Random Forest classification of crop type using multi-temporal TerraSAR-X dual-polarimetric data

Rei Sonobe*[a], Hiroshi Tani[b], Xiufeng Wang[b], Nobuyuki Kobayashi[c] and Hideki Shimamura[d]

[a] *Graduate School of Agriculture. Hokkaido University, Sapporo, Hokkaido 060-8589, Japan*

[b] *Research Faculty of Agriculture. Hokkaido University, Sapporo, Hokkaido 060-8589, Japan*

[c] *Hokkaido Intellect Tank, Sapporo, Hokkaido 060-0004, Japan*

[d] *PASCO Corporation, Meguro-Ku, Tokyo 153-0043, Japan*

Corresponding author

Kita 9, Nishi 9 Kita-ku, Sapporo, 060-8589, Japan

reysnb@gmail.com

Random Forest classification of crop type using multi-temporal TerraSAR-X dual-polarimetric data

The classification maps are required for management and for the estimation of agricultural disaster compensation; however, those techniques have yet to be established. Some supervised learning models may allow accurate classification. In this study, the Random Forest (RF) classifier and the classification and regression tree (CART) were applied to evaluate the potential of multi-temporal TerraSAR-X dual-polarimetric data, on the StripMap mode, for classification of crop type. Furthermore, comparisons of the two algorithms and polarizations were carried out. In the study area, beans, beet, grasslands, maize, potato and winter wheat were cultivated, and these crop types were classified using the data set acquired in 2009. The classification results of RF were superior to those of CART and the overall accuracies were 0.91 to 0.93.

1. Introduction

Land-cover classification is one of the most common applications of remote sensing. Crop type classification maps are useful for estimating the amount of crops harvested or the agricultural disaster compensation, in addition to the management of the agricultural field. However, those techniques have yet to be established. Optical remote sensing is still one of the most attractive options for obtaining biomass information, as new sensors are available with fine spatial and spectral resolutions (Sarker and Nichol 2011). In addition, some optical satellites such as RapidEye and Landsat have been used for crop type classification (Hartfield *et al*. 2013; Krahwinkler and Rossmann 2013). However, cloud cover strongly limits the number of available optical images. Radar provides a useful tool for monitoring agricultural fields, since it is unaffected by cloud cover or low solar zenith angles (Bindlish and Barros 2001). Furthermore, significant information about soil and vegetation parameters has also been obtained through microwave remote sensing, and these techniques are increasingly being used to manage land and water resources for agricultural applications (Fontanelli *et al*. 2013). Because, unlike passive systems, synthetic aperture radar (SAR) systems are not dependent on atmospheric influences or weather conditions, they are especially suitable for a multi-temporal classification approach. The first large backscatter coefficient change occurs as a result of plowing and seeding. Then, smaller changes occur due mainly to variations of biomass and plant water content, and, for SAR data, to changes in plant structure. Furthermore, harvesting causes large backscatter coefficient changes.

SAR signals acquired under different polarizations will obtain different backscatter responses, providing more information about vegetation (Brisco *et al*. 2013). TerraSAR-X was launched on June 15, 2007, and X-band SAR data are widely available and operated with several polarizations. Furthermore, several studies have proven the high geometric accuracy of TerraSAR-X (Ager and Bresnahan 2009). Within this framework, the main objective of the present study is to evaluate the potential of Terra-SAR-X dual-polarimetric data including Horizontal transmit - Horizontal receive polarization (HH) and Vertical transmit - Vertical receive polarization (VV) for crop type classification. The purpose of this study is to evaluate the classification results using TerraSAR-X data.

2. Materials and methods

The experimental area of this study is the farming area in western Tokachi plain, Hokkaido, Japan (142° 55′ 12″ to 143° 05′ 51″ E, 42° 52′ 48″ to 43° 02′ 42″ N). The 4,955 fields (1,053 beans fields, 709 beet fields, 623 grasslands, 254 maize fields, 831 potato fields and 1,485 winter wheat fields) covered all the areas. The mean size of the fields is 2.16 ha (the maximum area is 18.0 ha and the smallest area is 0.01 ha).

Sixteen TerraSAR-X images were acquired in StripMap mode (incidence angle of 42.3°) between May 2 and November 5, 2009 with 2.75 m pixel resolution. Multi-temporal sigma nought (σ^0) images, which are the the radiometrically calibrated power images referenced to the ground, have been revealed to be effective for crop type classification (Bargiel and Herrmann 2011). However, it is thought that the gamma nought (γ^0) images, which are the radiometrically calibrated power images spaced equally, are better than σ^0 images for crop classification since they are less dependent on the incidence angle. Therefore, in this study, Level 1B Enhanced Ellipsoid Corrected products were converted from digital numbers to γ^0. In order to reduce speckle and to avoid problems related to uncertainty in georeferencing, the pixel values of each fields were averaged and the γ^0 values of HH and VV were extracted from every acquisition (Bargiel and Herrmann 2011; Koppe *et al*. 2013).

The reference data was provided by Tokachi Nosai (http://www.tokachi-nosai.or.jp/) as a polygon shape file in which the position of the fields and attribute data such as crop types were included. It was based on the report of the farmers and we considered it correct.

The land cover classification accuracy of the Random Forest (RF) from optical imagery was superior to the result of the maximum likelihood classifier (Rodriguez-Galiano *et al*. 2012). Furthermore, RF is requires two parameters only to be set whereas the Support Vector Machine (SVM) requires a number of user-defined parameters the classification results are equally well to SVM (Pal 2005). Then, this algorithm was adopted in this study. RF is an ensemble learning technique that builds multiple trees based on random bootstrapped samples of the training data (Breiman 2001). Each tree is built using a different subset from the original training data, containing about two thirds of the cases, and the nodes are split using the best split variable out of randomly selected variables (Liaw and Wiener 2002). Through this strategy, RF is robust to over-fitting and can handle thousands of input variables (dependent or independent) without variable deletion. The output is determined by a majority vote of the trees. Two user-defined parameters are the number of trees (*k*) and the number of variables used to split the nodes (*m*); when increasing the number of trees, the generalization error always converges, and over-training is not a problem. On the other hand, a reduction in the number *m* of predictive variables results in each individual tree of the model being weaker; therefore, picking a large number of trees is recommended, as is using the square root of the number of variables for the value of *m* (Breiman 2001). The samples which are not present in the training subset are included as part of another subset called out-of-bag (OOB). These OOB elements, which are not considered for the training of the tree, can be classified by the tree to evaluate performance. The proportion between the misclassifications and the total number of OOB elements contributes an unbiased estimation of the generalisation error (Rodriguez-Galiano *et al*. 2012). RF uses the Gini Index as a measure for the best split selection, which measures the impurity of a given element with respect to the rest of the classes (Breiman *et al*. 1984). The data with a higher Gini Index is more important for discrimination. Thus, by using a given combination of features, a decision tree is made to grow up to its maximum depth with no pruning.

Hartfield *et al*. (2013) demonstrated that the classification and regression tree (CART) was the best algorithm among CART, the maximum likelihood classifier and Object-oriented classifiers for mapping crops in Arizona, using Landsat 5 Thematic Mapper (TM) imagery. Therefore, CART was also produced for comparison in this study.

These classifications algorithms were applied using R (R Core Team 2013), 'randomForest' package (Liaw and Wiener 2002) and 'rpart' package (Therneau *et al*. 2013). We used a stratified random sampling approach to select the fields used for training (Foody, 2009) and 20% of the crop fields were selected at random as training samples (Hartfield *et al*. 2013). Table 1 represents the numbers of fields of each crop type. In addition, the classifications were performed using HH, VV and the combination (HH+VV) in order to compare and understand the utilization of these polarizations.

The classification maps were evaluated in terms of their overall accuracy (OA), producer's accuracy (PA), user's accuracy (UA) and the kappa index of agreement (κ). Furthermore, the two simple measures of quantity disagreement (QD) and allocation disagreement (AD), which are much more useful to summarize a cross-tabulation matrix than the kappa index of agreement was used for evaluation (Pontius and Millones 2011). The Jeffries-Matusita distance (Richards 1999) and the Bhattacharyya distance (Fukunaga 1990) were also calculated to compare the statistical separability among crop types. The value of the Jeffries-Matusita distance measurement ranges from 0 to 2.0 and indicates how well the selected two crop types are statistically separated; values close to 1.9 indicate that the two crop types have good separability. Bhattacharyya distance is a distance measure in multidimensional space that takes into consideration the means and covariances of the samples. Lee and Choi (2000) presented a way to estimate the classification accuracy from Bhattacharyya distance values. According to their method, values greater than 1.47 correspond to less than 5% classification error. Laurin *et al.* (2013) used the Jeffries-Matusita distance to evaluate the separability of vegetations using SAR and lidar data and Miettinen and Liew (2011) used Bhattacharyya distance to evaluate the Separability of insular Southeast Asian woody plantation species in the Advance Land Observing Satellite (ALOS) / Phased Array L-band Synthetic Aperture Radar (PALSAR) data. Then, they indicate that these two separability measurements are useful for SAR data.

3. Results and Discussion

For application of RF, the number of trees was tuned and Figure 1 represents the relationships between the number of trees and the error rate for OOB. Since the results indicate that the number of more than approximately 50 is suitable for the three data sets (HH, VV and HH+VV), 50 was chosen as the number of trees in this study.

< Figure 1>

Figure 2 shows the relative importance of the contribution to the RF classification model. According to the Gini index, the features with the greater contribution to the classification model are the data acquired during June to August (the highest values were observed on 7 July for HH and on 26 June for VV). In this period, all kinds of crops were cultivated and the influence of soil were ignorable for all fields while the SAR data had high sensitivity to soil moisture or roughness due to sparse vegetation covers before June (seedlings or transplanting of beans, beet and maize) and after August (the harvest season of winter wheat).

< Figure 2>

< Table 2>

< Table 3>

The classifications accuracies are given in Table 2 and the results of the classification comparisons using *Z*-test (Congalton and Green 1999) are given in Table 3. RF is superior to CART and the best result was obtained using RF for HH+VV and the overall accuracy of the classification was 0.93 with a kappa of 0.91 (this case also offers the best results in terms of QD and AD). However, the difference of the classification results was not meaningful ($p < 0.05$) between HH+VV and HH. The worst result was obtained using CART for VV and the difference of the classification results was not meaningful between VV and HH. In particular, the PA and the UA for maize were 0 for VV. The Jeffries-Matusita distance and the Bhattacharyya distance were calculated to compare the statistical separability among beans and maize (Figure 3). Due to the low values for both polarization (the Jeffries-Matusita distance measurements were less than 0.20 and the Bhattacharyya distance measurements were less than 0.11), it is difficult to distinguish between beans and maize using backscattering coefficient values such as γ^0. The discrimination may be improved with the use of the polarimetric parameters such as averaged alpha angle and entropy.

<Table 3>

< Figure 3>

4. Conclusions

In order to generate classification maps, in this study sixteen StripMap images from TerraSAR-X were used. First, the performances of the two algorithms for crop type classifications were evaluated and the results of RF were superior to those of CART in terms of some indices of agreement. The overall accuracies were over 90 %. Next, the comparisons of polarizations were carried out and it was revealed that the γ^0 of HH polarization was more useful than that of VV and the combination of the two polarization was not effective for RF classification by *Z*-test.

Although it is possible to discriminate crop types from agricultural fields using RF, the low producer's accuracies were observed for maize because of low separability between them.

Acknowledgements

Thanks are addressed to Dr Timothy Warner and all reviews for their helpful comments.

References

AGER, T.P. and P.C. BRESNAHAN. 2009. Geometric precision in space radar imaging: Results from Terrasar-X. *In ASPRS 2009 annual conference*, pp.9-13. Baltimore, Maryland, USA.

BARGIEL, D. and S. HERRMANN. 2011. Multi-temporal land-cover classification of agricultural areas in two european regions with high resolution spotlight Terrasar-X data. *Remote Sensing*, **3**, pp.859-877.

BINDLISH, R. and BARROS, A.P. 2001. Parameterization of vegetation backscatter in radar-based, soil moisture estimation. *Remote Sensing of Environment*, **76**, pp.130-137.

BREIMAN, L. 2001. Random forests. *Machine Learning*, **45**, pp.5-32.

BREIMAN, L., FRIEDMAN, J., STONE, C. J. and OLSHEN, R. A. 1984. Classification and regression trees (Belmont, CA: Chapman and Hall/CRC).

BRISCO, B., A. SCHMITT, K. MURNAGHAN, S. KAYA and Roth, A. 2013. SAR polarimetric change detection for flooded vegetation. *International Journal of Digital Earth*, **6**, pp.103-14.

CONGALTON, R.G. and GREEN, K., 1999, Assessing the Accuracy of Remotely Sensed Data: Principles and Practices (Boca Raton, FL: Lewis Publishers).

FONTANELLI, G., PALOSCIA, S., ZRIBI, M. and CHAHBI, A. 2013, Sensitivity analysis of X-band SAR to wheat and barley leaf area index in the merguellil basin. *Remote Sensing Letters*, **4**, pp.1107-1116.

FOODY, G. M. 2009, Sample Size Determination for Image Classification Accuracy Assessment and Comparison. *International Journal of Remote Sensing*, **30**, pp.5273-5291.

FUKUNAGA, K., 1990, Introduction to Statistical Pattern Recognition. Second edition(Amsterdam, the Netherlands: Academic Press Inc.).

HARTFIELD, K., MARSH, S., KIRK, C. and CARRIERE, Y. 2013, Contemporary and historical classification of crop types in arizona. *International Journal of Remote Sensing*, **34**, pp.6024-6036.

KOPPE, W., M. GNYP, M., HUTT, C., YAO, Y., MIAO, Y., CHEN, X. and BARETH, G. 2013, Rice monitoring with multi-temporal and dual-polarimetric Terrasar-X data. *International Journal of Applied Earth Observation and Geoinformation*, **21**, pp.568-576.

KRAHWINKLER, P. and ROSSMANN, J. 2013, Tree species classification and input data evaluation. *European Journal of Remote Sensing*, **46**, pp.535-549.

LAURIN, G. V., FRATE, F. D., PASOLLI, L., NOTARNICOLA, C., GUERRIERO, L. and VALENTINI, R. 2013, Discrimination of vegetation types in alpine sites with ALOS PALSAR-, RADARSAT-2-, and lidar-derived information. *International Journal of Remote Sensing*, **34**, pp.6898-6913.

LEE, C. and CHOI, E., 2000, Bayes error evaluation of the Gaussian ML classifier. *IEEE Transactions on Geoscience and Remote Sensing*, **38**, pp.1471-1475.

LIAW, A. and WIENER, M. 2002, Classification and regression by random forest. *R News*, **2**, pp.18-22.

MIETTINEN, J. and LIEW, S.C. 2011, Separability of insular Southeast Asian woody plantation species in the 50 m resolution ALOS PALSAR mosaic product. *Remote Sensing Letters*, **2**, pp.299-307.

PAL, M. 2005, Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, **26**, pp.217-22.

PONTIUS JR., R.G., MILLONES, M. 2011, Death to Kappa: birth of quantity disagreementand allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*, **32**, pp.4407-4429.
R CORE TEAM. 2013, R, "A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org/.
RICHARDS, J. A. 1999, Remote Sensing Digital Image Analysis. Berlin: Springer-Verlag.
RODRIGUEZ-GALIANO, V., CHICA-OLMO, M., ABARCA-HERNANDEZ, F., ATKINSON, P. and JEGANATHAN, C. 2012, Random forest classification of mediterranean land cover using multi-seasonal imagery and multi-seasonal texture. *Remote Sensing of Environment*, **121**, pp.93-107.
SARKER, L. and NICHOL, J. 2011, Improved forest biomass estimates using alos avnir-2 texture indices. *Remote Sensing of Environment*, **115**, pp.968-977.
THERNEAU, T., ATKINSON, B. and RIPLEY, B. 2013, Package ‘rpart’ URL http://cran.r-project.org/web/packages/rpart/rpart.pdf.

Table 1. Crop type and number of fields.

Crop type	No. of fields	
	Test data	Training data
Beans	842	211
Beet	567	142
Grassland	499	124
Maize	204	50
Potato	664	167
Wheat	1188	297

Table 2. Accuracy results for RF and CART.

Statistic	Class	RF			CART		
		HH+VV	HH	VV	HH+VV	HH	VV
PA							
	Beans	0.919	0.905	0.898	0.792	0.827	0.822
	Beet	0.959	0.952	0.959	0.935	0.922	0.940
	Grasslands	0.938	0.940	0.938	0.844	0.872	0.844
	Maize	0.534	0.475	0.265	0.417	0.353	0.000
	Potato	0.935	0.947	0.946	0.899	0.756	0.890
	Wheat	0.982	0.981	0.974	0.943	0.945	0.943
UA							
	Beans	0.879	0.872	0.826	0.809	0.753	0.736
	Beet	0.932	0.923	0.943	0.826	0.853	0.826
	Grasslands	0.911	0.916	0.902	0.863	0.784	0.863
	Maize	0.886	0.829	0.659	0.733	0.661	0.000
	Potato	0.932	0.918	0.929	0.841	0.841	0.836
	Wheat	0.976	0.979	0.968	0.946	0.963	0.946
OA		0.929	0.924	0.910	0.863	0.845	0.847
κ		0.911	0.905	0.887	0.828	0.806	0.807
QD		2.043	2.195	3.078	3.052	4.642	5.525
AD		5.045	5.399	5.928	10.671	10.822	9.788

Note: PA, producer's accuracy; UA, user's accuracy; OA, overall accuracy; κ, kappa index of agreement; QD, quantity disagreement; AD, allocation disagreement.

Table 3. *Z*-test results.

		RF			CART		
		HH+VV	HH	VV	HH+VV	HH	VV
RF	HH+VV		0.87	3.22	9.92	12.17	12.09
	HH			2.35	9.06	11.32	11.24
	VV				6.74	9.00	8.92
CART	HH+VV					2.26	2.18
	HH						0.08
	VV						

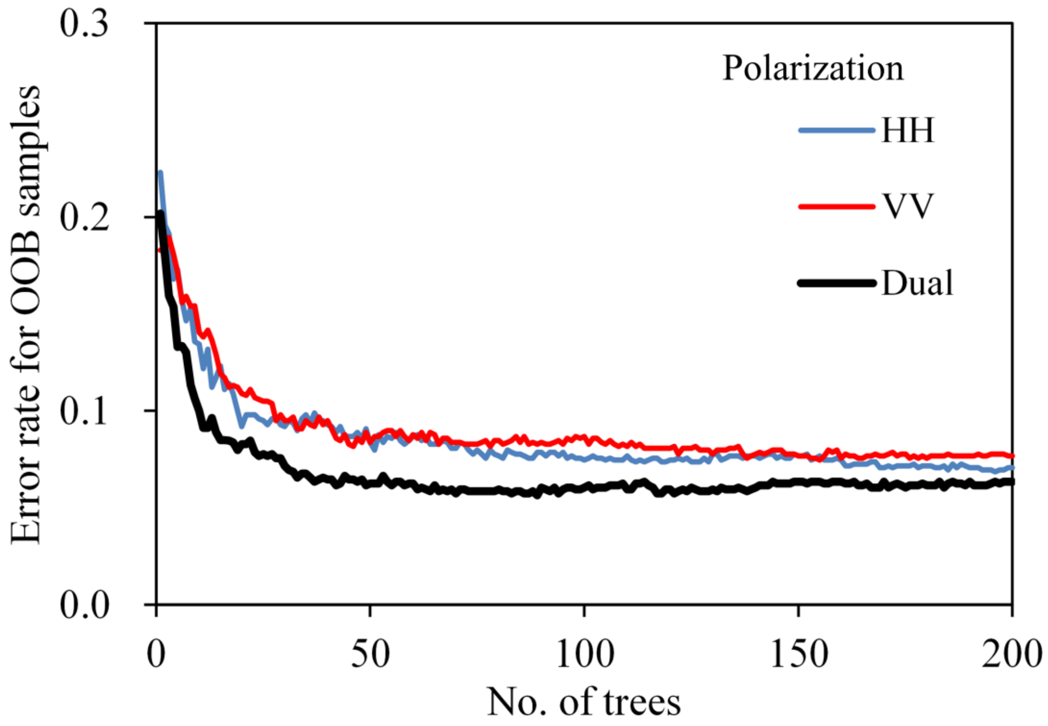

Figure1. Relationships between number of trees and error rate for OOB samples for HH, VV and Dual (HH+VV) polarization data.

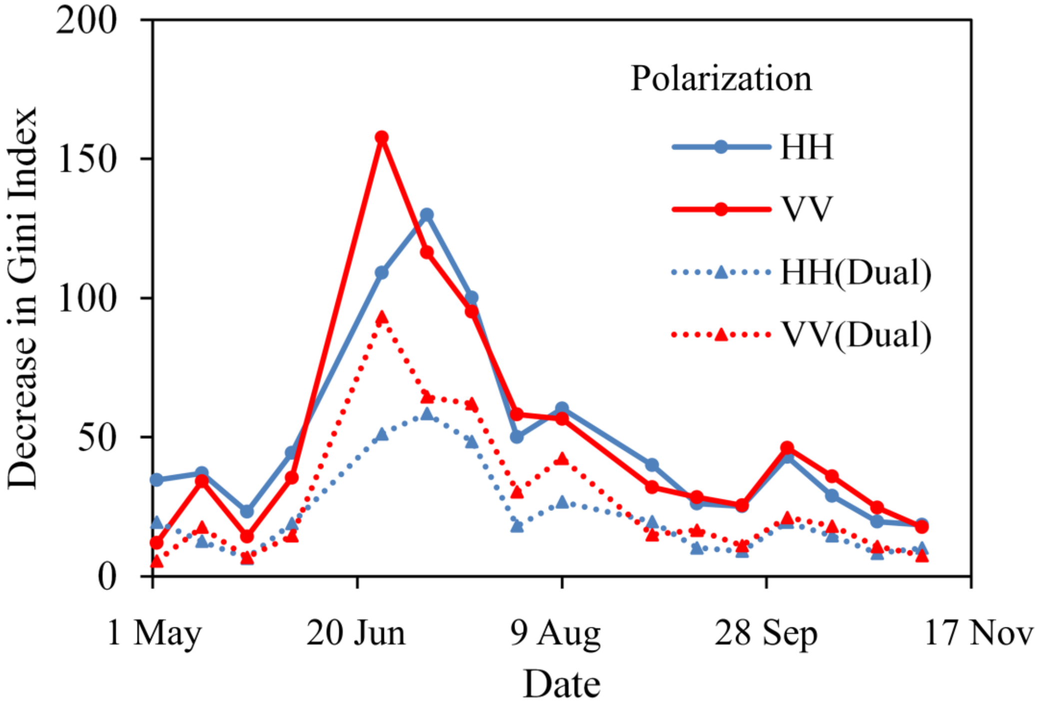

Figure 2. Importance of data acquisition date based on Gini measures.

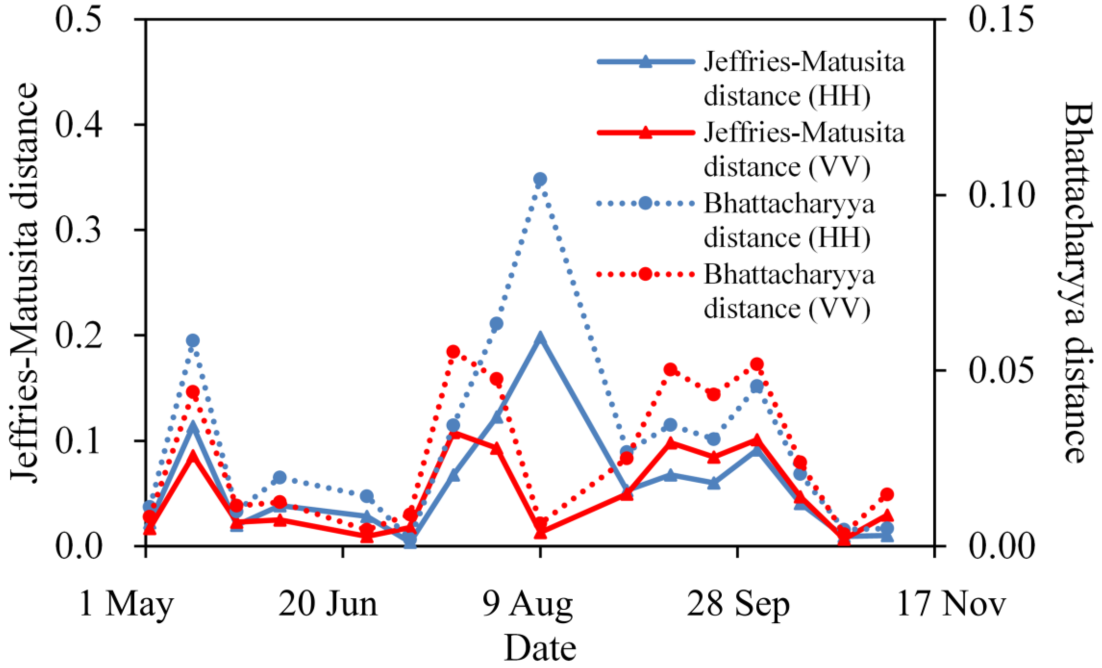

Figure 3. Jefferi-Matusita distance and Bhattacharyya distance between beans and maize for TerraSAR-X data.