Scheffler, D., Frantz, D., Segl, K. (2020): Spectral harmonization and red edge prediction of Landsat-8 to Sentinel-2 using land cover optimized multivariate regressors. - Remote Sensing of Environment, 241, 111723.

https://doi.org/10.1016/j.rse.2020.111723

SPECTRAL HARMONIZATION AND RED EDGE PREDICTION OF LANDSAT-8 TO SENTINEL-2 USING LAND COVER OPTIMIZED MULTIVARIATE REGRESSORS

*Daniel Scheffler[a], David Frantz[b], Karl Segl[a]

[a] Helmholtz Center Potsdam GFZ German Research Center for Geosciences, Section Remote Sensing, Telegrafenberg, Potsdam, Germany, 14473

[b] Geography Department, Humboldt-Universität zu Berlin, Unter den Linden 6, Berlin, Germany, 10099

*Correspondence: daniel.scheffler@gfz-potsdam.de; Tel.: +49-331-288-1198; Fax: +49-331-288-1192

ABSTRACT

Multi-sensor remote sensing applications consistently gain importance, boosted by a growing number of freely available earth observation data, increasing computing capacity, and increasingly complex algorithms that need as temporally dense data as possible. Using data provided by different sensors can greatly improve the temporal resolution of time series, fill data gaps and thus improve the quality of land cover monitoring applications. However, multi-sensor approaches are often adversely affected by different spectral characteristics of the sensing instruments, leading to inconsistencies in downstream products. Spectral harmonization, i.e., the transformation of one sensor into the spectral domain of another sensor, may reduce these inconsistencies. It simplifies workflows, increases the reliability of subsequently derived multi-sensor products and may also enable the generation of new products that are not possible with the initial spectral definition. In this paper, we compare the effect of multivariate spectral harmonization techniques on the inter-sensor reflectance consistency and derived products such as spectral indices or land cover classifications. We simulated surface reflectance data of Landsat-8 and Sentinel-2A from airborne hyperspectral data to eliminate any sources of error originating from unequal acquisition geometries, illumination or atmospheric state. We evaluate different methods based on linear, quadratic and random forest regression as well as linear interpolation, and predict not only matching but also unilaterally missing bands (red edge). We additionally consider material-dependent spectral characteristics in the harmonization process by using separate transformation functions for spectral clusters of the input dataset. Our results suggest that spectral harmonization is useful to improve multi-sensor consistency of remote sensing data and subsequently derived products, especially if multiple transformation functions are incorporated. There is a strong dependency between harmonization performance and the similarity of source and target sensor's spectral characteristics. For spectrally transforming Landsat-8 to Sentinel-2A, we achieved the lowest radiometric inter-sensor deviations with 50 spectral clusters and linear regression. Based on simulated data, deviations are below 1.7% reflectance within the red edge spectral region and below 0.3% reflectance for the remaining bands (RMSE). Regarding spectral indices, our results show a

reduction of inter-sensor deviation (vegetation pixels only) to 38% of the initial error for NDVI (Normalized Difference Vegetation Index) and to 43% for EVI (Enhanced Vegetation Index). Furthermore, we computed the REIP (Red Edge Inflection Point) with an accuracy of 3.1 nm from Sentinel-2 adapted Landsat-8 data. An exemplary multispectral classification use case revealed an increasing inter-sensor consistency of classification results from 92.3% to 97.3% mean error. Applied to time series of real Landsat-8 and Sentinel-2 data, we observed similar trends, albeit intermingled with non-sensor-induced inconsistencies.

Index Terms **–** spectral harmonization, satellite image harmonization, machine learning, time series analysis, analysis ready data

1. INTRODUCTION

Optical multispectral sensors like the Landsat legacy sensors deliver valuable information about the Earth's surface for more than four decades (Wulder et al., 2019). Many new sensors provide continuity with this record and continuously acquire huge amounts of image data, e.g., Landsat-8 (Irons et al., 2012) and Sentinel-2A/B (Drusch et al., 2012). All these data are suited for monitoring the changing Earth surface over time. Such time series analyses require in particular sensors with a short revisit time to detect and monitor slightest changes on the Earth's surface in a timely manner. However, this is often not sufficient, because frequently clouds mask the desired information. Therefore, it is obvious to additionally use data from multiple multispectral sensors to fill the temporal gaps (Brown et al., 2006) and hence to increase the chance of cloud-free acquisitions. Needless to say, this is a very challenging task due to different spectral or spatial sensor characteristics, different acquisition geometries or illumination conditions, or different atmospheric states (Pacifici et al., 2014; Roy et al., 2016a).

In this paper, we focus on the spectral domain. Solutions for homogenizing the spatial domain are already existing and well-tested, including methods for improving the co-registration between sensors (e.g., Scheffler et al., 2017) and for adjusting the spatial resolution of one sensor to the other (e.g., Gao et al., 2006). Different acquisition or illumination conditions as well as variations in the atmospheric conditions may also cause large variations in the measured signal (Roy et al., 2016a; Schaepman-Strub et al., 2006; Steven et al., 2003; Teillet, 1986). Therefore, images delivered in top-of atmosphere (TOA) radiance or reflectance have to be converted to bottom-of-atmosphere (BOA) reflectance for harmonization (Hall et al., 1991; Roy et al., 2016a). Numerous approaches exist, such as ATCOR (Richter and Schläpfer, 2002, 2019), LaSRC (Vermote et al., 2016), FORCE AC (Frantz et al., 2016) or Sen2Cor (Louis et al., 2016) and have been recently compared in Doxani et al., 2018. In the study presented herein, we rely on Level-2 data that already have undergone several steps of radiometric correction, normalization of acquisition and illumination geometries as well as atmospheric correction.

From the spectral side, multi-sensor, multispectral remote sensing applications are often affected by variations of the spectral band positions and width of the sensors used (Hong and Zhang, 2008; Steven et al., 2003; Teillet et al., 1997). Although many multispectral sensors, such as the Landsat sensors or Sentinel-2 have very similar bands in the visible and near infrared wavelength range, they are not strictly identical. Moreover, many sensors provide specific bands, e.g., in the red edge spectral region. The Sentinel-2 MSI provides 13 bands including three extra red edge bands (Drusch et al., 2012) compared to Landsat-8 OLI. Therefore, multi-sensor applications often use only the most similar bands, given that they are subject to the same or at least similar physical principles. Some applications may even leave out sensors that don't provide a band at a certain wavelength position. But even if spectral bands with a similar wavelength position exist, differences in the spectral responses of the sensors will create slightly different pixel values. This in turn creates variations in spectral index values or classification maps and thus it can lead to misinterpretations of the results that can solely be traced back to unequal spectral characteristics of the input sensors (Teillet et al., 1997).

Possible solutions are to evolve sensor specific processing chains, separate parameter retrieval models or classification approaches (e.g., Hagolle et al., 2010; Useya and Chen, 2018; Zhu et al., 2015) or to inter-calibrate multi-sensor products such as spectral indices or band ratios (e.g., Brown et al., 2006; Gallo et al., 2005; Miura et al., 2006; Steven et al., 2003). However, this is very time consuming and requires comprehensive validation techniques to achieve reliable results. Alternatively, spectral band harmonization may be used to approximate the spectral information at a certain wavelength position as it would have been acquired by another sensor. This allows to easily create inter-sensor data cubes, simplifies processing workflows and provides comparability of the harmonized datasets and subsequently generated analysis results. In this regard, research has been conducted to build up statistical relations between the acquired signal of equivalent multi-sensor bands. The following table summarizes some exemplary studies and highlights how our study is different from them:

Table 1. Examples from the literature for spectral harmonization techniques applicable to multispectral remote sensing data.

Reference	Harmonization target/source	Harmonization technique(s)	Main differences to the study presented herein
Chastain et al. (2019)	Landsat-8 from Landsat-7, Sentinel-2 from Landsat-7 Sentinel-2 from Landsat-8	Univariate linear regression	Harmonization includes common bands only; static transformation coefficients per spectral band
Claverie et al. (2017, 2018)	Landsat-8 from Sentinel-2	Univariate linear regression	Harmonization includes common bands only; static transformation coefficients per spectral band
Flood (2014)	Landsat-7 from Landsat-8	Univariate linear regression	Harmonization includes common bands only; static transformation coefficients per spectral band
Flood (2017)	Landsat-7 from Sentinel-2, Landsat-8 from Sentinel-2	Univariate linear regression	Harmonization includes common bands only; static transformation coefficients per spectral band
Hong & Zhang (2008)	IKONOS from QuickBird	Various tested, histogram matching worked best	Each IKONOS band covered by an equivalent QuickBird band with very similar relative spectral response
Roy et al. (2016a)	Landsat-8 from Landsat-7; Landsat-7 from Landsat-8	Univariate linear regression	Harmonization includes common bands only; static transformation coefficients per spectral band
Zhang et al. (2018)	Landsat-8 from Sentinel-2, Sentinel-2 from Landsat-8	Univariate linear regression	Harmonization includes common bands only; static transformation coefficients per spectral band

For example, Claverie et al. (2018) achieved good inter-sensor consistency for Sentinel-2 data transformed to the spectral domain of Landsat-8. However, the underlying techniques only use a single set of harmonization coefficients that have been approximated based on average band-to-band relationships but without respect to material-dependent spectral characteristics. This causes spatially variable prediction performances depending on the surface coverage since a single set of transformation coefficients does not perform equally well for all materials, but leads to larger errors the stronger spectra deviate from the global average. This effect was also mentioned by Flood (2014) as a limitation of his study to be addressed in future. To overcome this, separate transformation functions for different surface materials might be helpful. Apart from that, the largest differences in the spectral response of source and target bands have been investigated by Flood (2014, 2017), Roy et al. (2016a) and Chastain et al. (2019). They occur at the transformation between Landsat-7 and Landsat-8 data since the bandwidth of some bands has been significantly narrowed with Landsat-8 (Irons et al., 2012; Mishra et al., 2016). However, none of the mentioned studies attempted to predict data in unilaterally missing bands (e.g., red edge bands), in which case those approaches are not expected to work properly as the relationships were only estimated based on homologous bands.

For larger spectral differences, therefore, new methods are required that accurately estimate the reflectance signal within so far not covered wavelength regions. In this context machine learning seems to be particularly suitable – especially behind the background that there are direct material-dependent relations between the remotely sensed signal in covered and not covered spectral regions due to specific spectral characteristics of different surface materials (Flood, 2014; Miura et al., 2006; Roy et al., 2016a). This means, that even if, e.g., the red edge spectral region is not covered by a sensor, it might be predicted with some uncertainty, given that the material can be identified using the known signal.

In this study we investigate the potential to estimate the spectral information of Sentinel-2 from Landsat-8. These sensors are widely used and particularly suited for this study because they provide both bands with very similar spectral characteristics but also "spectral gaps", i.e., spectral regions only covered by a single sensor (Figure 1). A special focus is placed on wavelength ranges that are not covered by Landsat-8 (mainly the red edge). In particular, to draw attention to the specific spectral properties of different surface materials, we developed a new method for spectral harmonization using multivariate machine learning techniques combined with spectral clustering. More precisely, we utilize separate transformation functions for various spectral clusters to adequately transform the acquired signal of Landsat-8 into the spectral domain of Sentinel-2. The main research questions of this paper are:

1. How strong is the influence of spectral harmonization on the inter-sensor consistency of spectral information and subsequent products, such as spectral indices or classification maps?
2. May separate prediction functions for different spectral clusters be useful to more accurately predict the spectral information of the target sensor?

3. Which harmonization accuracies can be achieved within unilaterally existing wavelength regions and how does this accuracy vary with different numbers of spectral clusters?

2. DATA

2.1 Simulated Landsat-8 and Sentinel-2 data from hyperspectral images

A meaningful comparison of spectral harmonization techniques for multi-sensor remote sensing data requires a multi-sensor data basis where the individual acquisitions ideally only differ in the spectral responses of the input sensors. Any other side effects affecting the sensed signal, such as different spatial sensor characteristics or registration accuracies, acquisition, illumination or atmospheric conditions (section 1) should be minimized or not existing. In this manner, it is possible to accurately measure the deviation between the predicted and the actual reference spectral information. Therefore, we accomplished our study as a simulation study based on hyperspectral data that can be accurately transformed to artificial multispectral data as it would be acquired by different sensors. This has been similarly done by Claverie et al. (2018) and is generically applicable across different sensors requiring the knowledge of the sensor's spectral response functions.

For this purpose, we selected representative BOA reflectance spectra from nine hyperspectral airborne remote sensing images covering different climatic zones (Table 2, train data) to maximize the spectral variety. Selected spectra were used to generate a training database of multi-spectral signatures for Landsat-8 and Sentinel-2A. Additionally, independent verification spectra were selected from three additional hyperspectral test datasets to evaluate the quality of the harmonization results.

Table 2. Data characteristics of the hyperspectral input data used in this study.

Train/Test	Sensor	Location	UTM Zone	GSD [m]	Major land cover components								Provider
					snow/ ice	rocks / bare soil	clouds	rural/ farmland	forest/ bushland	urban	desert	water	
Train 1	AISA Eagle/ AISA Hawk	Isabena, Spain	31N	4		x		x	x				GFZ [1]
Train 2	APEX	Balaton lake, Hungary	34N	5			x	x	x	x		x	VITO
Train 3	APEX	Brussels, Belgium	31N	2					x	x		x	BELSPO
Train 4	APEX	Neusling, Germany	33N	4				x	x	x			LMU [2]
Train 5	HyMap	Northern Quebec, Canada	18N	2	x	x							DLR
Train 6	HyMap	Costa Rica	16N	15		x	x	x	x				DLR
Train 7	HyMap	Döberitzer Heide, Germany	33N	4		x		x	x	x		x	GFZ [3]
Train 8	HyMap	Mullewa, Western Australia	50S	4.2		x		x			x		CSIRO
Train 9	HyMap	Arundale, South Africa	35S	3.3		x			x		x		DLR
Test 1	HyMap	Berlin, Germany	33N	3.5		x		x	x	x		x	GFZ
Test 2	HyMap	Dresden, Germany	33N	4		x		x	x	x		x	GFZ
Test 3	HyMap	Potsdam, Germany	33N	4		x		x	x	x		x	GFZ

[1] Foerster et al. (2015)
[2] Hank et al. (2015)
[3] Neumann et al. (2015)

Explanation of provider abbreviations:
GFZ – German Research Centre for Geosciences (www.gfz-potsdam.de); BELSPO – Belgian Science Policy (www.belspo.be); LMU – Ludwig-Maximilians-Universität München (www.uni-muenchen.de); DLR – German Aerospace Centre (www.dlr.de); CSIRO - Commonwealth Scientific and Industrial Research Organisation (www.csiro.au)

The workflow to generate the spectral database is presented in the following. Multi-spectral signatures have been generated as in Steven et al. (2003) through a spectral up-sampling of the hyperspectral BOA reflectance to 1 nm resolution followed by the computation of weighted averages according to the spectral response functions of Landsat-8 and Sentinel-2A (Figure 1). An exemplary vegetation spectrum is shown in Figure 1 after transforming the hyperspectral signature into the multispectral domains (dashed line for each sensor). To ensure a high spectral variability of the sample spectra we intentionally chose images containing a high land cover variety without lacking any major component and separated 50 spectral clusters from each hyperspectral image using a K-Means clustering approach (implemented after Lloyd, 1982). Next, we randomly extracted 2,000 hyperspectral signatures from each cluster to guarantee a high statistical variety and convolved these spectra with the spectral response functions of both multispectral sensors as presented in Figure 1. This way, we extracted 100,000 spectra from each hyperspectral image and then combined the spectra of all images into a single "reference cube" for each sensor.

For this study, the generated reference cubes represent an ideal data basis as they contain a comprehensive variation of multispectral signatures from most important surface cover types around the world. They solely differ in their spectral dimension so that statistical relations between different multispectral sensors, as needed for their spectral harmonization, can be easily deduced.

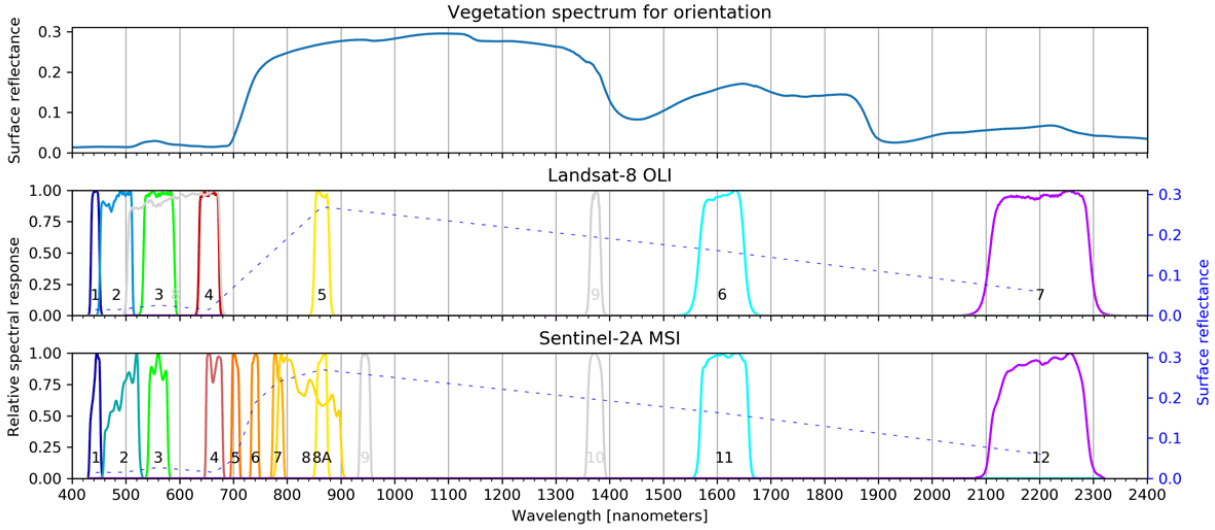


Figure 1. Spectral response functions of Landsat-8 OLI (Barsi et al., 2014) and Sentinel-2A MSI (European Space Agency (ESA), 2017). Bands excluded from regressor training are marked with light gray color. The dashed lines represent spectrally resampled versions of the hyperspectral signature above.

2.2 Real-world Landsat-8 and Sentinel-2 data

In addition to the hyperspectral data used to simulate training data for spectral harmonization and to evaluate sensor-induced spectral differences, we also evaluated the effect of spectral harmonization to real Landsat-8 and Sentinel-2A/B data (section 4.5). For this purpose, we used time series data acquired at two exemplary test sites. The first time series contains 9 Landsat-8 and 21 Sentinel-2

scenes acquired between 01.07.2018 and 31.12.2018 and represents an agricultural area (6 by 4.8 km) at the south coast of Crete, Greece. The second one consists of 20 Landsat-8 and 50 Sentinel-2 scenes acquired throughout 2018 and covers an agricultural area (22.5 by 18 km) in Brandenburg, Germany. The image data of both sensors have been preprocessed with the FORCE software (Frantz, 2019) that allows to apply the same preprocessing algorithm to generate homogenized Level-2 analysis ready data (ARD). The processing workflow converts Level-1 products to Bottom-of-Atmosphere Level-2 products. This includes cloud masking (Frantz et al., 2015, 2016, 2018; Zhu and Woodcock, 2012), co-registration (Rufin et al., in submission; Yan et al., 2016), resolution merging of Sentinel-2 bands, integrated corrections for atmospheric, topographic and adjacency effects (Frantz et al., 2016), nadir BRDF adjustment (Roy et al., 2016b, 2017a, 2017b), as well as re-projection and data cubing (Frantz et al., 2016). To homogenize the spatial resolution, we resampled the Sentinel-2 time series to 30m using an approximated Point Spread Function, i.e., the images were convoluted with a Gaussian lowpass filter with Full Width at Half Maximum of 30m.

3. HARMONIZATION ALGORITHM AND EVALUATION METHODS

3.1 Training of machine learning regressors for estimating spectral information

To harmonize the spectral domains of multiple sensors we evaluate three different machine learning techniques, multivariate linear regression (LR), multivariate quadratic regression (QR) and random forest regression (RFR). Equations for LR and QR are given in Draper and Smith (2014), RFR is explained in detail in Breiman (2001). We also compare these regression techniques with the performance of linear interpolation (LI) as the simplest approach to derive spectral information at a specific wavelength position (e.g., recently used in Griffiths et al., 2019). To be able to investigate the effect of separate, material-dependent estimators for various numbers of clusters for LR and QR, we not only trained one model for each band as, e.g., in Claverie et al. (2018) or in Flood (2017), but trained separate regressors for n subsets of the training data. The subsets were obtained by clustering (details in the following paragraphs). In contrast to previous studies (section 1), we use multi- instead of univariate regression here because univariate (i.e., band-to-band) regression is not expected to work well for unilaterally missing bands. The use of different numbers of clusters allows us to later assess the overall harmonization performance with and without consideration of different surface coverage types. In this study, we varied the number of clusters between 1 and 100 whereas a single cluster in fact is a special case because the sensor-specific spectra are not divided into sub-clusters at all. However, this case corresponds in principal to the harmonization techniques used by previous studies as listed in Table 1. In case of random forest regression (RFR), we do not subdivide the training data into subsets of similar spectra, since the random forest algorithm already models the spectral variability of the input data through a sufficiently high number of trees. In this study, we used 250 trees, which according to our tests represents a good compromise between prediction accuracy and

processing speed. We also tested RFR with 500 trees but could not find any significant improvements in the prediction result.

Figure 2 gives an overview about the workflow we used to train these machine learning regressors for predicting the spectral information at the target wavelength positions and thus to transform a Landsat-8 image into the spectral domain of Sentinel-2. As this approach might be applied to any sensor combination, the following description will use the terms source and target images, i.e., Landsat-8 and Sentinel-2, respectively.

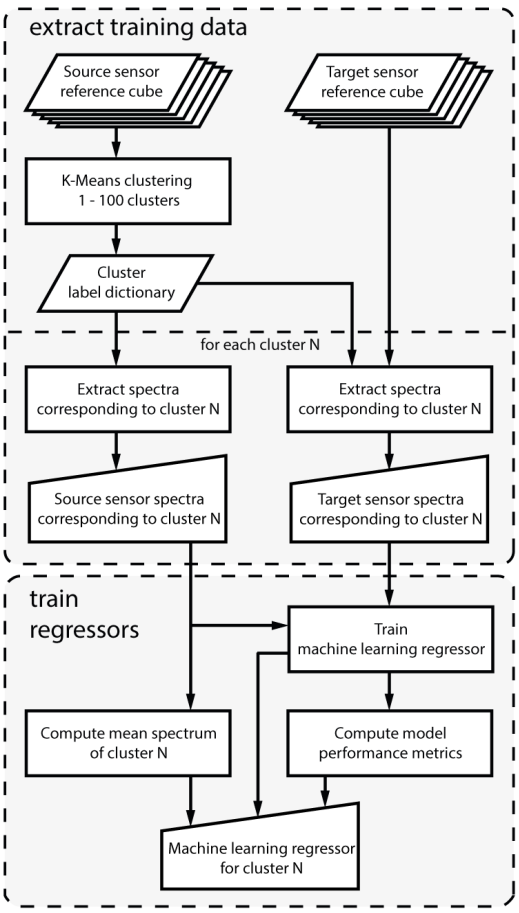


Figure 2. Workflow for training machine learning regressors for spectral harmonization with respect to spectral variations of surface cover types.

For clustering the source sensor reference cube, we used a combination of the K-Means algorithm for computing the cluster centers and the Spectral Angle Mapper (SAM) to assign each pixel to the corresponding spectral cluster. We also tested other clustering algorithms, however, this implementation turned out to be the fastest and therefore the most suitable solution in terms of algorithm operability. Using SAM here instead of the Euclidian Distance (K-Means) reduces the spectral outliers within the selected training spectra of each cluster and therefore increases the clustering quality as we found that spectral similarity was more important than the spectral distance. We iteratively repeated this for different numbers of clusters ranging between 1 and 100. We then used

the derived cluster label dictionary per iteration to extract pairs of spectral clusters from both the source sensor and the target sensor reference cube.

The spectra from each pair of spectral clusters were used to train individual spectral harmonization regressors, i.e., to separately fit the respective machine learning model for each spectral cluster. The computed prediction coefficients were stored along with the mean spectrum of each cluster and various model performance metrics, e.g., root mean square error (RMSE) or model score values.

3.2 Prediction of harmonized spectral information

The workflow we used for performing the spectral harmonization, i.e., for the transformation between different spectral domains, is presented in Figure 3. In contrast to previous studies, it is not limited to bands with similar wavelength in source and target sensor but also includes the prediction of unilaterally missing bands. As input images for the prediction we used multispectral images derived from the three hyperspectral test images (Table 2). To generate these multispectral images, we spectrally resampled the hyperspectral data as in section 3.1. The resulting multispectral images (three images per sensor) were used for two purposes in different combinations: (1) as source images for spectral harmonization and (2) as reference image for the target sensor to quantify the harmonization error.

For the prediction considering a specific number of spectral clusters, our goal was to ensure that we only use regressors trained with the most similar training data on each pixel. This allows an individual spectral transformation for different land cover types. As a measure for spectral similarity we utilized the spectral angle which is insensitive to illumination and albedo effects to find spectral signatures with similar shapes. Tests showed, however, that it is even better to consider several regressors per pixel, all of which are very similar to each other. We achieved good results with five regressors. However, using more may further improve the prediction performance but may also cause a much higher computational load. Consequently, here we first identified the five most appropriate machine learning regressors for each pixel of the source sensor multispectral image. We computed the spectral angle between each pixel spectrum of the source image and all mean spectra associated to the cluster regressors and incorporated only those regressors with the smallest spectral angles into the prediction of the target sensor spectrum. The final prediction result PR was then computed as the weighted average of the five selected prediction results PR_i with the highest weight corresponding to the smallest spectral angle:

$$PR = \frac{\sum_{i=1}^{i=5} w_i \times PR_i}{\sum_{i=1}^{i=5} w_i} \quad (1)$$

The weight of the selected prediction results w_i is defined as follows, where SA_i represents the spectral angle between the pixel spectrum and the regressor mean spectrum and SA_{min} and SA_{max} represent the minimum and maximum spectral angles calculated throughout the image:

$$w_i = 1 - \frac{SA_i - SA_{min}}{SA_{max} - SA_{min}} \quad (2)$$

To avoid using inappropriate regressors for spectra insufficiently represented in our training data, we added a maximum spectral angle of four degrees. Above this threshold we used global transformation coefficients as a fallback (as, e.g., used in Claverie et al., 2018). This procedure of using a weighted average of prediction results instead of using a single regressor per pixel helped us to effectively reduce artificial spectral edges in the predicted image and maximized the prediction performance. The threshold of four degrees spectral angle was determined iteratively by visual inspection of the predicted target sensor images along with corresponding error maps. A lower threshold (e.g., one degree) enlarges the area where global transformation coefficients are applied, which in turn can lead to larger prediction errors, as the global coefficients may not reflect the full spectral variability of the source image. A higher threshold value does not force the mean spectrum of the used regressor and the source sensor spectrum to be spectrally similar which also leads to poor prediction results.

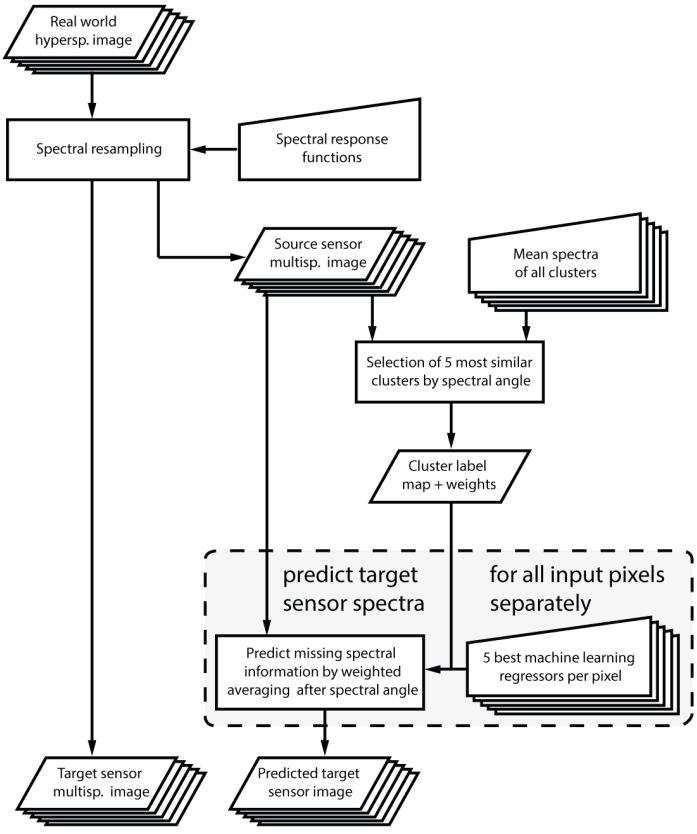


Figure 3. Workflow for prediction of spectrally harmonized images and generation of corresponding target sensor/reference images.

3.3 Methods used for harmonization performance evaluation

Based on the predicted target sensor images, we conducted evaluations on the spectral harmonization performance for the different techniques and with varying numbers of spectral clusters. Additionally, we assessed the effect on typical remote sensing applications as a practical use case by evaluating subsequently derived data products. Details are given in the following:

1. We quantified spectral deviations between the predicted spectral bands and the corresponding multispectral target sensor images as reference (generated by spectral convolution from the hyperspectral input data) by computing root mean square errors.
2. We analyzed the spatial distribution of harmonization errors by computing difference images and examining differing spectral signatures at prominent image positions.
3. We evaluated the effect of spectral harmonization on selected subsequently computed vegetation indices, namely the NDVI (Normalized Difference Vegetation Index; Tucker, 1979), the EVI (Enhanced Vegetation Index; Huete et al., 2002; Liu and Huete, 1995) and the REIP (Red Edge Inflection Point; linear four-point interpolation approach after Clevers et al., 2002 and Guyot and Baret, 1988).
4. We performed an exemplary multispectral random forest classification on test dataset 2 (Table 2) and compared the classification accuracy of native Sentinel-2 data with the accuracy based on predicted data originating from Landsat-8. For each classification scenario we trained one classifier incorporating 250 trees, always with the same spectral endmembers used as training data to achieve comparable results. These were derived by clustering the hyperspectral test dataset 2 into five main classes using a K-Means approach followed by a spectral convolution to Sentinel-2A and Landsat-8. We randomly selected 500 spectral signatures per cluster to adequately model the spectral variability of each endmember. We computed confusion matrices between the classification results and analyzed if an improvement of classification accuracy is detectable after spectral harmonization.
5. We evaluated the effect of spectral harmonization on real Landsat-8/Sentinel-2 data by computing difference images and investigating BOA reflectance deviations in time for different harmonization techniques.

4. RESULTS AND DISCUSSION

4.1 Spectral performance evaluation of different harmonization approaches

4.1.1 Harmonization performance using global prediction coefficients

We directly compared the harmonization performance without spectral sub-clustering between multivariate linear regression (LR), quadratic regression (QR) and random forest regression (RFR) with 250 decision trees. Because of its simplicity and the low computational effort, we also added linear interpolation (LI) here.

Figure 4 shows the spectral deviation for LR, QR, RFR and LI between Sentinel-2A reference images and artificial Sentinel-2A data generated from Landsat-8 (harmonization results). The deviation is quantified as an overall RMSE per spectral band in percent BOA reflectance, averaged over all test image datasets used in this study (Table 2). High harmonization performance is indicated by an RMSE close to zero. For comparison, we also added the average RMSE values between Landsat-8 and Sentinel-2 without any harmonization (only for spectrally overlapping bands). Note, that these spectral differences represent the sensor-induced component, i.e., solely originate from unequal spectral responses of the individual bands. Therefore, they cannot be directly compared with deviations found in previous studies. Chastain et al. (2019) reported TOA reflectance differences between 1.3% and 3.1% and Flood (2017) detected them to be between 0.5% and 2.1% for BOA reflectance (homologous bands only, without harmonization). Both studies only incorporated image pairs with a maximum time difference of one day. However, as these studies were based on real-world data, sensor-induced differences cannot be completely separated from additional biases due to varying illumination or observation geometries or changing atmospheric conditions.

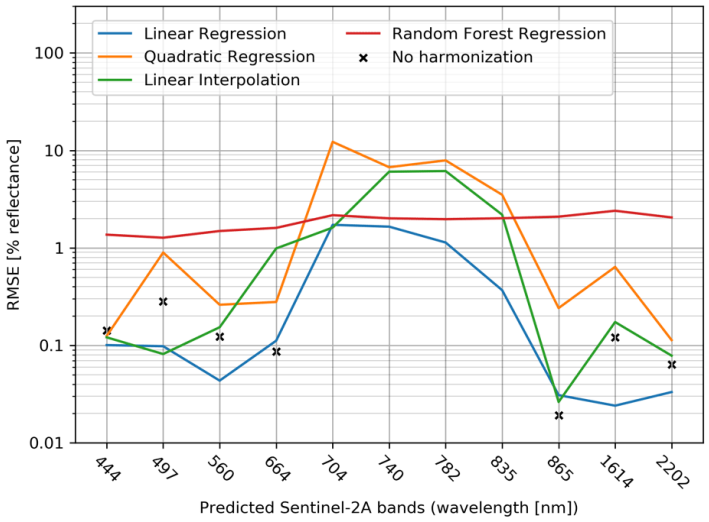


Figure 4. Band-wise reflectance deviation between Sentinel-2A reference sensor bands and artificial Sentinel-2A data as predicted by LR, QR, LI and RFR harmonization from Landsat-8. Crosses represent deviations without any harmonization for similar bands in both sensors. Note that the y-axis is drawn with logarithmic scale.

We found that, generally the harmonization performance is clearly dependent on the spectral wavelength of the target sensor band. LR outperforms QR, RFR and LI for nearly all target sensor bands. The largest errors appear in the red edge spectral region (700–750 nm) and the first near infrared bands (below 850 nm), i.e., at wavelengths where the source sensor (Landsat-8) features no spectral information similar to the Sentinel-2 bands to be predicted. Within this wavelength region LI causes errors up to 6.1% reflectance and QR even exceeds 12.2%. The latter is because QR tends to produce spectral artefacts (strong over- or underestimations) if input spectra differ too much from the

regressor training data. In this regard, LR offers more outlier robustness whereas QR tends to overfitting. However, also the harmonization errors of LR increase by a factor of 15 (1.7% reflectance) in this spectral region compared with the remaining spectral bands (below 0.12% reflectance). For the spectrally overlapping bands in both sensors, LR increases the inter-sensor similarity compared with the similarity without any spectral harmonization (black crosses in Figure 4) for most bands. This has also been reported by Flood (2014, 2017). However, there might not be an accuracy improvement if the gray value difference between not harmonized bands is smaller than the error caused by spatially variable prediction performance. This applies to Sentinel-2´s 664 nm and 865 nm bands. Nevertheless, we note that LR achieves very good prediction performances for these bands (0.11% and 0.03% reflectance RMSE). RFR causes errors around 2% reflectance in the red edge spectral region and, in contrast to the other techniques, it does not show significantly higher performance for bands where source and target sensor have similar wavelengths. Additionally, the harmonized output images showed clearly visible artefacts and edges at positions where the input image contained spatially nearly homogenous pixel values (not shown here). This means, that RFR could not fully model the spectral complexity of our test datasets which might be an issue of overfitting. However, we could not improve that with different regressor configurations.

LI and RFR can only be applied to the whole dataset at once. So in the following, we evaluate the sub-clustering approach for LR and QR only.

4.1.2 Effect of spectral sub-clustering to harmonization performance

To assess the effect of separate transformation functions for different spectral clusters (sub-clustering), we analyzed the deviation between the harmonized image results and their corresponding reference images with a varying number of spectral clusters between 1 and 100. Figure 5 shows the results for LR and QR with Landsat-8 as source sensor and Sentinel-2A as target sensor. The deviation is again quantified per band as RMSE in percent reflectance, averaged over all test images.

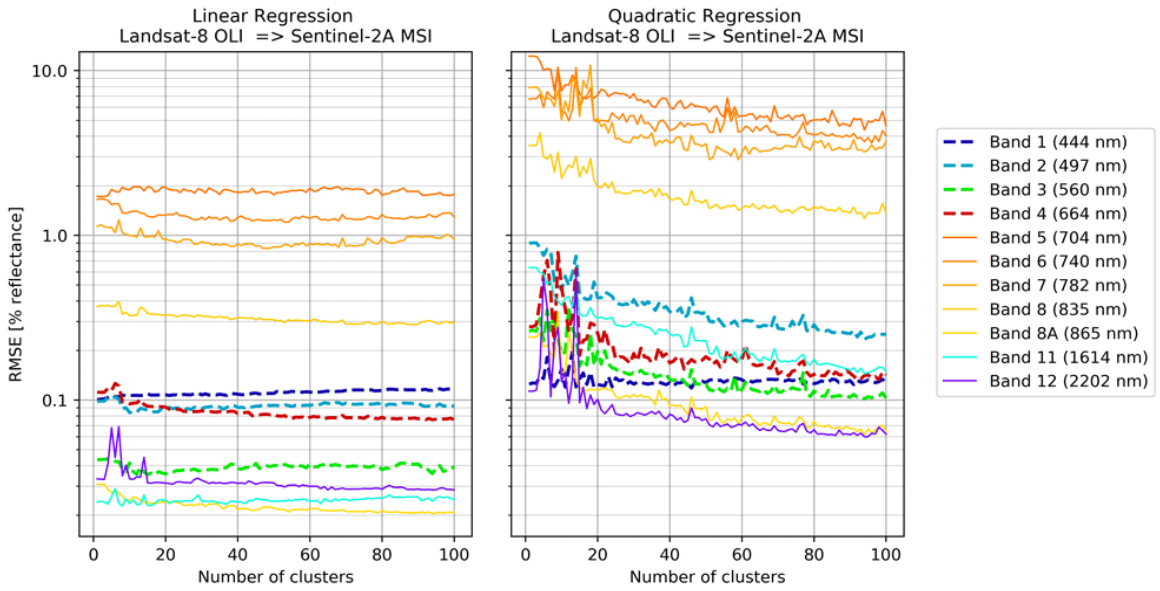


Figure 5. Band-wise reflectance deviation between Sentinel-2A reference sensor bands and artificial Sentinel-2A data as predicted by LR and QR harmonization from Landsat-8. Deviation is plotted as RMSE values (% reflectance) against different numbers of spectral sub-clusters. The bands in the visible spectral region are represented by dashed lines whereas solid lines represent all remaining bands. Note that the y-axis is drawn with logarithmic scale.

For both, LR and QR, the predicted bands at 704, 740, 782 and 835 nm show the lowest harmonization performance, independently from the number of spectral clusters. As mentioned before, this is due to the missing spectral information of Landsat-8 in the red edge and near infrared region below 850 nm. All other bands perform much better with RMSE values 10 times lower and better.

However, in case of LR, we found that the RMSE clearly decreases for most of the predicted bands if more and more spectral clusters are involved in the prediction. We observed this trend up to about 50 clusters above which the RMSE slowly increases again for the predicted 740 and 782 nm bands (lighter orange lines in Figure 5, left). We attribute this rebound to the size of the training library in our study (reference cubes as described in section 2), which in some cases cannot provide enough training signatures for some land cover types at a large number of spectral clusters. The regressors for those underrepresented clusters then perform worse and lead to a re-increasing average RMSE as shown above. Nevertheless, this is not an issue until up to 50 clusters. For example, for the 740 nm band we observed an RMSE-decrease from initially 1.7% reflectance without sub-clustering (also visible in Figure 4) to 1.2% reflectance. This means that we could reduce the prediction uncertainty by around 30% (average for all test datasets). Against the background of a mean overall reflectance of 18.8% in the 740 nm band, this corresponds to an error reduction from 9.0% to 6.4%. In our test data, the 704 nm band did not show this decreasing RMSE with more clusters but rather kept stable with some small variations. The RMSE values of some bands also show slightly increasing trends below 50 clusters, which we also attribute to the decreasing number of training spectra per regressor with more and more clusters used. Additionally, changing cluster center positions in the spectral feature space

might also negatively affect the harmonization performance of certain materials in the test images for these bands. However, we point out that this only applies to some of those bands with a direct equivalent in the source and target sensor where RMSE values generally do not exceed 0.12% reflectance. So, the accuracy of the harmonization is already very high there. Regarding the variation of harmonization errors among our test images (Figure 12, left, in the supplements), we observed similar trends for all of them with standard deviations between 0.9% and 0.4% reflectance for the 704 nm and 740 nm bands and below 0.2% for the remaining bands. This equals around 30% to 50% of the RMSE values mentioned above suggesting a slightly varying harmonization performance from test image to test image. This is due to different surface material compositions of the test images which more or less match the mean spectra of the used regressors. The slightly increasing RMSE of, e.g., the 560 nm band is also visible there as it occurs in only one test image which consequently causes an increasing standard deviation among all test images.

For QR, we observed decreasing RMSE values for all predicted bands (Figure 5, right) which implies an increasing spectral harmonization performance if more material-dependent regressors are incorporated. However, compared with LR, QR showed much larger harmonization errors and could not outperform LR, not even if a large number of spectral clusters was used. For example, the 740 nm band showed reflectance errors of 6.7% without sub-clustering, which could be reduced to 4.0% reflectance with 100 clusters. Nevertheless, compared with the mean band reflectance, we consider this amount of errors as critical for subsequent analysis. Similar to LR, we also observed standard deviations between 30% and 50% of the RMSE values per band (Figure 12, right, in the supplements).

Summarizing the results, the best harmonization performance can be achieved using LR and a number of clusters between 20 and 50. For that reason, we focus on LR harmonization in our subsequent evaluations to spectrally transform Landsat-8 data to Sentinel-2.

4.2 Spatial distribution of harmonization errors

In addition to our quantitative analysis of spectral harmonization performance, we also analyzed how the deviation between the harmonized data and the reference data varies in space. Figure 6 visualizes this deviation between Sentinel-2A (reference) and Landsat-8 (1) using linear spectral interpolation/LI, (2) harmonized using LR but without sub-clustering and (3) harmonized using LR with 5, 15 and 50 spectral sub-clusters. Note, that we provide an animated version of this figure for 1-100 spectral sub-clusters in the supplements. The figure shows the results for the predicted Sentinel-2A band 3 at 560 nm and band 6 at 740 nm, i.e., for the green band with very similar wavelength at both sensors (561 and 560 nm) and for a band within the red edge region which is not covered by Landsat-8. The image (a part of test dataset 2, see Table 2) contains forest/bushland, urban/sealed areas, bare soil and water (see true color composite in the upper left of the figure). The classification maps in left column indicate the most appropriate spectral cluster for each input image pixel, i.e., demonstrate which pixel has been transformed by which regressor. White areas represent input pixels, where no suitable

regressor was available in our training data (spectral angle above the four degrees threshold, as explained in section 3.2). These pixels were harmonized using global LR coefficients.

Linear interpolation (Figure 6, top row) causes the largest deviations in the red edge due to the strongly increasing reflection of vegetation and the missing coverage of Landsat-8 in this spectral region (absolute spectral differences of up to 17.25% reflectance, 7.54% RMSE). This is critical for any remote sensing application incorporating the red edge and will lead to large errors in subsequent analyses. The predicted green band does not exceed 0.41% reflectance error (0.18% RMSE) for LI due to the similar spectral response of both sensors in that spectral region (see Figure 1). It is evident that the deviations are highly dependent on the surface cover type (Figure 6, first row, column two and three). Densely vegetated areas have large deviations due to the steep red edge slope of vegetation spectra whereas sealed surfaces as well as water show small deviations due to much smaller spectral gradients. We note that using the spectrally closest Landsat bands (without any harmonization) instead of linearly interpolated bands would most likely not be an option in practice for many users as this is expected to cause even larger spectral differences depending on the surface coverage.

If LR harmonization without spectral sub-clustering is performed (Figure 6, second row), the deviations are reduced to <0.2% and 4.16% reflectance in maximum for band 3 and 6 (0.07% and 1.52% RMSE). However, deviations are not uniform after spectral harmonization. There are still areas with higher or lower deviations. We attribute this to the global, band-wise transformation function as used in Claverie et al. (2018) that cannot sufficiently consider all spectral individualities of different surface coverages within the image (as reported by, e.g., Flood, 2014, 2017).

In case of the 740 nm band, this effect is reduced by using separate transformation functions for spectral clusters (sub-clustering approach, Figure 6, rows three to five). The 560 nm band also benefits from additional sub-clustering, mainly in densely vegetated image areas. But we note that the prediction accuracy with global coefficients is already very high due to the similarity of the Landsat-8 signal. In contrast, the deviations in the red edge band can be reduced to 1.50%, 1.06% and 0.94% reflectance RMSE with 5, 15 and 50 spectral clusters. This clearly demonstrates the benefit of considering spectral sub-clusters for the harmonization. In case of the example image in Figure 6, it reduces the deviations especially for the vegetated areas and ensures a significantly more homogeneous distribution of prediction errors. It has to be noted that in contrast to the RFR harmonization approach (not shown here) no brightness artefacts originating from the clustered harmonization approach are visible in the harmonized output image. In addition, the output spectra fit well to the Sentinel-2A reference spectra, especially for the 50 clusters case (Figure 6, lower right). Only the water spectrum (blue line) keeps unchanged in the sub-clustering cases because no material-specific regressor with a spectral angle smaller than four degrees (section 3.2) was found. This also applies to a few other dark surfaces which were instead harmonized with global regressor coefficients. From the spectral angles point of view, dark areas are extremely variable and therefore often exceed the threshold. Additionally, it is possible that we could not fully capture their entire spectral variability

in our reference cubes. Further developments are needed to improve the harmonization performance for dark surfaces in future.

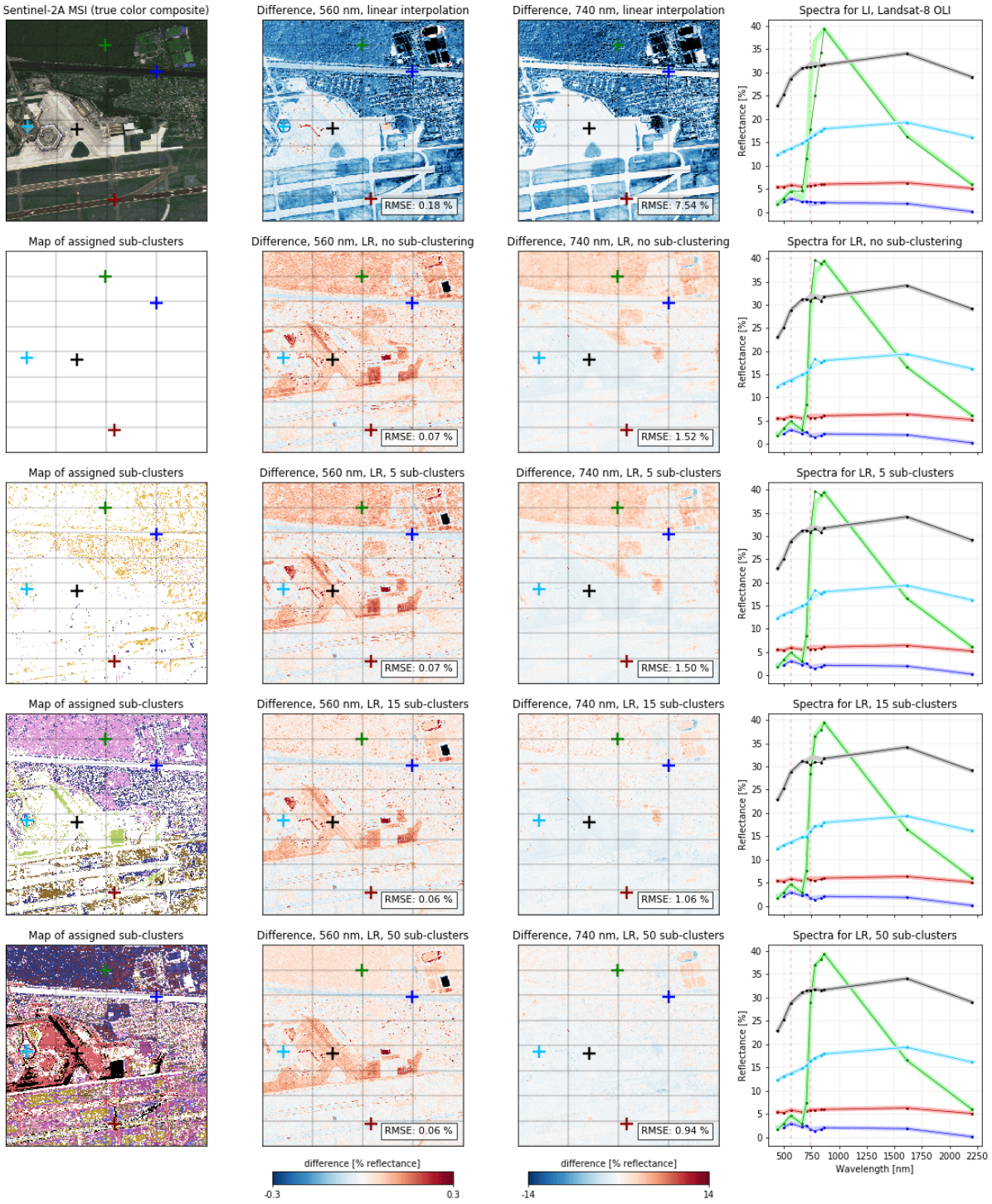


Figure 6. Reflectance deviation between the Sentinel-2A reference image and artificial Sentinel-2A data as predicted from Landsat-8 for the Sentinel-2A bands 3 (560 nm) and 6 (740 nm). Rows contain different harmonization scenarios: using linear spectral interpolation (upper row), with LR-harmonized Landsat-8 data, no sub-clustering (second row) and with LR-harmonized Landsat-8 data using 5, 15 and 50 sub-clusters. Note, that the difference images have separate color ranges per column due to strongly differing value ranges. The spectra in column 4 are extracted at the positions indicated in the images to the left. Brighter, thicker and darker, thinner

lines indicate reference and predicted spectra. The classification maps in column 1 indicate the regressor assignment whereas white areas represent the global LR regressor (used as fallback). Note, that we provide an animated version of this figure for 1-100 spectral sub-clusters in the supplements.

4.3 Effect of spectral harmonization on selected spectral indices

To assess the benefit of spectral harmonization to vegetation indices, we generated NDVI, EVI and REIP products based on (1) simulated Sentinel-2A and Landsat-8 data without harmonization, (2) Landsat-8 data, LR-harmonized to Sentinel-2A without sub-clustering and (3) Landsat-8 data, LR-harmonized to Sentinel-2A using 50 sub-clusters.

Figure 7 compares the results for the NDVI for the same image subset as used in section 4.2; there is no visible difference for the three states of harmonization (second column). However, the difference images (third column) reveal NDVI differences in the range of -0.04 to +0.04 which corresponds to 4% absolute NDVI difference. The scatter plots (Figure 7, fourth column) allow to numerically assess the effect of spectral harmonization with regard to NDVI differences.

In case no harmonization is applied to the input data, the NDVI is computed from Sentinel-2A's 664 nm and 835 nm bands and Landsat-8's 655 nm and 864 nm bands. This implies that the Near Infrared (NIR) center wavelength position differs by 29 nm. Although Sentinel-2 also features a narrower band at 865 nm, we note that this band is acquired by a different focal plane with 20 m spatial resolution and therefore does not match the 10 m red band without resampling. To compute a 10 m NDVI product, the 835 nm band is frequently used in the literature (e.g., Belgiu and Csillik, 2018; Gao et al., 2017; Van der Meer et al., 2014). The largest NDVI differences appear in sparsely vegetated pixels (Figure 7, third column), i.e., at the light green areas in the second column of Figure 7. This is due to the individual spectral response functions of the red and NIR band of both sensors and mainly caused by a higher slope of the vegetation signatures between 835 nm (NIR band of Sentinel-2) and 865 nm (NIR band of Landsat-8) of sparse compared with dense vegetation. The same pattern was also observed in previous studies between Landsat-7 and Landsat-8 (Roy et al., 2016a; Xu and Guo, 2014) where Landsat-8 NDVI values were slightly higher for sparse vegetation and nearly equal to Landsat-7 for dense vegetation.

The difference images proof, that the NDVI differences can be reduced if the input images are spectrally harmonized using LR (lighter red areas of sparse vegetation in Figure 7, third column), with an even stronger reduction when using 50 clusters (Figure 7, bottom row). If LR is applied without sub-clustering (Figure 7, center row), the NDVI differences even increase for dense vegetation because these spectra are only insufficiently represented by the global transformation coefficients. The black areas in the difference image indicate NDVI deviations outside of the data range and correspond to water that cannot be accurately predicted using global transformation coefficients, too. If multiple clusters are used within LR harmonization, the NDVI differences decrease to <0.01 (except for the water areas for which we could only apply global coefficients). This is corroborated by the scatter plots for NDVI values above 0.3 (vegetation pixels). These points converge to the red line of zero-

difference if LR harmonization is applied and even more if multiple spectral clusters are used. The RMSE of these pixels decreases from 0.0094 to 0.0082 to 0.0038 NDVI values which equals 1.6%, 1.4% and 0.6% of NDVI values referred to the NDVI value range that is realistic for vegetation (0.3–0.9). This suggests that even if the center wavelength position of the NDVI input bands does not change much through harmonization, an LR harmonization is useful to eliminate multi-sensor inconsistencies within NDVI products, especially if material specific transformation coefficients are used (sub-clustering approach). Consequently, it allows to compute 10 m NDVI products from Sentinel-2 without large inconsistencies compared with Landsat-8.

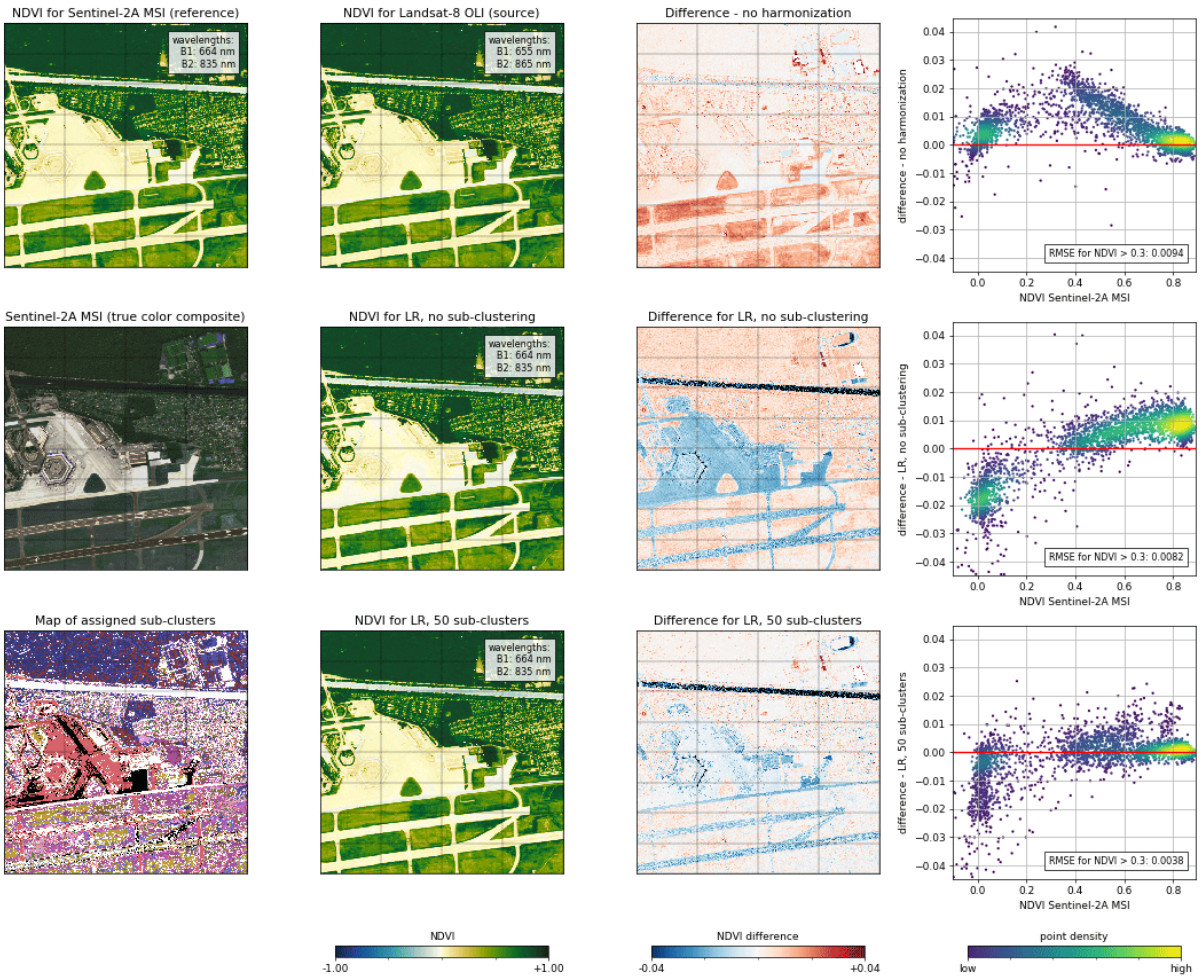


Figure 7. Comparison of NDVI values computed from Sentinel-2A and (harmonized) Landsat-8 data. Rows contain different harmonization scenarios for Landsat-8: without spectral harmonization (upper row), with LR-harmonized Landsat-8 data, no sub-clustering (center row) and with LR-harmonized Landsat-8 data, 50 sub-clusters. The true color Sentinel-2 image and the NDVI product computed from it are shown for reference in column 1.

To validate that the above findings also apply to large scale satellite data with various land cover types, we computed NDVI and EVI products from a full HyMap image (test dataset 1; 1.8 by 26.5 km, 3.5 m spatial resolution). Additionally, we generated REIP products from Landsat-8 data, LR harmonized to Sentinel-2A. To compute the REIP, at least two bands in the red edge spectral region

are needed. Therefore, this is only made possible for Landsat-8 by harmonization to Sentinel-2A. So the goal was to assess how the REIP based on Landsat-8 differs from a native Sentinel-2A product if the red edge bands are artificially generated by LR spectral harmonization.

Figure 8 shows the index value deviations limited to vegetation pixels only (identified by NDVI >0.3). The results for the NDVI confirm the above findings of clearly decreasing deviations if LR harmonization including spectral clustering is applied. For global LR coefficients, we observed an increasing RMSE from 0.0073 (no harmonization) to 0.0087 (LR harmonization without clustering) which is due to too large spectral differences between dense vegetation spectra and the global mean spectrum of the training data (as mentioned above). However, using 50 material-dependent regressors for prediction reduced these deviations to 0.0028, i.e., to only 38% of the initial error without harmonization. On the one hand this demonstrates the limitations of a global LR regressor and on the other hand shows that the sub-clustering approach can effectively improve the inter-sensor consistency of harmonized NDVI products.

The EVI uses the same spectral bands like the NDVI plus an additional blue band. It shows a similar deviation like the NDVI suggesting an advantage of spectral harmonization but mainly if material-dependent regressors are used. (95% of the initial error with LR harmonization applied; 43% for LR, under the use of 50 clusters).

Regarding the REIP, which is expressed as a wavelength position, the scatterplots show that it can be computed from LR harmonized Landsat-8 data with a mean accuracy of 4.25 nm. Spectral sub-clustering improves that to 3.12 nm accuracy (RMSE). A REIP-based estimation of biophysical plant parameters has been studied several times in the past (Clevers et al., 2002; Gitelson et al., 1996; Herrmann et al., 2011; Jago et al., 1999; Main et al., 2011) but is dependent on the crop type. Hermann et al. (2011) estimated the relation between REIP and the leaf area index (LAI) for a mix of wheat and potato crops and according to their studies, a REIP at 717.9 nm (as the mean in this study with 50 spectral clusters involved) would correspond to an LAI of around 1.1. An uncertainty of 3.12 nm would lead to an LAI uncertainty of around 1.8.

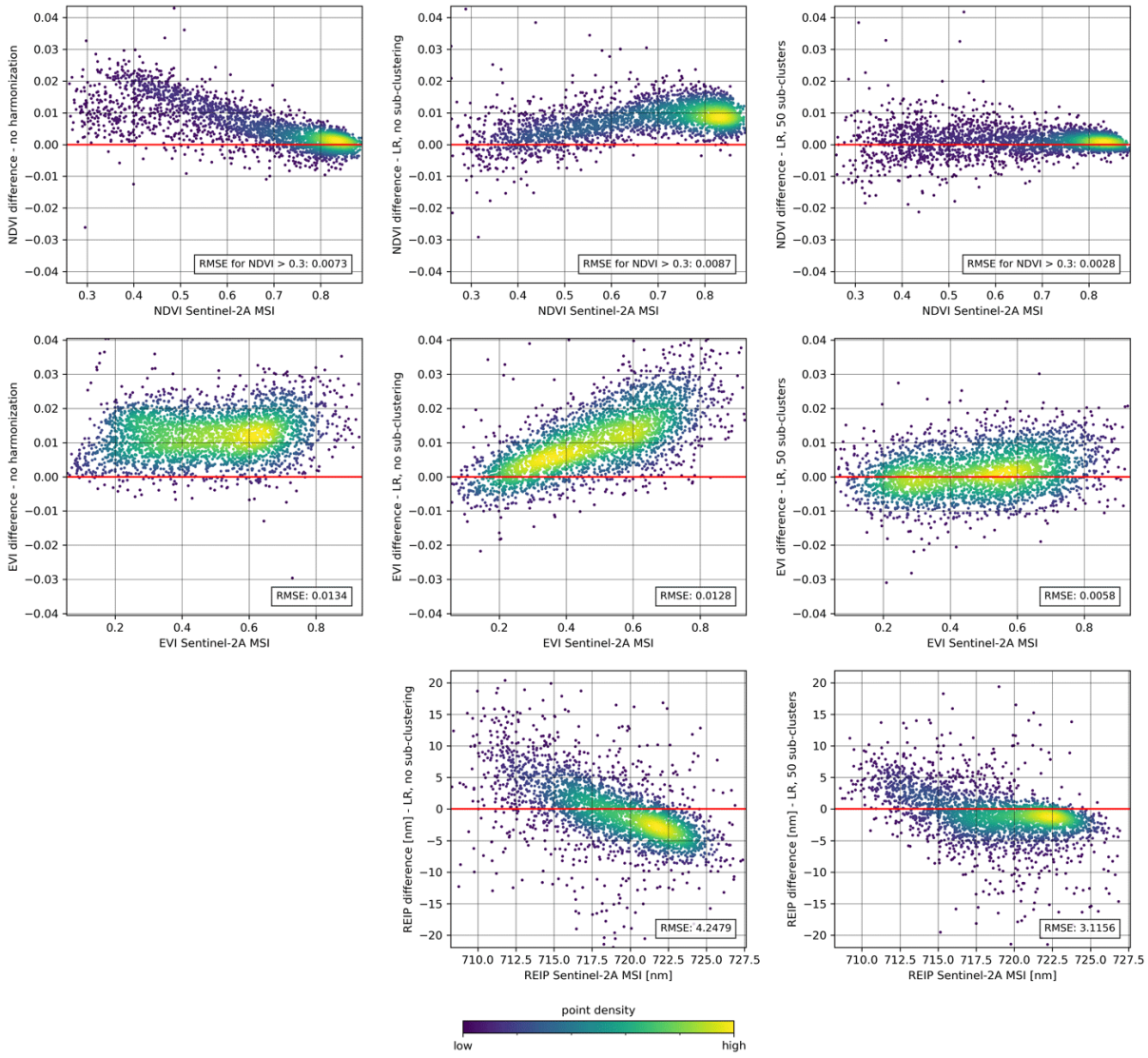


Figure 8. Effect of spectral harmonization to NDVI, EVI and REIP generated from Sentinel-2A and (harmonized) Landsat-8 data (vegetation pixels only). Columns contain different harmonization scenarios for Landsat-8: without spectral harmonization (first column), with LR-harmonized Landsat-8 data, no sub-clustering (second column) and with LR-harmonized Landsat-8 data, 50 sub-clusters (third column).

Further improvements could even be achieved by adding a single additional band in the red edge spectral region for upcoming Landsat sensors. To reinforce this statement, we used the same methodology as described in section 3.1 to simulate RapidEye-5 data and to train corresponding regressors, as this sensor provides an additional band at 713 nm. We estimated Sentinel-2 from RapidEye-5 data as we did for the Landsat-8 data in this study, and applied the same LR harmonization approach (see section 3.2). We achieved a REIP estimation accuracy of 0.96 nm without sub-clustering and 0.62 nm with 50 spectral clusters incorporated (Figure 13 in the supplements). This corresponds to an LAI uncertainty of around 0.55 and 0.35. It clearly shows that even a single additional red edge band can highly improve the estimation accuracy of biophysical parameters such as LAI, and we strongly suggest this spectral region to be covered in upcoming

satellite missions. Nevertheless, even without a red edge band we were able to estimate LAI with an uncertainty of 1.8 from Landsat-8 data, harmonized to Sentinel-2A.

4.4 Effect of spectral harmonization on the multi-sensor consistency of land cover classifications

Finally, we analyzed the effect of spectral harmonization to the consistency of multi-sensor land cover classifications. We performed an exemplary multispectral random forest classification on Sentinel-2A and Landsat-8 data using test dataset 2 (Table 2) and compared the classification accuracy (see section 3.3 for details on the classification). We show again three scenarios: (1) without any harmonization of Landsat-8, (2) with LR harmonized data but without sub-clustering and (3) with LR harmonization incorporating 50 spectral sub-clusters. The classification result directly computed from Sentinel-2A data is taken as the reference, since a classification based on perfectly harmonized Landsat-8 data should ideally lead to the same classification map. Differing spectral information due to harmonization uncertainties should instead cause deviations in the classification maps.

Figure 9 shows the confusion matrices between the classification results and illustrates the similarity between classifying a native Sentinel-2A dataset and the three Landsat-8 harmonization results. Without spectral harmonization the classification result of Landsat-8 is at least 82.9% consistent with Sentinel-2A. One class achieves an accuracy of 100% but the mean value is 92.3%. However, if an LR harmonization is performed to transform the Landsat-8 data to the spectral domain of Sentinel-2A, the accuracies can be improved by 7.5% to at least 90.4% or to 96.3% average. Using spectral sub-clustering with 50 spectral clusters can even further improve that to a minimum and mean consistency of 94.2% and 97.3%, respectively.

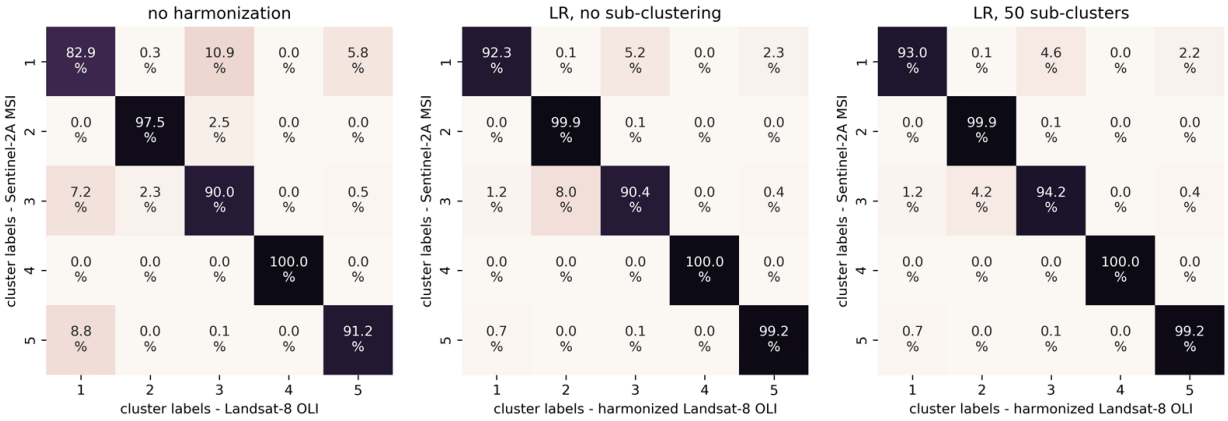


Figure 9. Classification confusion matrices to compare the similarity of a random forest classification (5 spectral classes) between Sentinel-2A and (left) Landsat-8 without harmonization, (center) Landsat-8, harmonized to Sentinel-2 using LR, without sub-clustering and (right) Landsat-8, harmonized to Sentinel-2 using LR, with 50 sub-clusters.

4.5 Application of the proposed harmonization to real Landsat-8 and Sentinel-2 data

After evaluating the effect of the proposed harmonization algorithm on simulated data, we tested the approach on real Landsat-8 and Sentinel-2 data acquired at two test sites in 2018: (1) at the south coast of Crete, Greece and (2) in Brandenburg, Germany, in the northwest of Berlin. However, we point out that numerous additional effects occur in real data which can cause differences between dates but are unrelated to the sensor-induced reflectance differences investigated in this study. Even small differences in the acquisition time, the observation or illumination geometry lead to visible spectral differences due to material-dependent BRDF properties or changing atmospheric states (variable aerosol or water vapor contents, cloud and cloud shadow positions, etc.) (Roy et al., 2016a).

As mentioned in section 2, the data were preprocessed by the FORCE software (Frantz, 2019) which runs several processing steps to minimize the above described spectral differences. However, a perfect correction is not possible (Claverie et al., 2015; Doxani et al., 2018; Ju et al., 2012; Zhang et al., 2018), i.e., remaining spectral differences are intermingled by the purely sensor-induced deviations that we aim to correct in this study.

Similar to Figure 6, Figure 10 compares these spectral differences depending on several spectral harmonization techniques for band 6 of Sentinel-2 (740 nm), as predicted from Landsat-8. The image pair of the Crete site has been acquired with a 1-day separation in August 2018 and the one of the Brandenburg site represents a same-day acquisition from early May 2018. In both cases, there is a difference in daytime of only about 15 min. This minimizes spectral differences due to surface coverage or atmospheric dynamics or changes in the illumination geometry. Small-scale elevations are usually below 30 m mitigating topographic effects. The Sentinel-2 image, spatially resampled to Landsat-8, is taken as the reference.

At the Crete test site (Figure 10, upper row), surface materials mainly consist of bare soils, shrub and rangelands, olive trees and urban areas. In case of LI, the 740 nm reflectance values of Sentinel-2 are clearly underestimated from Landsat-8 with an RMSE of 3.64% reflectance. This is because LI cannot model the shape of the spectral signature in the red edge (section 4.2). Using LR without spectral sub-clustering (as used by Claverie et al., 2018), the deviations can be reduced to an RMSE of 1.36% reflectance. If, additionally, 50 material-specific regressors are used (proposed sub-clustering approach), they can be further reduced to 1.09% reflectance. Moreover, the difference image becomes smoother suggesting that deviations due to material-specific individualities could be reduced. Unlike our evaluations with simulated data (section 4.2), the remaining spatial variability of harmonization errors is not only due to different prediction performances depending on the surface material, but also due to uncertainties in preprocessing (as described above).

An even stronger effect of the proposed spectral harmonization approach can be seen at the Brandenburg site (Figure 10, bottom row). Here, a much larger image fraction is covered by vegetation (various agricultural areas, forests, few water bodies and urban areas). This causes larger

deviations in case of LI (RMSE of 8.65% reflectance), particularly at densely vegetated areas. But also if LR with a global regressor is used for harmonization, there are a lot of areas in the predicted image that show larger harmonization errors (darker red areas in the difference image in Figure 10; RMSE of 3.63% reflectance). These areas correspond to surface materials that are spectrally more different from the global average spectrum resulting in larger errors with the global harmonization coefficients. Using 50 material-specific regressors instead of a single set of harmonization coefficients reduces the RMSE to 1.8% reflectance and produces a much more homogenous difference image with errors below 1.5% reflectance in the most part of the image. However, some patches with higher spectral differences remain. With a view to the true color image, it becomes evident that these patches occur mainly at the bright cropland areas which we interpret as rape fields, as rape reaches its full bloom at the beginning of May in Germany. Further investigations revealed that our threshold of a maximum spectral angle (section 3.2) was exceeded there during harmonization so that the global regressor was used as fallback (equivalent errors as in the previous global regressor case). This is because we had no spectra of blooming rape in our training database. However, this effect can be avoided by a training database that also takes greater account of phenological changes over the year.

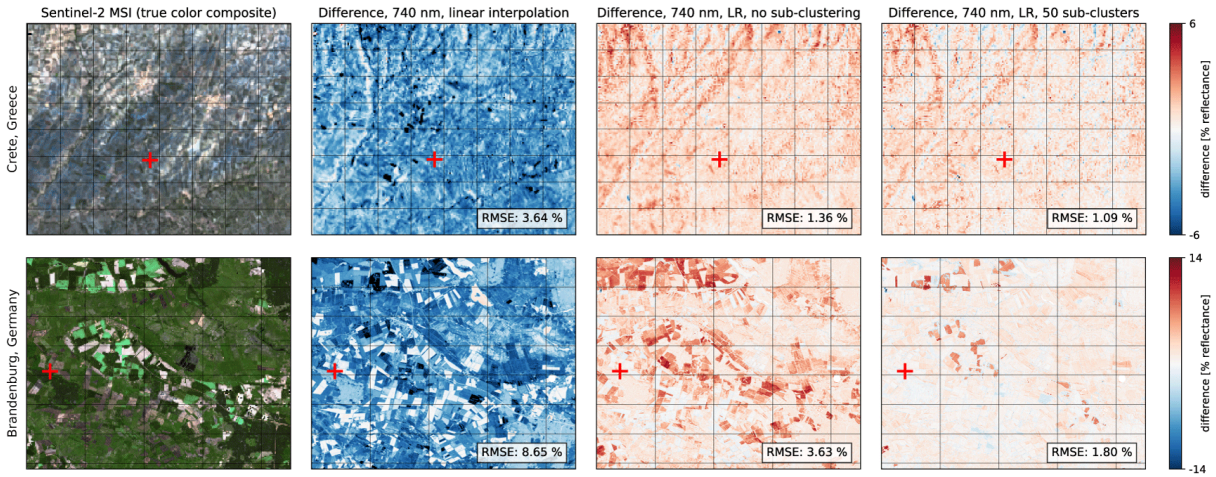


Figure 10. Reflectance deviation in space between real and artificial Sentinel-2 data as predicted from Landsat-8 for the Sentinel-2 band 6 at 740 nm. The upper row represents the test site in Crete, Greece and the bottom row the one in Brandenburg, Germany. Deviations are shown for different harmonization scenarios: using linear spectral interpolation, with LR-harmonized Landsat-8 data (no sub-clustering) and with LR-harmonized Landsat-8 data using 50 sub-clusters. The red crosses indicate the positions of the BOA reflectance time series shown in Figure 11.

To analyze the effect of spectral harmonization to inter-sensor spectral differences in time, Figure 11 visualizes the BOA reflectance values of exemplary pixels out of the above shown Landsat-8/Sentinel-2 time series of both test sites. The positions of the pixels are indicated in Figure 10. In case of LI, the above mentioned underestimation is also visible in Figure 11 (left) as a clear offset between the reference reflectance of Sentinel-2 and the values predicted from Landsat-8. This offset can be reduced by using LR instead of LI (Figure 11, center column). Nevertheless, some predicted values still differ

systematically from Sentinel-2. This can be further improved by using a material-dependent LR regressor instead of global transformation coefficients (Figure 11, right).

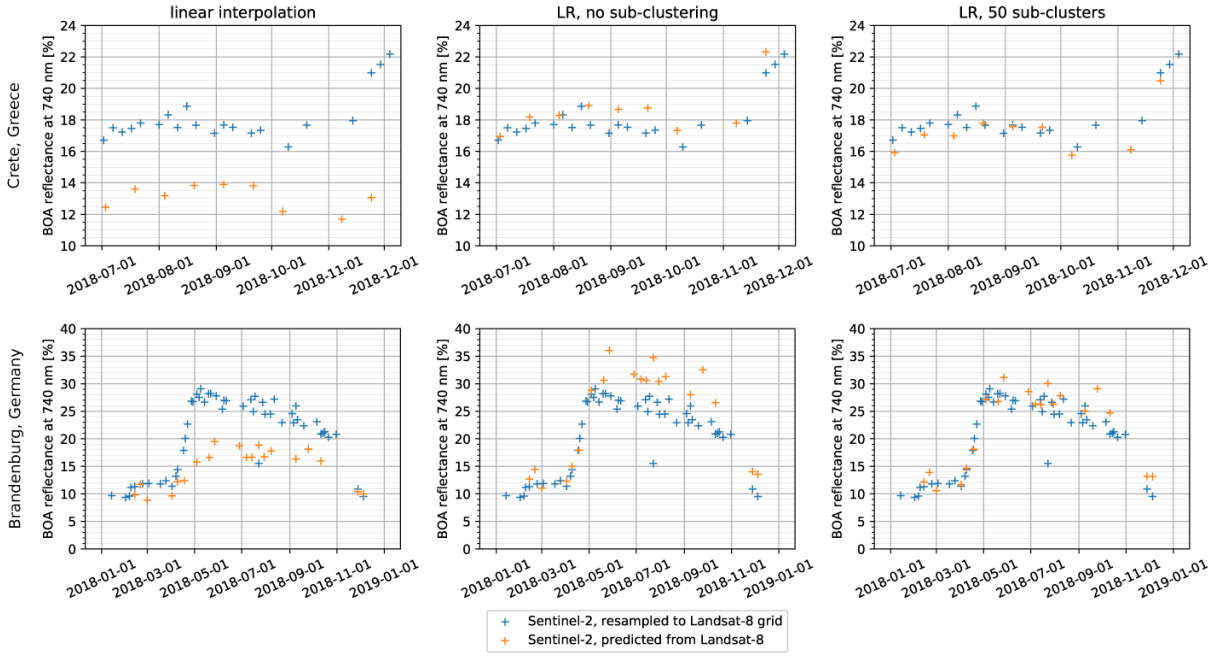


Figure 11. Reflectance deviation in time between real and artificial Sentinel-2 data as predicted from Landsat-8 for the Sentinel-2 band 6 at 740 nm. The upper row contains data from the test site in Crete, Greece and the bottom row from the one in Brandenburg, Germany. Deviations are shown for different harmonization scenarios: using linear spectral interpolation (left), with LR-harmonized Landsat-8 data, no sub-clustering (center) and with LR-harmonized Landsat-8 data using 50 sub-clusters (right).

4.6 Error assessment and limitations of the study

As limitations of this study, we note some concerns regarding the nine hyperspectral datasets we used for the simulation of our multispectral data basis. These hyperspectral data have been radiometrically unified prior to spectral convolution as far as possible (section 2). However, different algorithms have been used for atmospheric correction in advance of this study and we were not able to correct for BRDF effects. This is because a BRDF correction would have to be applied to the hyperspectral data prior to the simulation of the multispectral data basis and would require detailed information about land cover types (Collings et al., 2010), which was not available for this study. Moreover, we did not equalize spatial resolutions to avoid spectral degradation. This means that we might have included purer spectra in our training data than usually recorded by Landsat-8 and Sentinel-2A under the same acquisition and illumination conditions. Nevertheless, we note that these differences between the underlying hyperspectral datasets have been incorporated into our multispectral input database (reference cube) for each individual sensor. Therefore, they don´t affect the inter-sensor deviations we derive here. On the contrary, they increase the spectral variability of our data basis and therefore contribute to a higher robustness of the machine learning techniques we used.

As a technical drawback of using multivariate instead of univariate (band-to-band) regressors, we would like to mention, that in case of multivariate regressors (as used in our study) all spectral bands of a source image have to be read to generate a harmonization result in the target sensor spectral domain. This may increase the processing time for disk access intensive workflows. In contrast, univariate regression, e.g., only requires two bands to be read from each image to obtain an NDVI result based on harmonized spectral information. We therefore also do not recommend to harmonize different spectral bands with multiple numbers of clusters as this would severely slow down the harmonization process. Apart from that, separate regressors per band might also cause spikes in the predicted spectra and consequently affect downstream products.

Regarding LR and QR based on separate harmonization functions per spectral cluster, we assign the most appropriate machine learning regressors for each pixel of the source sensor image by computing the spectral angle as similarity measure between input image spectra and the mean spectra of the training data associated with each regressor (section 3.2). However, the spectral angle is mainly sensitive to a similar shape of spectral signatures but not to brightness differences. This might cause some wrong regressor assignments and hence increase deviations at certain positions in the predicted image. Generally, this effect is reduced by our procedure of computing weighted averages of multiple regressor results. However, further research should be conducted to study the effect of more reliable spectral similarity metrics that are sensitive to both, spectral shape and brightness differences and also consider spatial adjacencies.

Additionally, our evaluations revealed that the proposed material-specific regressors could rarely be applied to dark surfaces such as water bodies or asphalt since the spectral angle threshold was exceeded there. We see two reasons for this. First, we had only insufficient training data to adequately train material specific regressors that cover the entire spectral variability of these surfaces in our independent test data. Second, due to the low signal, dark surfaces have an extreme spectral variability in terms of the spectral angle. Even small reflectance variations may lead to spectral angles exceeding our threshold of four degrees. For these materials a more suitable process must therefore be developed in future. The hard fallback we used in case of an exceeded threshold may also cause gray value edges in the prediction result if the global harmonization coefficients are unsuitable to the corresponding surface material. This was especially apparent in our Brandenburg real data use case (section 4.5) where we had no suitable material-specific classifier for some rape fields which were then spectrally transformed with larger errors. Future versions of the proposed harmonization approach must therefore improve the use of this fallback to further reduce the variation of harmonization errors in space.

Finally, it must be noted that the harmonization coefficients derived in this study may not be suitable for all geographic positions of the world due to different land cover (Chastain et al., 2019; Flood, 2014, 2017; Mandanici and Bitelli, 2016). Nevertheless, we intentionally chose training images acquired under different climatic conditions and with a high land cover variability to achieve as much model robustness as possible. Since our algorithm tolerates spectral deviations between input spectra

and the assigned regressors by handling larger deviations with lower weights, users can still expect good harmonization results for surfaces that were not explicitly included in our training data. Apart from that, users are able to constrain this spectral deviation with a threshold above which global transformation coefficients (as, e.g., in Claverie et al., 2018) are used as a fallback (trained to model entire bands instead of material-specific clusters). This keeps the proposed harmonization algorithm applicable to "incompatible" surfaces, even though it will not achieve the harmonization accuracy of other surface coverages there.

5. SUMMARY AND CONCLUSION

This paper presents a thorough investigation of the benefit and limitations of spectral band harmonization for optical multispectral satellite imagery at the example of simulated Landsat-8 and Sentinel-2 data. It particularly addresses the prediction of the Sentinel-2 spectral information at those wavelengths that are not covered by Landsat-8, i.e., the red edge and the spectral region up to 850 nm. Different harmonization techniques were used, such as multivariate linear regression (LR), multivariate quadratic regression (QR), random forest regression (RFR) and linear interpolation (LI). Additionally, we developed a new prediction approach to improve the harmonization accuracy by respecting material-dependent spectral individualities. It incorporates separate transformation functions for different spectral clusters of the input dataset and was examined for LR and QR. We evaluated the harmonization performance by computing root mean square errors to simulated reference data and quantified NDVI and EVI differences with and without spectral harmonization to demonstrate the effect for typical remote sensing applications. Furthermore, we investigated if an improvement of classification accuracy is detectable if unequal spectral sensor characteristics are unified. In addition to these evaluations based on simulated data, we also applied the harmonization techniques to real Landsat-8 and Sentinel-2 data to investigate the effectiveness for realistic remote sensing applications.

Our results show that the quality of the harmonized image data highly depends on the spectral wavelength or more precisely on the similarity of the spectral characteristics of source and target sensor. Hence, the prediction accuracy of spectral bands with center wavelength positions close to existing bands is generally much higher than at spectral positions where the source sensor lacks spectral information. Besides that, it is dependent on the shape of the spectral signatures, i.e., high reflectance gradients like within the red edge spectral region cause larger errors.

With global regressor coefficients, i.e., without the proposed sub-clustering approach, LR generally outperformed QR, RFR and LI and could also improve the inter-sensor consistency compared with non-harmonized data. LI seems to be only useful for already spectrally overlapping bands but is highly prone to errors in case of larger "spectral gaps", i.e., if the source sensor does not provide spectral information at the targeted wavelength. Using RFR (250 decision trees in this study) the harmonized output data contained clearly visible artefacts and could not outperform the harmonization results of

LR. Regarding LR and QR incorporating global harmonization coefficients (as, e.g., in Claverie et al., 2018), our analysis revealed that deviations highly depend on the surface coverage and vary from pixel to pixel. This confirms the presumptions of Flood (2014). The deviations can be reduced by using multiple transformation functions for different spectral clusters. To be able to analyze the effect of different numbers of material-specific regressors in detail we gradually increased the number of clusters from 1 to 100. We found out that there is an improvement of harmonization quality for both, LR and QR, if more and more spectral clusters are involved. In case of LR, we achieved the maximum harmonization performance with 50 spectral clusters and the spectral angle as technique to assign each input pixel to a corresponding spectral cluster and hence to a specific LR regressor. Although we mainly attribute the re-increasing harmonization errors at >50 clusters in our study to our limited total number of training spectra (section 4.1.2), we note that a larger number of clusters also increases the computational load during harmonization. Nevertheless, with 50 material-specific regressors we could reduce the inter-sensor deviations by about 30% in the red edge spectral region (averaged for all our test datasets, simulated from hyperspectral data) and achieved a much more homogenous distribution of remaining errors. QR produces much larger deviations compared with LR. Therefore, we limited our further evaluations to LR harmonization.

Our analyses regarding NDVI products from simulated data revealed that the purely sensor-induced deviations between NDVI index values of Landsat-8 and Sentinel-2A are in the range of 4% with reference to the whole NDVI value range. Spectral harmonization with a global regressor could not reduce them, because material-dependent variations are not considered. However, by using 50 separate LR regressors instead of a single one, we could reduce the NDVI deviations to 38% of the initial error without harmonization (vegetated image pixels only). For EVI, we observed a slight reduction of deviations to 95% using a single LR regressor for homogenization which could be further reduced to 43% by using 50 spectral clusters. Based on Landsat-8 simulations, LR-harmonized to Sentinel-2A, we were able to compute the red edge inflection point (REIP) with an accuracy of 3.1 nm. The REIP cannot be computed from native Landsat-8 data due to missing red edge spectral bands. With regard to land cover classifications we observed an improvement of multi-sensor consistency of the classification maps from 92.3% to 96.3% with LR harmonized Landsat-8 data and to even 97.3% under the use of 50 spectral clusters (mean consistencies).

When applied to real Landsat-8 and Sentinel-2 data, the reduction of purely sensor-induced deviations is difficult to quantify because inter-sensor deviations are not only due to spectral response differences but also caused by unequal observation and illumination geometries or atmospheric conditions which can never be perfectly corrected in the pre-processing of the data. However, with regard to the 740 nm red edge band of Sentinel-2 predicted from Landsat-8, we observed a reduction of inter-sensor BOA reflectance deviations in both of our test sites. At the Crete site, they decreased from 3.64% using LI harmonization to 1.36% using LR with global transformation coefficients to 1.09% using LR with 50 material-dependent regressors (sub-clustering approach). At the Brandenburg site, the effect was even

stronger: deviations reduced from 8.65% to 3.63% to 1.8% RMSE reflectance. This reduction was also evident when comparing the harmonized inter-sensor reflectance over time.

In summary, the study demonstrates on simulated Landsat-8 and Sentinel-2 data that spectral harmonization is useful to improve the multi-sensor consistency of remote sensing data, especially if multiple transformation functions are incorporated. Whether it is also worthwhile in real world applications depends on the individual radiometric accuracy requirements of the application and on the question if unilaterally missing bands are incorporated. However, we showed that spectral harmonization directly increases the inter-sensor similarity of reflectance values and consequently the reliability of all subsequent data products. We suggest linear regression as a robust and easy to implement technique to gain unified spectral characteristics of actual multi-sensor data. However, with global regressor coefficients, LR has the drawback of remaining material-dependent deviations that in some cases exceeded 10% reflectance with our test data simulations. Our proposed algorithm accounting for these spectral individualities does exist as a Python package (Scheffler 2020; this study is based on version 0.5.1) which will be published as open source code at the following URL soon: https://gitext.gfz-potsdam.de/geomultisens/spechomo. It is self-contained, generic and works out-of-the-box to harmonize Landsat-8 and Sentinel-2A as well as other sensors such as Landsat-5 TM, Landsat-7 ETM+, Sentinel-2B, RapidEye-5, SPOT-4 and SPOT-5 using the included material-specific machine learning regressors. Moreover, it features an algorithm to train additional regressors (as presented section 3.1) for custom sensor combinations based on user provided hyperspectral training data. This also allows users to further improve harmonization results according to their specific study areas. The proposed approach is considered to be incorporated in the next release of FORCE, a toolset for generating and analyzing Landsat and Sentinel-2 Analysis Ready Data (Frantz, 2019). Future work may further investigate the effect of spectral and spatial harmonization on real world remote sensing applications.

ACKNOWLEDGEMENTS

This work was funded by the German Federal Ministry of Education and Research (BMBF 01IS14010A-C/GeoMultiSens) within the framework of GeoMultiSens – Scalable Multi-Sensor Analysis of Remote Sensing Data. We thank CSIRO (Commonwealth Scientific and Industrial Research Organization), DLR (German Aerospace Centre), GFZ (German Research Centre for Geosciences), BELSPO (Belgian Science Policy Office), LMU (Ludwig Maximilian University of Munich) and VITO (Flemish Institute for Technological Research) for providing the hyperspectral input data of this study that we used to simulate multispectral data for Sentinel-2A, Landsat-8 and RapidEye-5. Moreover, we thank USGS for providing Landsat-8 L1T data and ESA for providing Sentinel-2 data. We are grateful to the anonymous reviewers and the editor for their constructive and insightful comments that helped to improve the quality of the manuscript. We also thank Patrick Hostert (HU Berlin/IRI THESys) for commenting on the manuscript.

REFERENCES

Barsi, J., Lee, K., Kvaran, G., Markham, B., Pedelty, J., Barsi, J.A., Lee, K., Kvaran, G., Markham, B.L., Pedelty, J.A., 2014. The Spectral Response of the Landsat-8 Operational Land Imager. Remote Sens. 6, 10232–10251. https://doi.org/10.3390/rs61010232

Belgiu, M., Csillik, O., 2018. Sentinel-2 cropland mapping using pixel-based and object-based time-weighted dynamic time warping analysis. Remote Sens. Environ. 204, 509–523. https://doi.org/10.1016/J.RSE.2017.10.005

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32. https://doi.org/https://doi.org/10.1023/A:1010933404324

Brown, M.E., Pinzón, J.E., Didan, K., Morisette, J.T., Tucker, C.J., 2006. Evaluation of the consistency of long-term NDVI time series derived from AVHRR,SPOT-vegetation, SeaWiFS, MODIS, and Landsat ETM+ sensors. IEEE Trans. Geosci. Remote Sens. 44, 1787–1793. https://doi.org/10.1109/TGRS.2005.860205

Chastain, R., Housman, I., Goldstein, J., Finco, M., Tenneson, K., 2019. Empirical cross sensor comparison of Sentinel-2A and 2B MSI, Landsat-8 OLI, and Landsat-7 ETM+ top of atmosphere spectral characteristics over the conterminous United States. Remote Sens. Environ. 221, 274–285. https://doi.org/10.1016/j.rse.2018.11.012

Claverie, M., Ju, J., Masek, J.G., Dungan, J.L., Vermote, E.F., Roger, J.-C., Skakun, S. V., Justice, C., 2018. The Harmonized Landsat and Sentinel-2 surface reflectance data set. Remote Sens. Environ. 219, 145–161. https://doi.org/10.1016/J.RSE.2018.09.002

Claverie, M., Masek, J.G., Ju, J., Dungan, J.L., 2017. Harmonized Landsat-8 Sentinel-2 (HLS) Product User's Guide. https://doi.org/10.13140/RG.2.2.33017.26725

Claverie, M., Vermote, E.F., Franch, B., Masek, J.G., 2015. Evaluation of the Landsat-5 TM and Landsat-7 ETM + surface reflectance products. Remote Sens. Environ. 169, 390–403. https://doi.org/10.1016/J.RSE.2015.08.030

Clevers, J.G.P.W., De Jong, S.M., Epema, G.F., Van Der Meer, F.D., Bakker, W.H., Skidmore, A.K., Scholte, K.H., 2002. Derivation of the red edge index using the MERIS standard band setting. Int. J. Remote Sens. 23, 3169–3184. https://doi.org/10.1080/01431160110104647

Collings, S., Caccetta, P., Campbell, N., Wu, X., 2010. Techniques for BRDF correction of hyperspectral mosaics. IEEE Trans. Geosci. Remote Sens. 48, 3733–3746. https://doi.org/10.1109/TGRS.2010.2048574

Doxani, G., Vermote, E., Roger, J.-C., Gascon, F., Adriaensen, S., Frantz, D., Hagolle, O., Hollstein, A., Kirches, G., Li, F., Louis, J., Mangin, A., Pahlevan, N., Pflug, B., Vanhellemont, Q., Doxani, G., Vermote, E., Roger, J.-C., Gascon, F., Adriaensen, S., Frantz, D., Hagolle, O., Hollstein, A., Kirches, G., Li, F., Louis, J., Mangin, A., Pahlevan, N., Pflug, B., Vanhellemont, Q., 2018.

Atmospheric Correction Inter-Comparison Exercise. Remote Sens. 10, 352. https://doi.org/10.3390/rs10020352

Draper, N.R., Smith, H., 2014. Applied Regression Analysis., 3rd ed. Wiley India. https://doi.org/10.1002/9781118625590

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., Bargellini, P., 2012. Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. Remote Sens. Environ. 120, 25–36. https://doi.org/10.1016/J.RSE.2011.11.026

European Space Agency (ESA), 2017. Sentinel-2 Spectral Response Functions (S2-SRF) [WWW Document]. URL https://earth.esa.int/web/sentinel/user-guides/sentinel-2-msi/document-library/-/asset_publisher/Wk0TKajiISaR/content/sentinel-2a-spectral-responses (accessed 4.12.18).

Flood, N., 2014. Continuity of Reflectance Data between Landsat-7 ETM+ and Landsat-8 OLI, for Both Top-of-Atmosphere and Surface Reflectance: A Study in the Australian Landscape. Remote Sens. 6, 7952–7970. https://doi.org/10.3390/rs6097952

Flood, N., 2017. Comparing Sentinel-2A and Landsat 7 and 8 Using Surface Reflectance over Australia. Remote Sens. 9, 659. https://doi.org/10.3390/rs9070659

Foerster, S., Brosinsky, A., Wilczok, C., Bauer, M., 2015. Isábena 2011 - An EnMAP Preparatory Flight Campaign (Datasets). https://doi.org/http://doi.org/10.5880/enmap.2015.007

Frantz, D., 2019. FORCE—Landsat + Sentinel-2 Analysis Ready Data and Beyond. Remote Sens. 11, 1124. https://doi.org/10.3390/rs11091124

Frantz, D., Haß, E., Uhl, A., Stoffels, J., Hill, J., 2018. Improvement of the Fmask algorithm for Sentinel-2 images: Separating clouds from bright surfaces based on parallax effects. Remote Sens. Environ. 215, 471–481. https://doi.org/10.1016/j.rse.2018.04.046

Frantz, D., Röder, A., Stellmes, M., Hill, J., 2016. An operational radiometric Landsat preprocessing framework for large-area time series applications. IEEE Trans. Geosci. Remote Sens. 54, 3928–3943. https://doi.org/10.1109/TGRS.2016.2530856

Frantz, D., Röder, A., Udelhoven, T., Schmidt, M., 2015. Enhancing the Detectability of Clouds and Their Shadows in Multitemporal Dryland Landsat Imagery: Extending Fmask. IEEE Geosci. Remote Sens. Lett. 12, 1242–1246. https://doi.org/10.1109/LGRS.2015.2390673

Gallo, K., Ji, L., Reed, B., Eidenshink, J., Dwyer, J., 2005. Multi-platform comparisons of MODIS and AVHRR normalized difference vegetation index data. Remote Sens. Environ. 99, 221–231. https://doi.org/10.1016/j.rse.2005.08.014

Gao, F., Masek, J., Schwaller, M., Hall, F., 2006. On the blending of the Landsat and MODIS surface reflectance: predicting daily Landsat surface reflectance. IEEE Trans. Geosci. Remote Sens. 44, 2207–2218. https://doi.org/10.1109/TGRS.2006.872081

Gao, Q., Zribi, M., Escorihuela, M., Baghdadi, N., Gao, Q., Zribi, M., Escorihuela, M.J., Baghdadi, N., 2017. Synergetic Use of Sentinel-1 and Sentinel-2 Data for Soil Moisture Mapping at 100 m

Resolution. Sensors 17, 1966. https://doi.org/10.3390/s17091966

Gitelson, A.A., Merzlyak, M.N., Lichtenthaler, H.K., 1996. Detection of Red Edge Position and Chlorophyll Content by Reflectance Measurements Near 700 nm. J. Plant Physiol. 148, 501–508. https://doi.org/10.1016/S0176-1617(96)80285-9

Griffiths, P., Nendel, C., Hostert, P., 2019. Intra-annual reflectance composites from Sentinel-2 and Landsat for national-scale crop and land cover mapping. Remote Sens. Environ. 220, 135–151. https://doi.org/10.1016/j.rse.2018.10.031

Guyot, G., Baret, F., 1988. Utilisation de la Haute Resolution Spectrale pour Suivre L'etat des Couverts Vegetaux, in: Guyenne, T.D., Hunt, J.J. (Eds.), 4th International Colloquium on Spectral Signatures of Objects in Remote Sensing. European Space Agency, Aussois, France, pp. 279–286.

Hagolle, O., Huc, M., Villa Pascual, D., Dedieu, G., 2010. A multi-temporal method for cloud detection, applied to FORMOSAT-2, VENμS, LANDSAT and SENTINEL-2 images. Remote Sens. Environ. 114, 1747–1755. https://doi.org/10.1016/j.rse.2010.03.002

Hall, F.G., Strebel, D.E., Nickeson, J.E., Goetz, S.J., 1991. Radiometric Rectification: Toward a Common Radiometric Response Among Multidate, Multisensor Images. Remote Sens. Environ. 35, 11–27.

Hank, T.B., Richter, K., Locherer, M., Frank, T., Mauser, W., 2015. Neusling (Landau a.d. Isar) 2011 - An Agricultural EnMAP Preparatory Flight Campaign Using the APEX Instrument (Datasets). https://doi.org/http://doi.org/10.5880/enmap.2015.003

Herrmann, I., Pimstein, A., Karnieli, A., Cohen, Y., Alchanatis, V., Bonfil, D.J., 2011. LAI assessment of wheat and potato crops by VENμS and Sentinel-2 bands. Remote Sens. Environ. 115, 2141–2151. https://doi.org/10.1016/j.rse.2011.04.018

Hong, G., Zhang, Y., 2008. A comparative study on radiometric normalization using high resolution satellite images. Int. J. Remote Sens. 29, 425–438. https://doi.org/10.1080/01431160601086019

Huete, A., Didan, K., Miura, T., Rodriguez, E.., Gao, X., Ferreira, L.., 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sens. Environ. 83, 195–213. https://doi.org/10.1016/S0034-4257(02)00096-2

Irons, J.R., Dwyer, J.L., Barsi, J.A., 2012. The next Landsat satellite: The Landsat Data Continuity Mission. Remote Sens. Environ. 122, 11–21. https://doi.org/10.1016/J.RSE.2011.08.026

Jago, R.A., Cutler, M.E.., Curran, P.J., 1999. Estimating Canopy Chlorophyll Concentration from Field and Airborne Spectra. Remote Sens. Environ. 68, 217–224. https://doi.org/10.1016/S0034-4257(98)00113-8

Ju, J., Roy, D.P., Vermote, E., Masek, J., Kovalskyy, V., 2012. Continental-scale validation of MODIS-based and LEDAPS Landsat ETM+ atmospheric correction methods. Remote Sens. Environ. 122, 175–184. https://doi.org/10.1016/J.RSE.2011.12.025

Liu, H.Q., Huete, A., 1995. A Feedback Based Modification of the NDVI to Minimize Canopy

Background and Atmospheric Noise. IEEE Trans. Geosci. Remote Sens. 33, 457–465. https://doi.org/10.1109/TGRS.1995.8746027

Lloyd, S.P., 1982. Least Squares Quantization in PCM. IEEE Trans. Inf. Theory 28, 129–137. https://doi.org/10.1109/TIT.1982.1056489

Louis, J., Debaecker, V., Pflug, B., Main-Knorn, M., Bieniarz, J., Mueller-Wilm, U., Cadau, E., Gascon, F., 2016. Sentinel-2 Sen2Cor: L2A processor for users, in: Proceedings of the Living Planet Symposium. Prague, Czech Republic, pp. 9–13.

Main, R., Cho, M.A., Mathieu, R., O'Kennedy, M.M., Ramoelo, A., Koch, S., 2011. An investigation into robust spectral indices for leaf chlorophyll estimation. ISPRS J. Photogramm. Remote Sens. 66, 751–761. https://doi.org/10.1016/J.ISPRSJPRS.2011.08.001

Mandanici, E., Bitelli, G., 2016. Preliminary Comparison of Sentinel-2 and Landsat 8 Imagery for a Combined Use. Remote Sens. 8, 1014. https://doi.org/10.3390/rs8121014

Mishra, N., Helder, D., Barsi, J., Markham, B., 2016. Continuous calibration improvement in solar reflective bands: Landsat 5 through Landsat 8. Remote Sens. Environ. 185, 7–15. https://doi.org/10.1016/j.rse.2016.07.032

Miura, T., Huete, A., Yoshioka, H., 2006. An empirical investigation of cross-sensor relationships of NDVI and red/near-infrared reflectance using EO-1 Hyperion data. Remote Sens. Environ. 100, 223–236. https://doi.org/10.1016/j.rse.2005.10.010

Neumann, C., Weiss, G., Itzerott, S., 2015. Döberitzer Heide 2008/2009 - An EnMAP Preparatory Flight Campaign (Datasets). https://doi.org/http://doi.org/10.5880/enmap.2015.001

Pacifici, F., Longbotham, N., Emery, W.J., 2014. The Importance of Physical Quantities for the Analysis of Multitemporal and Multiangular Optical Very High Spatial Resolution Images. IEEE Trans. Geosci. Remote Sens. 52, 6241–6256. https://doi.org/10.1109/TGRS.2013.2295819

Richter, R., Schläpfer, D., 2019. Atmospheric / Topographic Correction for Satellite Imagery. DLR Report DLR-IB 564-04/2019. German Aerospace Center (DLR). Wessling, Germany.

Richter, R., Schläpfer, D., 2002. Geo-atmospheric processing of airborne imaging spectrometry data. Part 2: Atmospheric/topographic correction. Int. J. Remote Sens. 23, 2631–2649. https://doi.org/10.1080/01431160110115834

Roy, D.P., Kovalskyy, V., Zhang, H.K., Vermote, E.F., Yan, L., Kumar, S.S., Egorov, A., 2016a. Characterization of Landsat-7 to Landsat-8 reflective wavelength and normalized difference vegetation index continuity. Remote Sens. Environ. 185, 57–70. https://doi.org/10.1016/J.RSE.2015.12.024

Roy, D.P., Li, J., Zhang, H.K., Yan, L., Huang, H., Li, Z., 2017a. Examination of Sentinel-2A multi-spectral instrument (MSI) reflectance anisotropy and the suitability of a general method to normalize MSI reflectance to nadir BRDF adjusted reflectance. Remote Sens. Environ. 199, 25–38. https://doi.org/10.1016/J.RSE.2017.06.019

Roy, D.P., Li, Z., Zhang, H.K., Roy, D.P., Li, Z., Zhang, H.K., 2017b. Adjustment of Sentinel-2

Multi-Spectral Instrument (MSI) Red-Edge Band Reflectance to Nadir BRDF Adjusted Reflectance (NBAR) and Quantification of Red-Edge Band BRDF Effects. Remote Sens. 9, 1325. https://doi.org/10.3390/rs9121325

Roy, D.P., Zhang, H.K., Ju, J., Gomez-Dans, J.L., Lewis, P.E., Schaaf, C.B., Sun, Q., Li, J., Huang, H., Kovalskyy, V., 2016b. A general method to normalize Landsat reflectance data to nadir BRDF adjusted reflectance. Remote Sens. Environ. 176, 225–271. https://doi.org/10.1016/j.rse.2016.01.023

Rufin, P., Frantz, D., Yan, L., Hostert, P.. Operational co-registration of the Sentinel-2A/B image archive using multi-temporal Landsat spectral averages. In submission.

Schaepman-Strub, G., Schaepman, M., Painter, T., Dangel, S., Martonchik, J., 2006. Reflectance quantities in optical remote sensing - definitions. Remote Sens. Environ. 103, 27–42. https://doi.org/10.1016/j.rse.2006.03.002

Scheffler, D., 2020. SpecHomo Python package, Version 0.5.1. Zenodo doi:10.5281/zenodo.3678713. https://doi.org/10.5281/zenodo.3678713.

Scheffler, D., Hollstein, A., Diedrich, H., Segl, K., Hostert, P., 2017. AROSICS: An Automated and Robust Open-Source Image Co-Registration Software for Multi-Sensor Satellite Data. Remote Sens. 9, 676. https://doi.org/10.3390/RS9070676

Steven, M.D., Malthus, T.J., Baret, F., Xu, H., Chopping, M.J., 2003. Intercalibration of vegetation indices from different sensor systems. Remote Sens. Environ. 88, 412–422. https://doi.org/10.1016/J.RSE.2003.08.010

Teillet, P.M., 1986. Image correction for radiometric effects in remote sensing. Int. J. Remote Sens. 7, 1637–1651. https://doi.org/10.1080/01431168608948958

Teillet, P.M., Staenz, K., William, D.J., 1997. Effects of spectral, spatial, and radiometric characteristics on remote sensing vegetation indices of forested regions. Remote Sens. Environ. 61, 139–149. https://doi.org/10.1016/S0034-4257(96)00248-9

Tucker, C.J., 1979. Red and photographic infrared linear combinations for monitoring vegetation. Remote Sens. Environ. 8, 127–150. https://doi.org/10.1016/0034-4257(79)90013-0

Useya, J., Chen, S., 2018. Comparative Performance Evaluation of Pixel-Level and Decision-Level Data Fusion of Landsat 8 OLI, Landsat 7 ETM+ and Sentinel-2 MSI for Crop Ensemble Classification. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 11, 4441–4451. https://doi.org/10.1109/JSTARS.2018.2870650

Van der Meer, F.D., Van der Werff, H.M.A., Van Ruitenbeek, F.J.A., 2014. Potential of ESA's Sentinel-2 for geological applications. Remote Sens. Environ. 148, 124–133. https://doi.org/10.1016/j.rse.2014.03.022

Vermote, E., Justice, C., Claverie, M., Franch, B., 2016. Preliminary analysis of the performance of the Landsat 8/OLI land surface reflectance product. Remote Sens. Environ. 185, 46–56. https://doi.org/10.1016/J.RSE.2016.04.008

Wulder, M.A., Loveland, T.R., Roy, D.P., Crawford, C.J., Masek, J.G., Woodcock, C.E., Allen, R.G., Anderson, M.C., Belward, A.S., Cohen, W.B., Dwyer, J., Erb, A., Gao, F., Griffiths, P., Helder, D., Hermosilla, T., Hipple, J.D., Hostert, P., Hughes, M.J., Huntington, J., Johnson, D.M., Kennedy, R., Kilic, A., Li, Z., Lymburner, L., McCorkel, J., Pahlevan, N., Scambos, T.A., Schaaf, C., Schott, J.R., Sheng, Y., Storey, J., Vermote, E., Vogelmann, J., White, J.C., Wynne, R.H., Zhu, Z., 2019. Current status of Landsat program, science, and applications. Remote Sens. Environ. 225, 127–147. https://doi.org/10.1016/j.rse.2019.02.015

Xu, D., Guo, X., 2014. Compare NDVI Extracted from Landsat 8 Imagery with that from Landsat 7 Imagery. Am. J. Remote Sens. 2, 10–14. https://doi.org/10.11648/j.ajrs.20140202.11

Yan, L., Roy, D., Zhang, H., Li, J., Huang, H., 2016. An Automated Approach for Sub-Pixel Registration of Landsat-8 Operational Land Imager (OLI) and Sentinel-2 Multi Spectral Instrument (MSI) Imagery. Remote Sens. 8, 520–543. https://doi.org/10.3390/rs8060520

Zhang, H.K., Roy, D.P., Yan, L., Li, Z., Huang, H., Vermote, E., Skakun, S., Roger, J.-C., 2018. Characterization of Sentinel-2A and Landsat-8 top of atmosphere, surface, and nadir BRDF adjusted reflectance and NDVI differences. Remote Sens. Environ. 215, 482–494. https://doi.org/10.1016/J.RSE.2018.04.031

Zhu, Z., Wang, S., Woodcock, C.E., 2015. Improvement and expansion of the Fmask algorithm: cloud, cloud shadow, and snow detection for Landsats 4–7, 8, and Sentinel 2 images. Remote Sens. Environ. 159, 269–277. https://doi.org/10.1016/j.rse.2014.12.014

Zhu, Z., Woodcock, C.E., 2012. Object-based cloud and cloud shadow detection in Landsat imagery. Remote Sens. Environ. 118, 83–94. https://doi.org/10.1016/j.rse.2011.10.028

SUPPLEMENTS

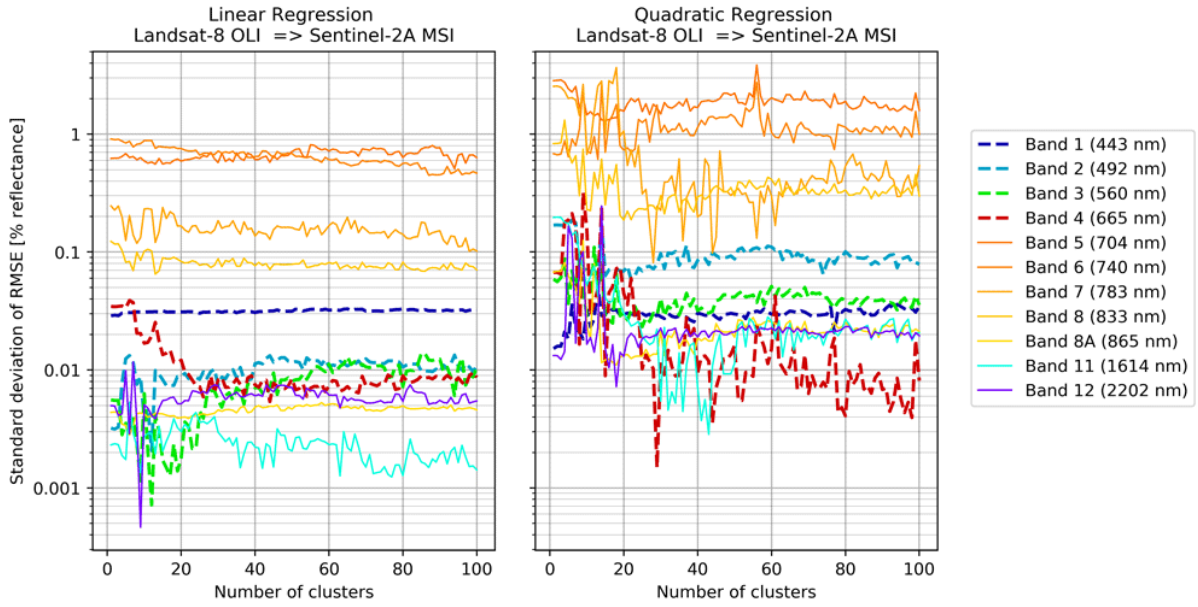


Figure 12. Band-wise variation of harmonization errors among the test datasets used in this study in case of Sentinel-2A data as predicted by LR and QR harmonization from Landsat-8. Variation is plotted as standard deviation of RMSE values.

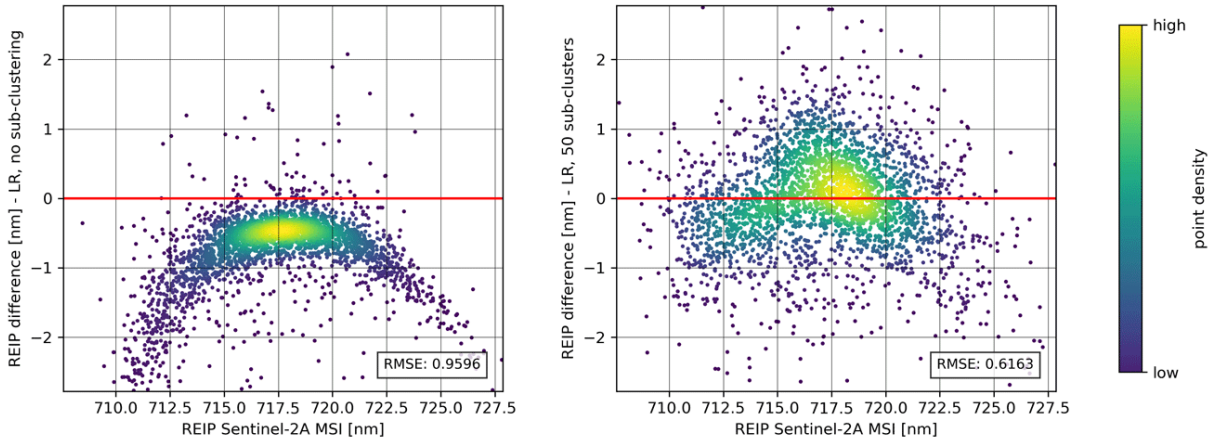
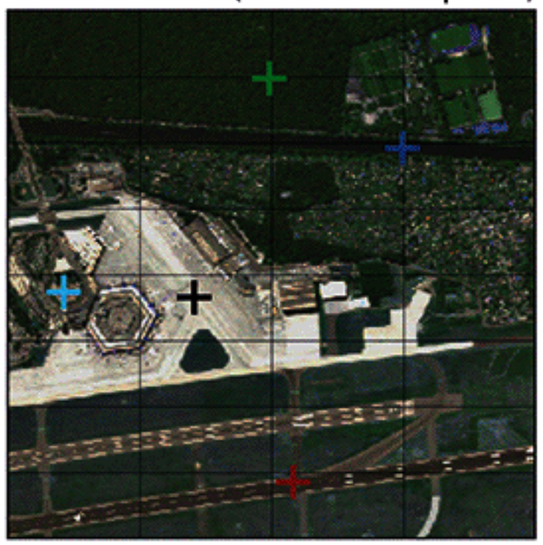
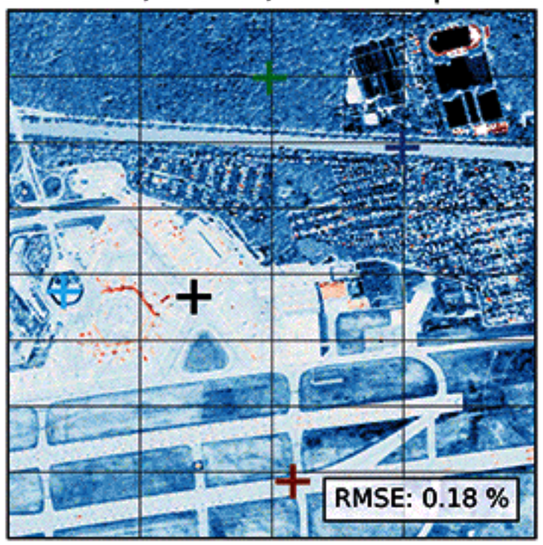


Figure 13. REIP accuracy for vegetation pixels only compared between Sentinel-2A (reference) and RapidEye-5, spectrally harmonized to Sentinel-2A using LR. Left: without sub-clustering; right: with 50 spectral clusters.

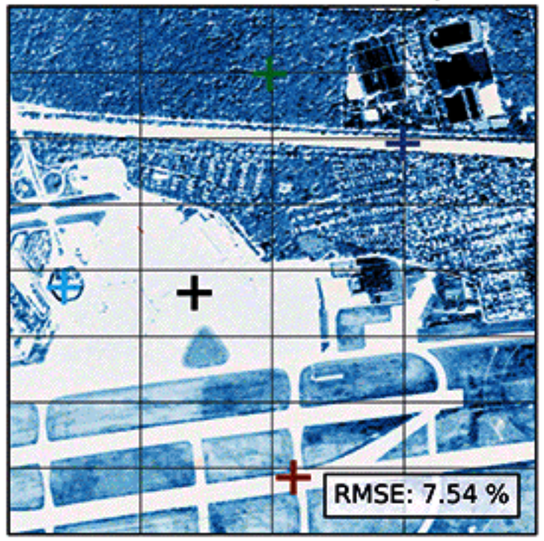
Sentinel-2A MSI (true color composite)



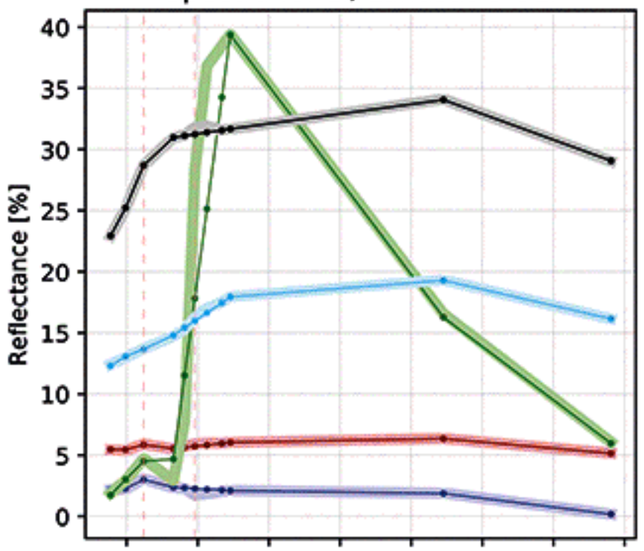
Difference, 560 nm, linear interpolation



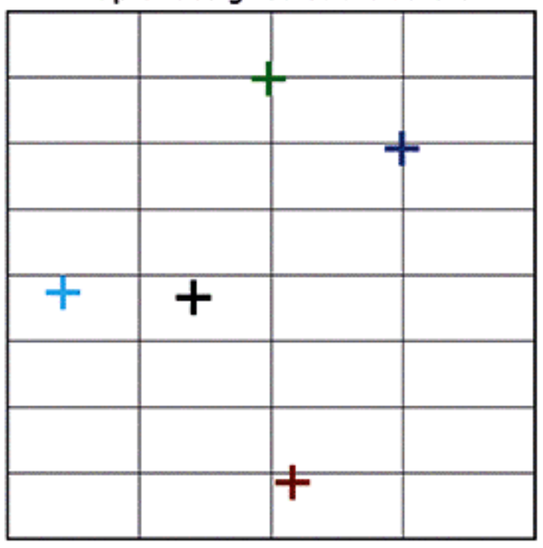
Difference, 740 nm, linear interpolation



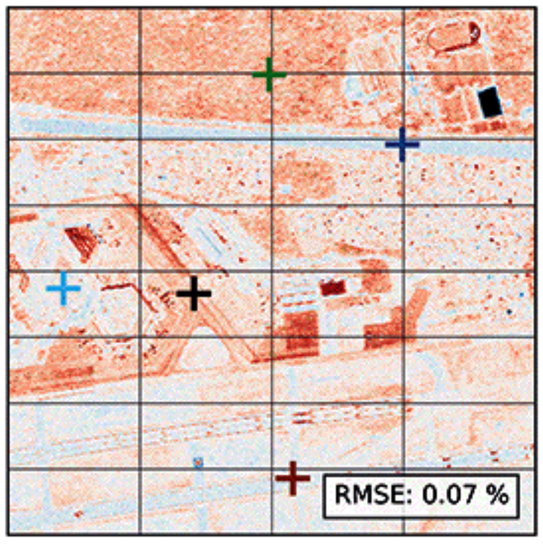
Spectra for LI, Landsat-8 OLI



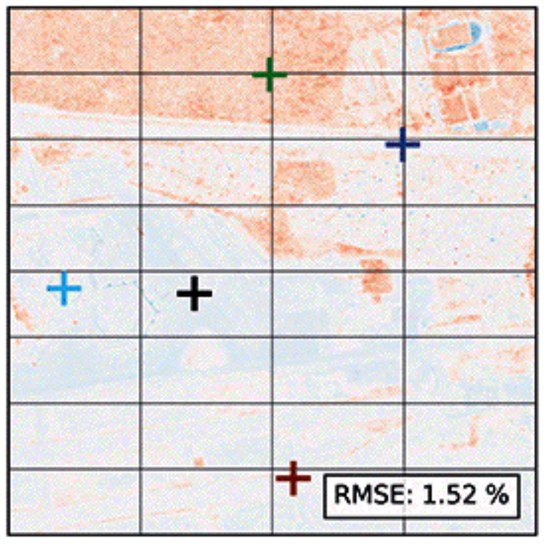
Map of assigned sub-clusters



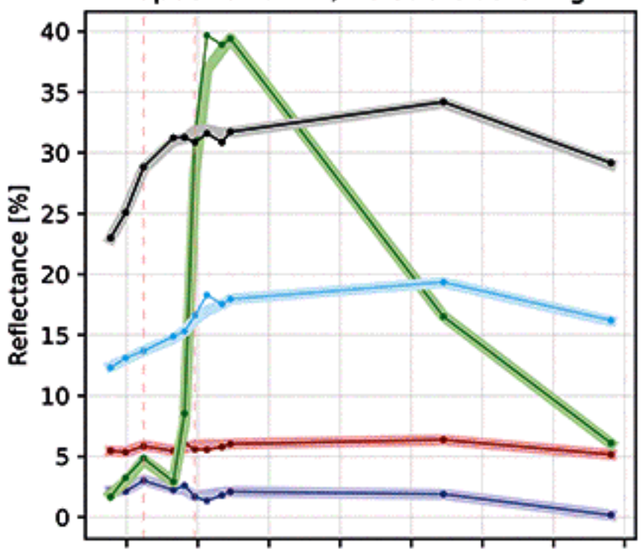
Difference, 560 nm, LR, no sub-clustering



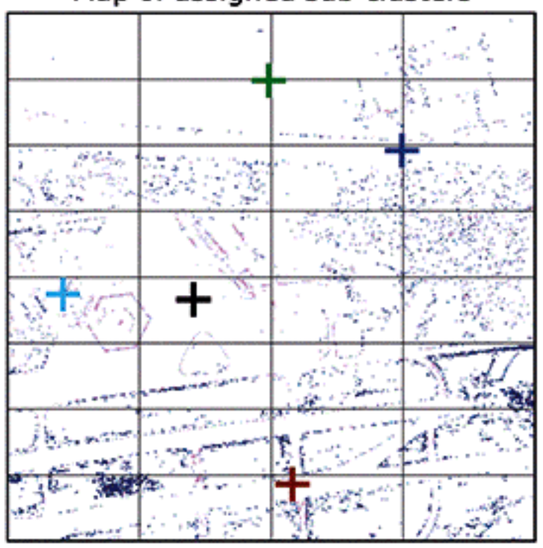
Difference, 740 nm, LR, no sub-clustering



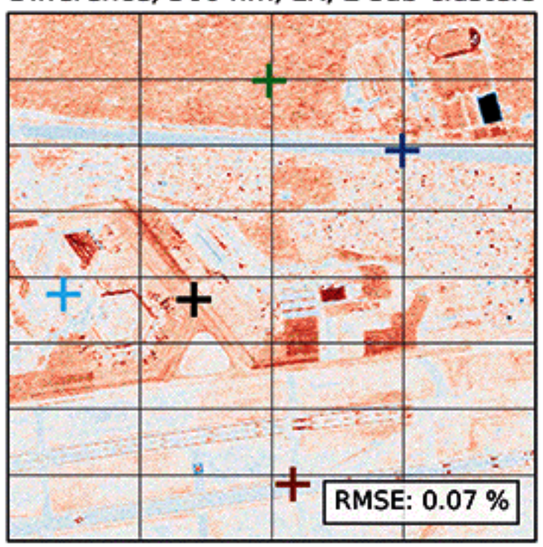
Spectra for LR, no sub-clustering



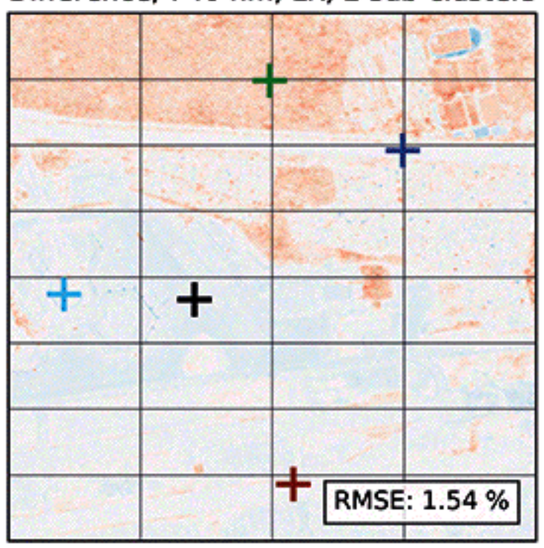
Map of assigned sub-clusters



Difference, 560 nm, LR, 2 sub-clusters



Difference, 740 nm, LR, 2 sub-clusters



Spectra for LR, 2 sub-clusters