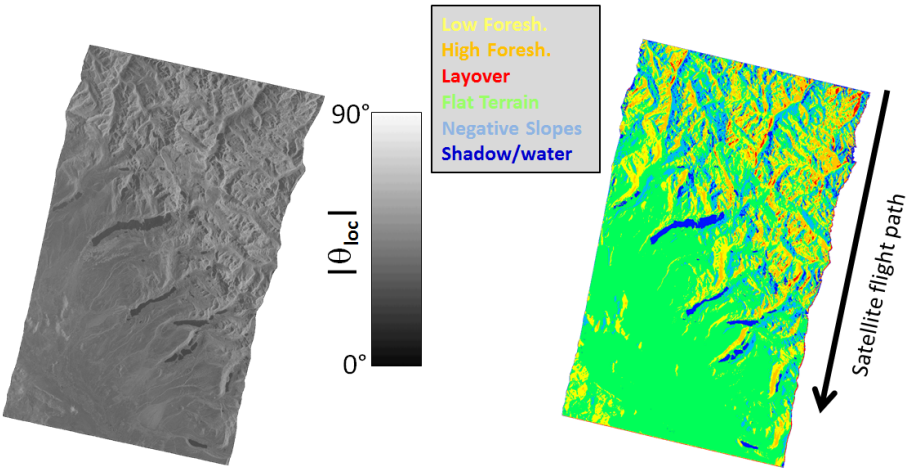

Figure 5: (Left) Local incidence angle map derived from the TanDEM-X quicklook DEM of a mountainous area in South-West Brazil and (right) geometry mask. The satellites operate in right looking geometry and their path is indicated on the right. For the identification of shadow, layover, as well as water areas the local slope together with the amplitude and coherence information is exploited.

volume decorrelation relies on the DEM height, which may be less accurate. As an example, on the left-hand side of Fig. 5 the local incidence angle map derived from the TanDEM-X DEM quicklook of a mountainous area in South-West Brazil is shown. From this, together with the local coherence and amplitude information [10], areas affected by geometrical distortions as well as water bodies can be identified, which are shown with different colors on the right-hand side of the figure.

4.2 Comparison with External Data

As a first validation, we have compared the classification results obtained from a TanDEM-X acquisition over the Amazon rainforest with the forest density provided by Landsat (about 1.5 million pixels). It has to be highlighted that the two estimation methods are completely independent, provide different information, and, therefore, one has to be very careful when comparing them. In particular, we set a reasonable threshold t_{TDM} = 50% for the TanDEM-X vegetation probability (P_{TDM}) and evaluate the probability of detection (P_d) and false alarm P_{fa} depending of different threshold values (t_{Lsat}) for the Landsat density information (D_{Lsat}), i.e.

$$P_d = P(P_{TDM} > t_{TDM} | D_{Lsat} > t_{Lsat}), \quad (6)$$

$$P_{fa} = P(P_{TDM} > t_{TDM} | D_{Lsat} < t_{Lsat}). \quad (7)$$

Looking at the resulting ROC (Receiver Operating Characteristics) curves we obtained the best performance for t_{Lsat} values already around 15%, for which a $P_d = 91\%$ and $P_{fa} = 14\%$ are obtained. These promising results need to be further investigated but somehow confirm that X-band InSAR is particularly sensitive to the presence of even very sparse forest.

Up to now, TanDEM-X has acquired the global land masses at least twice, exploiting different baselines. Additionally, many densely forested areas have been covered up to four times in a time frame of several years (from 2010 to 2015). Hence, it becomes clear that a proper mosaicking algorithm is needed to improve the final classification accuracy. A first idea would be to use an approach similar to the one used for the optimum combination of N multiple DEM height estimates [3]

$$\hat{Y} = \sum_{i=1}^{N} \alpha_i y_i. \quad (8)$$

For the definition of the α_i weights a sort of classification reliability can be derived as a function of the local SNR and of the inherent intercluster separation, which depends on the particular acquisition geometry (see Fig. 2). This aspect is currently under investigation and will be object of present and future studies. As an alternative, the manifold, multi-temporal TanDEM-X data set represents a powerful source to detect and monitor possible changes in the forest cover. An optical image acquired in the 1970's by the first Landsat satellites over an area located in the Amazon rainforest (and available in the GoogleEarth image archive) is shown in Fig. 6 (a). At that time the region was still predominantly covered by forest, but some clear cuts are already visible. Fig. 6 (b) and (c) show the coherence quicklooks acquired by TanDEM-X over the same area in January 2011 and October 2012, respectively. High-coherence (white) is obtained on bare soil areas, whereas lower coherence is observed in correspondence of the remaining forest. If compared to Fig. 6 (a), the effects of about 30 years of overwhelming deforestation activities are clearly visible. The slightly different baselines explain the higher coherence observed in Fig. 6 (c) with respect to Fig. 6 (b). Finally, Fig. 6 (d) shows the forest loss occurring between the two TanDEM-X acquisitions (brown spots), which has been derived by simply comparing the two coherence maps. The red circles in the figures highlight the areas where forest cover changes are most visible.

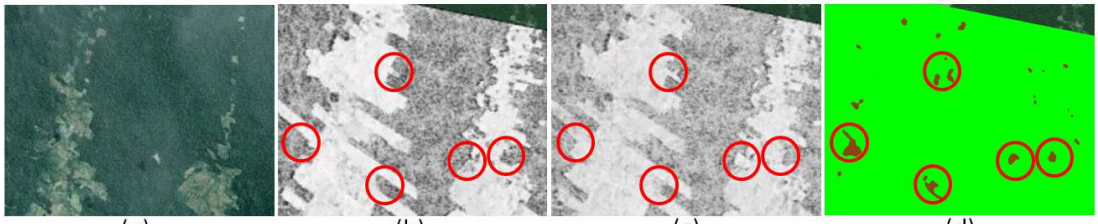

Figure 6: (a) Optical image acquired by the Landsat sensor in the 1970's (and available in the GoogleEarth image archive) over a region located in the Amazon rainforest and characterized by dense forest as well as by clear cuts. Coherence quicklook maps of a TanDEM-X acquisition commanded on (b) January 2^{nd}, 2011, and (c) on October 23^{rd}, 2012. (d) Forest loss map (brown: loss, green: no loss) occurred between the two TanDEM-X acquisitions. The areas of major interest are highlighted by the red circles in the figures (b)-(d).

5 Conclusions and Outlook

In this paper we have introduced a method to derive forest classification maps from TanDEM-X InSAR data. We exploit a classification algorithm based on fuzzy clustering, for which the coherence-derived volume decorrelation γ_{Vol} is used as input feature. Since the geometric acquisition configuration strongly influences the considered InSAR quantity, a multi-clustering approach is proposed: the data are divided into subsets according to the specific interferometric baseline and incidence angle, and the classification algorithm is independently implemented for each of them. Particular focus is given to the identification of additional information layers, represented by e.g. urban regions, water or areas affected by strong geometrical distortions. During the whole TanDEM-X mission duration, the global land masses have been acquired at least twice. In order to further improve the performance, many densely forested areas have been covered up to four times in a time frame of several years (from 2010 to 2015). Such a unique and manifold data set can be exploited to (1) improve the classification accuracy by properly mosaicking of multiple observations and, at the same time, to (2) get up-to-date information and (3) for detecting possible changes in the forest cover, as shown in Fig. 6. As next steps, the obtained results will be quantitatively validated with existing vegetation maps and by means of external land cover classification data.

References

[1] J. O. Sexton, X.-P. Song, M. Feng, P. Noojipady, A. Anand, C. Huang and D.-H. Kim, K. M. Collins, S. Channan, C. DiMiceli, and J. R. Townshend: *Global, 30-m resolution continuous fields of tree cover: Landsat-based rescaling of MODIS Vegetation Continuous Fields with lidar-based estimates of error*, International Journal of Digital Earth, Vol. 6, N. 5, pp. 427-448, 2013.

[2] M. Shimada, T. Itoh, T. Motooka, M. Watanabe, T. Shiraishi, R. Thapa, and R. Lucas, *Newglobal forest/non-forest maps from ALOS PALSAR data (2007−2010)*, Remote Sens. of Env., Vol. 155, pp. 13-31, May 2014.

[3] G. Krieger, A. Moreira, H. Fiedler, I. Hajnsek, M. Werner, M. Younis, and M. Zink: *TanDEM-X: A Satellite Formation for High-Resolution SAR Interferometry*, IEEE Trans. Geosci. Remote Sens., Vol. 1, N. 11, pp. 3317–3341, Nov. 2007.

[4] P. Rizzoli, M. Martone, and B. Bräutigam: *Global Interferometric Coherence Maps From TanDEM-X Quicklook Data*, IEEE Geosci. and Remote Sens. Lett., Vol. 11, N. 11, pp. 1861–1865, Nov. 2014.

[5] P. Rizzoli, M. Martone, and B. Bräutigam: *Global Mosaics of the Relative Height Error From TanDEM-X Quicklooks*, IEEE Geosci. and Remote Sens. Lett., Vol. 12, N. 9, pp. 1928–1932, Sept. 2015.

[6] M. Martone, P. Rizzoli, B. Bräutigam, and G. Krieger: *A method for generating forest/non-forest map from TanDEM-X interferometric data*, IEEE Int. Geosci. Remote Sens. Symp., Milan (Italy), July 2015.

[7] M. Martone, B. Bräutigam, P. Rizzoli, C. Gonzalez, M. Bachmann, and G. Krieger: *Coherence Evaluation of TanDEM-X Interferometric Data* ISPRS J. of Photogr. Remote Sens., Vol. 73, pp. 21-29, Sep. 2012.

[8] J. Bezdek, R. Ehrlich, and W. Full: *FCM: the fuzzy c-means clustering approach*, Computers and Geosciences, Vol. 10, pp. 191-203, December 1984.

[9] P. Rizzoli, M. Martone, and B. Bräutigam: *Greenland ice sheet snow facies identification approach using TanDEM-X interferometric data*, IEEE Int. Geosci. Remote Sens. Symp., Milan (Italy), July 2015.

[10] M. Martone, P. Rizzoli, B. Bräutigam and G. Krieger: *First two years of TanDEM-X mission: interferometric performance overview*, Radio Science, Vol. 48, pp. 617-627, Oct. 2013.