Sentinel-2 Agriculture

Design Justification File

Benchmarking for L3 monthly composite product

Milestone	Milestone 2
Version	1.0
Authors	CESBIO - Olivier Hagolle, David Morin

Contents

1 Introduction

Cloud free composites or syntheses of surface reflectance (also called Level 3A products) may be useful for several reasons:

- they cover surfaces larger than that of a single satellite image;
- they can be provided at the same date every year and do not depend on a cloud free acquisition date;
- they enable a data volume reduction compared to the level 2A products (but - - they also represent a data loss compared to the level 2A);

many algorithms for classification or segmentation do not easily handle the presence of data gaps in the time series.

The Sentinel-2 Agri User Requirement Document requires monthly Level 3A composites produced every month, provided with several masks[1]

It is important here to recall that L3A syntheses are not mosaics. It means that a synthesis aims at keeping a physical meaning to the reflectance value of each pixel, while a mosaic aims only at producing a visually homogeneous with as little artifacts as possible. For instance, the RapidEye composite product is advertised as "All RapidEye Mosaic™ products will be tonally balanced to create a visually pleasing, seamless product"[2]. Mosaicking methods are based on image harmonization, using or not a reference image from another sensor such as MODIS for instance[3]. They often search for seamlines that avoid uniform zones to hide the change of image in the contours of objects, however, these methods are really not adapted to multi-temporal imagery, in the sense that they are not meant to preserve the multi-temporal consistency of images. This is clearly not what is seeked here.

Many algorithms may be used to produce a cloud free synthesis. The most traditional method is the famous NDVI Maximum Value Composite (NDVI MVC) method[4], initially developed to obtain composite products from moderate resolution optical satellites. The main advantage of this method is to select the date the most likely to be cloud free among the list of available dates in the compositing period. This method also does not require heavy computing resources. But this type of composite, which only selects one date for each pixel and discards the others, always results in very noisy surface reflectance composites, because for a given pixel :

- the selected date may have been acquired under a different viewing angle compared to the neighbourhood;
- the selected date may be affected by a cloud shadow or observed in the vicinity of a cloud;
- surface reflectance may have changed with during/within the compositing period, and if the date changes from one pixel to the other, the image appears noisy.

More recent methods make use of all the valid observations within a compositing period, either by averaging all the valid data during the compositing period (for instance the method used by UCL to produce the Globcover products[5]), or by fitting a directional model to cope with the directional effects problem (such as the CYCLOPES composite [6] developed by CNES, or the MODIS NBAR composite [7]). All these methods provide much better results than the basic max-NDVI method (see Figure 1 and Section (Preliminary benchmarking phase).

Compared to very wide field of view instruments, in the case of Sentinel-2, the problem is really simplified for two main reasons:

- the cloud detection is much more accurate when performed at a high resolution, with images acquired with viewing angles close to nadir, and with a large diversity of spectral bands including the 1.38 µm spectral band, able to detect thin cirrus clouds;
- the directional effects, although still present with Sentinel-2, are largely reduced thanks to the limited viewing angle of the acquisitions.

However, since the revisit cycle is not daily but 5 days (10 days in the initial phase with one satellite), the amount of valid dates is somewhat reduced compared to medium resolution instruments.

[1] Sen2-Agri$_{TS1}$.0 : Sen2Agri Technical specification

[2] http://blackbridge.com/rapideye/upload/RapidEye$_{MosaicProductSpecifications}$.pdf

[3] Krauß, T., 2014: Six Years Operational Processing of Satellite data using CATENA at DLR: Experiences and Recommendations. In: Kartographische Nachrichten, Journal of Cartography and Geographic Information, Vol. 2/2014

[4] Holben, B. N. (1986). Characteristics of maximum-value composite images from temporal AVHRR data. International Journal of Remote Sensing, 7(11), 1417-1434.

[5] Vancutsem, C., Pekel, J. F., Bogaert, P., & Defourny, P. (2007). Mean Compositing, an alternative strategy for producing temporal syntheses. Concepts and performance assessment for SPOT VEGETATION time series. International Journal of Remote Sensing, 28(22), 5123-5141.

[6] Hagolle, O., Lobo, A., Maisongrande, P., Cabot, F., Duchemin, B., & De Pereyra, A. (2005). Quality assessment and improvement of temporally composited products of remotely sensed imagery by combination of VEGETATION 1 and 2 images. Remote sensing of environment, 94(2), 172-186.

[7] Schaaf, C. B., Gao, F., Strahler, A. H., Lucht, W., Li, X., Tsang, T., ... & Roy, D. (2002). First operational BRDF, albedo nadir reflectance products from MODIS. Remote sensing of Environment, 83(1), 135-148.

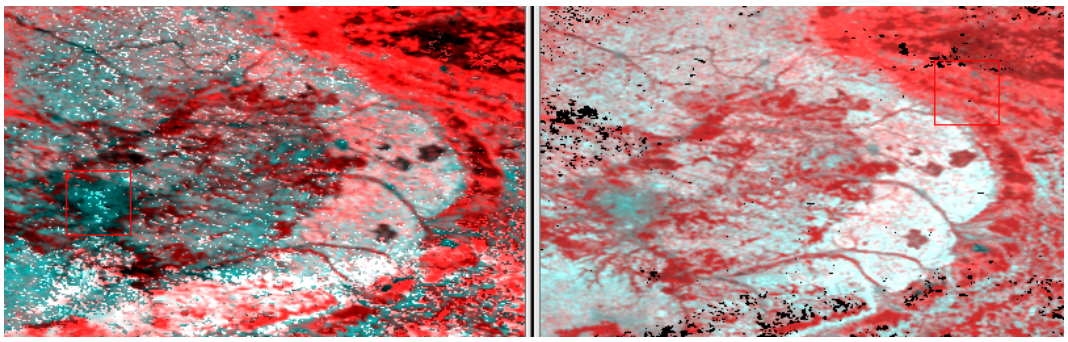

Figure 1: Example of an artifact (left), due to the use of different sets of dates depending on the pixel (middle). Our criterion computes the difference between the averages of values for the inner contour and outer contour of each zone with the same set of dates (right)

2 Issues and methods

2.1 Clouds and shadows

- most clouds and cloud shadows are detected in the Level-2A product, but of course, some

of them may be missed by the cloud masking. It is easy to use the cloud and shadows masks to discard pixels from the level 3A product, but the undetected clouds or shadows may be the cause of artefacts in the the level 3A products. The existing compositing methods propose strategies to minimize these artifacts.

To summarize, the compositing methods may be split in two categories:

- **Best pixel composites**, for which the best pixel according to several criteria is selected among the available dates within a certain time distance. There are several classical ways of selecting the best pixels, mostly to avoid undetected clouds or cloud shadows :
  - choose the maximum NDVI (which will enable to avoid clouds);

choose the minimum blue band (but it will tend to select also undetected shadows);

- choose the maximum of temperature (as in WELD products[8]), but this methods is not available for Sentinel-2 which lacks thermal bands;
- the S2-PAD ATBD[9] proposes other variants which will be analyzed below and which are:
  - choose the most recent cloud free date
  - choose the date with the minimum cloud cover
  - choose the date with the minimum AOT
- **Average syntheses**, for which the reflectance value is the average of surface reflectances of cloud free pixels. Here the idea is to rely on the repetitivity of observations to statistically reduce errors that could happen due to undetected clouds or cloud shadows or atmospheric correction errors. Weighted average may be used to favour dates with low aerosol content, low cloudiness and pixels far from clouds. However, the weighting must be light enough so that it does not finally select only one date, and finally looks like a best pixel method.

2.2 Snow and water

The surface covered by water and especially snow is highly variable from date to date, and their reflectances are very different from those of a vegetated area or of a bare soil. For SEN2AGRI project, For each pixel, it has been decided to discard water and snow dates except for pixels always covered by water and snow. This method has the drawback to provide sharp limits around the water or snow regions. This should not be an issue for water as these sharp limits are naturally observed, but could be more problematic for snow where a

[8]D. P. Roy, J. Ju, K. Kline, P. L. Scaramuzza, V. Kovalskyy, M. Hansen, T. R. Loveland, E. Vermote, and C. Zhang, "Web-enabled Landsat Data (WELD): Landsat ETM+ composited mosaics of the conterminous United States," Remote Sensing of Environment, vol. 114, no. 1, pp. 35–49, 2010.

[9]Sentinel-2 spatio-temporal synthesis ATBD: Sentinel-2 MSI - Level 3 Products Algorithm Theoretical Basis. Document, Volume A, issue 2.0, 18 Jun 2010, ref. S2PAD-VEGA-ATBD-0005

transition is usually observed between regions covered or not by snow. Yet, including snow reflectance values in the compositing is found more critical than these potential sharp limits.

As a result, here is the method proposed :

- For each pixel
  - for all dates
    - if at least one date is cloud/cloud shadow free
      - discard all cloud and cloud shadow pixels
      - if at least one of the remaining dates is snow or water free
        - discard snow and water dates
        - compute the weighted average of remaining dates
        - flag the pixel as land
      - else :
        - if at least one date is snow free
          - discard snow dates
          - compute the average of remaining dates
          - flag the pixel as water
        - else :
          - compute the weighted average of snow date
          - flag the pixel as snow
    - else :
      - flag the pixel as Cloud/CloudShadow

2.3 Directional correction

2.3.1 Issue description

As there is often some overlap between the Sentinel-2 images acquired from adjacent orbits, users may want to have directionally normalized composites that would allow producing seamless multi-swath composites. 4 sites of the SPOT4 (Take5) experiment were observed from 2 viewing different directions, in order to enable us testing a directional correction method that will be necessary for Sentinel-2.

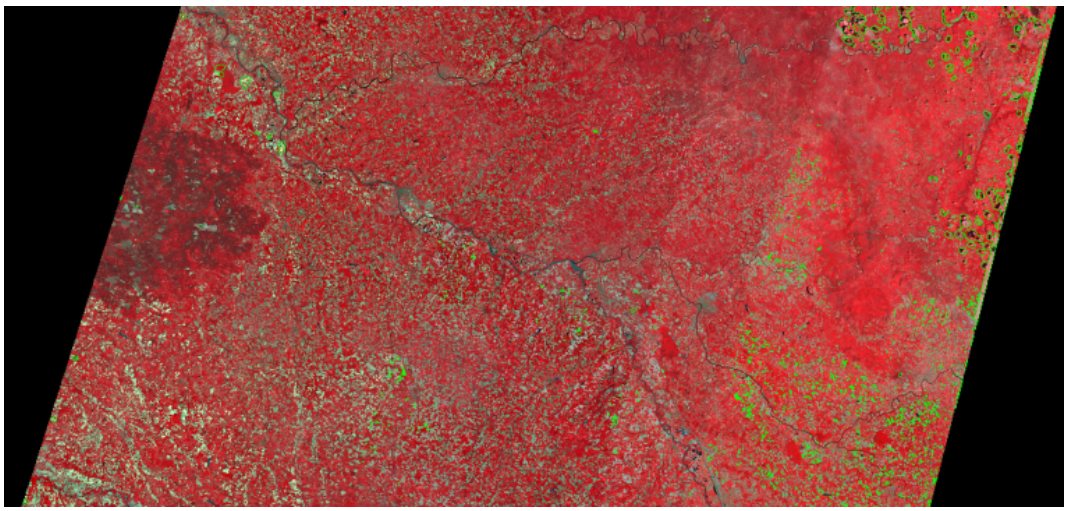

Figure 2: Example of a multi-swath weighted average synthesis from 2 SPOT4 (Take5) swaths acquired under different viewing angles (angle difference twice higher than that of Sentinel-2). The pixels outlined in green correspond to invalid data (due to a very cloudy spring in France in 2013, or to saturation in SPOT4 images)

In the image above, three zones are visible: the left three quarters of the image were taken from the west, while the East quarter was taken from the East, in between is the overlap region. The left and right parts have different mean reflectance values, and the overlap region average reflectance is in the middle.. A simple directional correction could somewhat reduce the observed artifacts.

2.3.2 Solution adopted for benchmarking

Directional variations of surfaces at a high resolution are highly dependent on the surface nature. For instance, 15 days after sowing, a wheat field will have a Bidirectional Reflectance Distribution Function (BRDF) different from the one it will have one month after sowing. Moreover, the BDRF of a field with ranks oriented North South will differ from that of a field oriented East West[10]. This is even worse when comparing different species.

This explains why directional correction at a high resolution is still in the research domain and why up to now, only basic algorithms exist[11], such as applying a constant average directional model. As the angle variation in Sentinel-2 images is not too large, it is worth trying this approach to see the amplitude of the residual artefacts and foster subsequent research. A more complex approach, quoted in our Sen2Agri proposal, was tested: it consists in computing global directional models at a low resolution (5km) using MODIS data. According to the authors, these models can be used all year long thanks to a parametrization with NDVI. However, it turned out that the coefficients provided by Vermote and al did not much enhance the directional homogeneity of the SPOT4 (Take5). It turned out that the coefficients obtained with MODIS were highly correlated with the presence of slopes within the 5 km MODIS pixels, resulting in very bad results when used at a high resolution.

Three very different SPOT4 (Take5) sites acquired from two viewing directions (Sudmipy, Maricopa, Provence) were used to estimate the parameters of a directional model (derived from [12]). This model is a kernel model which has the following form :

$$\rho(\theta_s, \theta_v, \phi) = \rho_0.(1 + V.F_V(\theta_s, \theta_v, \phi) + R.F_R(\theta_s, \theta_v, \phi))$$

where θ_s is the sun zenith angle, θ_v is the view zenith angle and ϕ is the relative azimuth. F_V is the kernel function that represents volume scattering, while F_R estimates the directional reflectance of a flat surface with randomly distributed and oriented protrusions. ρ_0, R and V are coefficients.

The directional correction consists in using this same model with constant coefficients R and V for all pixels, dividing by the model value for the actual viewing conditions and multiplying by the model value at NADIR conditions, for the mean zenith solar angle of the compositiong period.

R and V have been determined so that they enable to minimize differences between observations of the 3 multi-directional SPOT4 (Take5) sites.

2.4 Gap filling

Even if a month of data is used, there may still be data gaps with pixels for which all the observations were cloudy. Users might be interested to have these gaps filled using some sort of interpolation.

We implemented a simple linear time interpolation that uses the composite products before and after the composite date. This method has an important drawback, which is the fact that it is not possible to deliver the gap filled products in real time, but one month later.

3 Level 3A criteria

We have defined 3 quality criteria to assess and compare the performances of various synthesis methods :

- gaps : the remaining proportion of data gaps after the synthesis
- fidelity : the fidelity of the composite Level 3A image to a Level 2A image close to the Level 3A medium date
- artifacts : the amplitude of artifacts observable at the limits of zones obtained with the same set of dates

In order to obtain a sufficient number of measurements of the performances of Level 3A products, these quality criteria were computed for several SPOT4 (Take5) sites, and a Level 3A was produced every seventh day.

[10] S.Duthoit, Prise en compte de l'agrégation des cultures dans la simulation du transfert radiatif: importance pour l'estimation de l'indice foliaire (LAI), de la parcelle au paysage, PhD Mabuscript, Université Paul Sabatier France,2006

[11] N. Flood, T. Danaher, T. Gill, and S. Gillingham, "An operational scheme for deriving standardised surface reflectance from Landsat TM/ETM+ and SPOT HRG imagery for eastern Australia," Remote Sensing, vol. 5, no. 1, pp. 83–109, 2013.

[12] F. Maignan, F.-M. Breon, et R. Lacaze, «Bidirectional reflectance of Earth targets: Evaluation of analytical models using a large set of spaceborne measurements with emphasis on the Hot Spot», Remote Sensing of Environment, vol. 90, n 2, p. 210–220, 2004.

3.1 Remaining proportion of data gaps

This criterion is rather simple : it consists in counting the pixels with dummy values within the image footprint, and divide by the number of pixels which should have been observed if at least an image had been completely cloud free.

$$residualgaps = \frac{numberofpixelsindatagapwithinfootprint}{numberofpixelswithinfootprint}$$

In the plots, we provide the average value of the remaining data gaps for all the Level 3A products available for a time series.

3.2 Fidelity to the central date

If everything was perfect, the Level 3A synthesis of the 15th of February should be identical to a cloud free Level 2A acquired at that date, if it existed.

As a result, our fidelity criterion measures the differences between the Level 3A surface reflectance and the Level 2A surface reflectance, when a relatively cloud free Level 2A image is available for a date close to the central date of the composite (+/- 8 days). For each SPOT4 (Take5) site, we automatically selected all the images with less than 50% cloud cover, but computed the comparison only for the cloud free pixels of course.

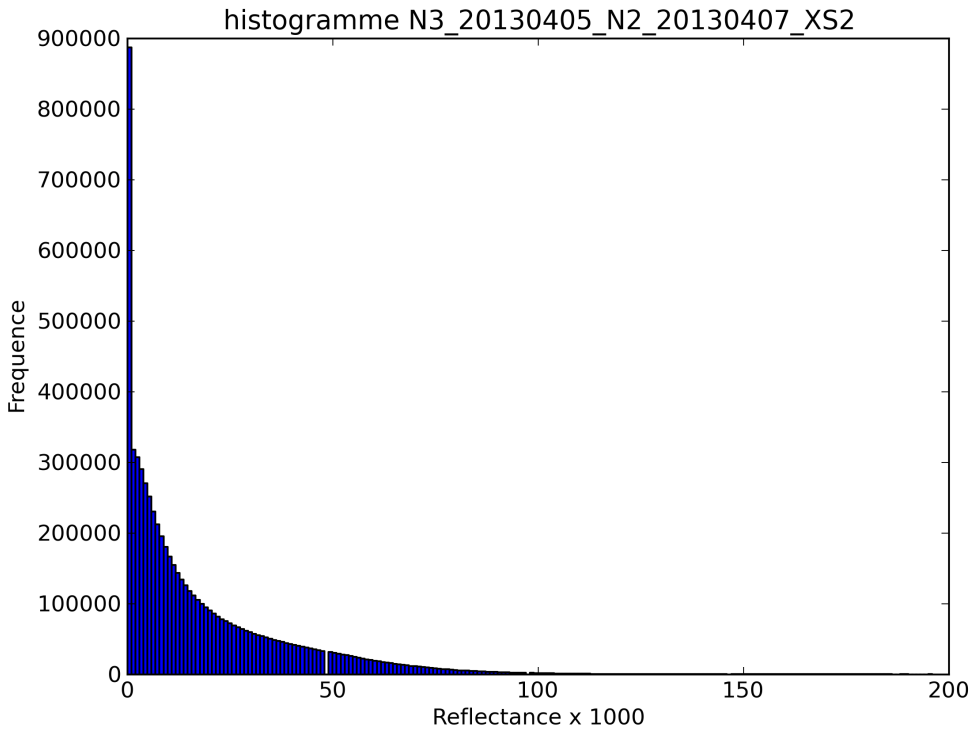

Figure 3: Error histogram for the comparison of a Level 2A and a level 3A image using the NDVI-MVC method.

When a selected Level 2A product is available close to the synthesis central date, the difference image between the level 3A and level 2A images is computed and for all the pixels which are cloud free in the level 2A images, the following statistics are computed :

- 70% percentile : Maximum value of the absolute value of the difference between level 3A and level 2A, for the 70% of pixels which have the lowest absolute value of difference.
- 95% percentile : Maximum value of the absolute value of the difference between level 3A and level 2A, for the 90% of pixels which have the lowest absolute value of difference.

One may object that this quality criterion is optimistic since it is only computed when a good quality level 2A image is available at the center of the level 3A period. But we have compared the performances obtained with or without including this level 2A image in the level 3A composite. the performances are of course a little worse when the image is not used, but the difference is quite low, as can be seen in the figure 4.

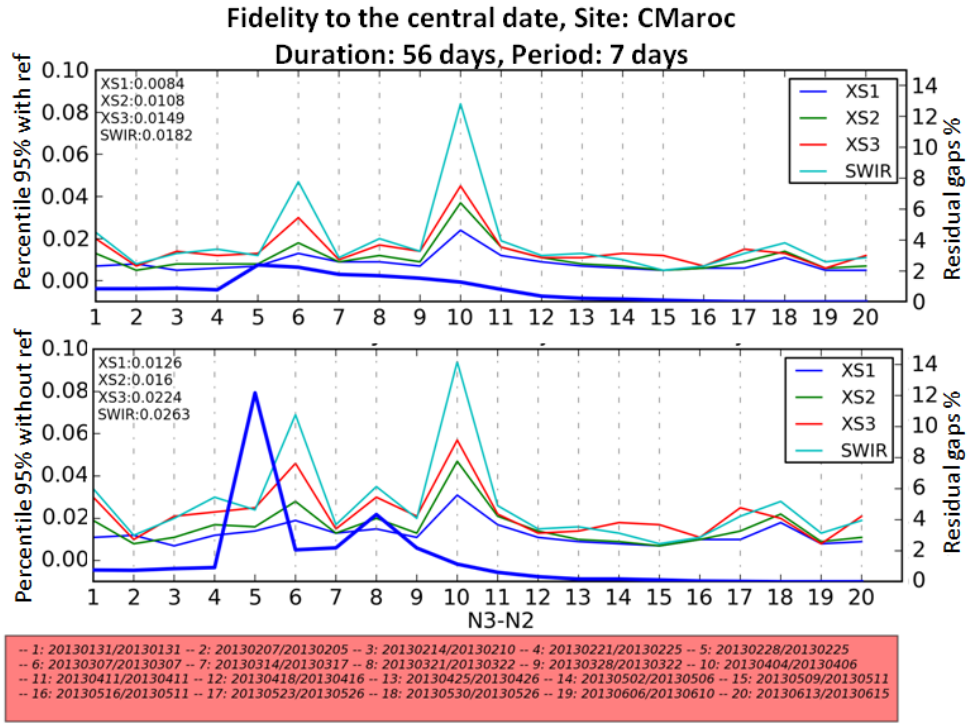

Figure 4: Comparison of the 95% fidelity criterion as a function of Level3A/Level2A couples of dates, computed using (top) or not (bottom) the central date in the composite. For most dates the difference is small except for two of them where the number of available data was very low without the central date.

3.3 Artifacts

Due to the presence of clouds in the level 2A images, a level 3A image is obtained :

- for the "best pixel" algorithms, from dates which may differ from one pixel to the next one
- for the "weighted average" algorithms, from sets of dates which may differ from one pixel to the next one

Artifacts may appear in the level 3A products along the limits of contiguous zones obtained with the same set of dates.

For all the connected groups of pixels with the same set of dates, the average difference between the external border and the internal border of the contiguous zone is computed. Our artifact criterion is the standard deviation of this average for all the contiguous zones with the same date or same set of dates.

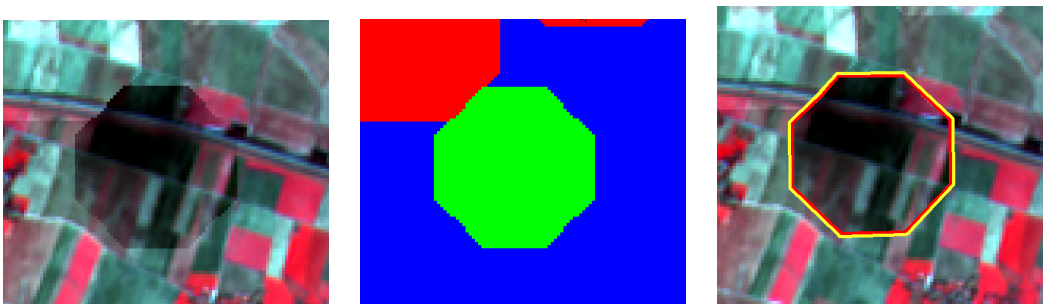

Figure 5: Example of an artifact (left), due to the use of different sets of dates depending on the pixel (middle). Our criterion computes the difference between the averages of values for the inner contour and outer contour of each zone with the same set of dates (right)

4 Preliminary benchmarking phase

Although the benchmarking phase will allow us to test the performances of 5 strategies to produce monthly syntheses on 12 sites, the number of strategies that would need to be tested is much larger. For this reason, the Sen2Agri project decided to hold a selection phase on a little number of sites before selecting the 5 strategies. The idea was to compare the performances of best pixel and weighted average approaches, to focus the detailed benchmarking phase on the approach that yields the best results.

4.1 selected methods

In this phase, we compared two variants of the best pixel method to the weighted average method (decribed above) :

4.1.1 The NDVI MVC

In this method, for a given compositing period, for each pixel, the selected date is the one of the valid pixel which has the greatest NDVI.

4.1.2 The priority to the least cloudy image

This method is one of the methods proposed within the S2-PAD project [9]. It works as follows : the selected date corresponds to the date with the lowest cloud cover, if the pixel is valid (not cloud, no snow, no cloud shadow)

Variables

- $refl_{composite}$(band,line,pixel) : table with the output values of the composite image
- refl2A(date,band,line,pixel) : table with the input reflectance value of L2A products for all dates

Pseudocode

- Rank the available dates during the compositing period by increasing cloud cover(or cloud shadow) : D1 =>DN
- For each pixel (l,p)
  - For each spectral band b
    - Di=D1
    - while $refl_{composite}$ (b,l,p)=Null and (Di<DN)
      - If (l,p) is valid (no cloud, no shadow, no saturation)
        - $Refl_{composite}$ (b,l,p)=refl2A(D,b,l,p)
      - Else :
        - DI+=1

4.1.3 Weighted average

The selected weighting strategy for our method was defined for the Venμs ground segment. In order to enhance the fidelity to the central date, and to reduce artifacts due to undetected clouds or shadows, it gives more weight :

- to the images closer to the Level 3A date.
- to the images with a low aerosol content
- to the pixels far from a cloud

4.2 Comparison

All 3 methods were compared using the same compositing period: 42 days. Given that the same compositing period was used for the three methods, the 3 methods have exactly the same amount of remaining data gaps. The methods have been compared for the two remaining criteria, i.e. the fidelity to the central date and the artifact criterion.

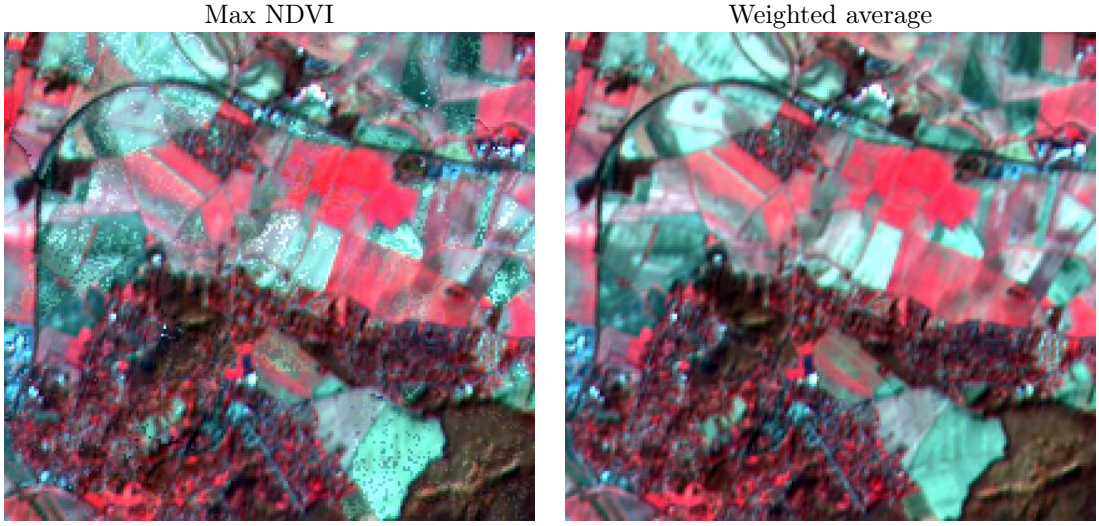

Figure 6: NDVI MVC composite product (left) compared with weighted average composite product (right). A large noise is clearly visible on the NDVI MVC composite due to the change of selected date from one pixel to the next

4.2.1 Visual comparison

As said above, the NDVI MVC shows a very large noise, that we call artifacts, due to the selection of pixels from different dates from one pixel to the other. Between two dates, even separated by a few days, many things may change, such as the vegetation cover or the soil moisture and its colour. The selection criterion of the NDVI MVC leads to select different dates from one pixel to the other, which explains a very large noise, which is not observed on the weighted average images. This can be easily seen on any NDVI MVC composite, provided several images are available in the compositing period. For Instance, figure fig:NDVI_MoyPond compares a detail of a NDVI MVC composite with a weighted average synthesis, with a much higher noise on the NDVI MVC composite.

Contrarily to the NDVI MVC composite, Min cloud method is designed to minimize artifacts, since it selects the image with the largest number of valid points within the available set of dates. As a result, the possibility to observe artifacts when the set of dates changes is reduced. But this method has an important drawback, as the selected date in the compositing period is selected on a criterion of nebulosity, which from a vegetation phenology point of view is equivalent to a random selection of the date within a month period.

As an illustration, we often found periods when successive Min Cloud composites produced every two weeks, based on a 6 weeks synthesis period where identical for 3 images in a row, as for the 3 compositing periods, it was the same date which had the greatest amount of cloud free pixels (see figures 4.2.1 and **??**). Compared to that, the same series of (Weighted Average composites) show a normal evolution, with vegetation development growing during spring.

4.2.2 quantitative analysis

What we just showed is of course visible in the quantitative analysis using the criteria used in section Level 3A criteria. The figure 8 shows a summary of the fidelity and artifact criteria for Versailles site, for the weighted average, the ESA min cloud and the $NDVI_{MVC}$, computed for the composite products produced every week with a 42 days period. The graphs confirm the visual impression observed in the figure above, with a very large amount of artefacts for the $NDVI_{MVC}$, while both other method have similar performances. The NDVI MVC has also a worse fidelity, especially in the NIR and SWIR, because in this spring season, the vegetation is growing and for vegetation pixels the NDVI MVC tends to select the latest dates with the greatest NDVI, which are there fore different from the images at the center of the compositing period.

Regarding the Min Cloud and the Weighted average, the observed performances are similar, with a small advantage to the weighted average. It is necessary to look more closely at the detailed results to understand why the Min cloud nearly matches the performances of the weighted average.

The conclusions observed for the Versailles site are also valid for two other sites (Belgium and Morocco, 9) for which we processed syntheses over the whole time series, for the three methods. The NDVI-MVC produces artefacts at least 10 times higher than the other methods and it is usually also much worse in terms of fidelity. There is an exception to that regarding the Morocco site due to the presence of snow. Snow is a highly variable cover and is not well suited to produce monthly syntheses. It often produce random results depending whether

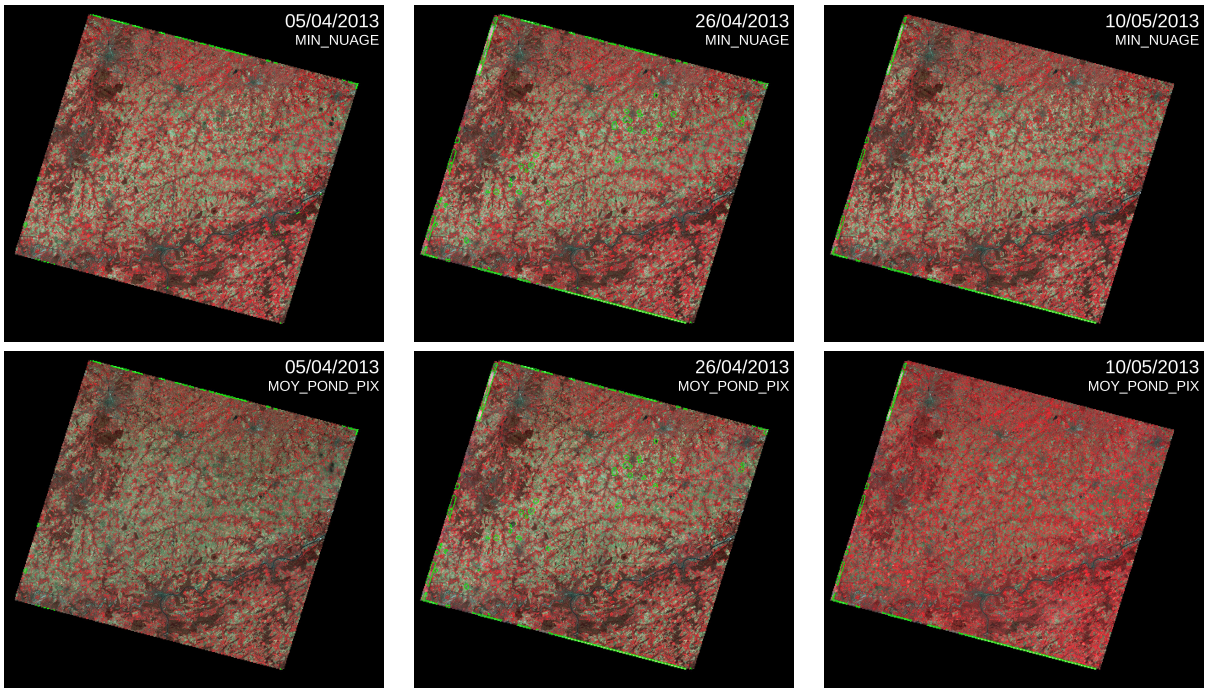

Figure 7: Min cloud composites (top) and weighted average (bottom), for Versailles site, for 3 successive dates of synthesis. On the Min cloud composites, 90% of the pixels come from the same image and are identical for the 3 dates, which is not the case for the weighted average

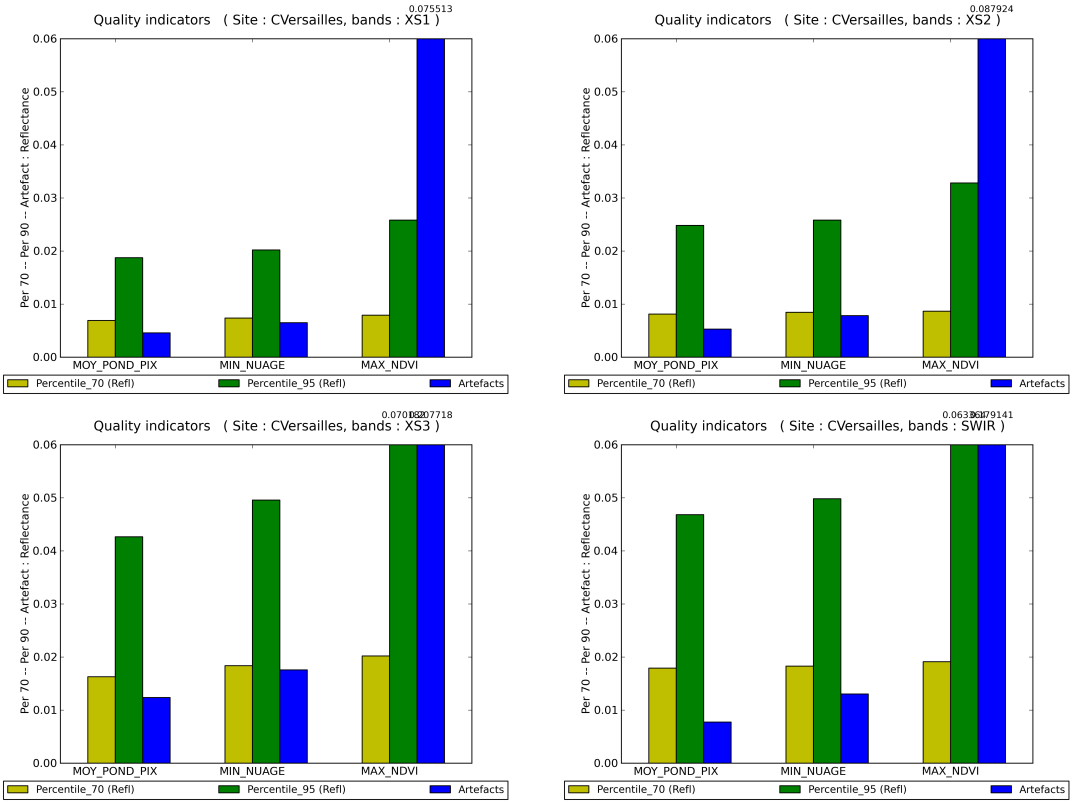

Figure 8: Compilation of quality indices for the 3 selected methods (from left to right, Weighted average, Min cloud, and NDVI MVC) for Versailles site, and for the 4 bands of SPOT4

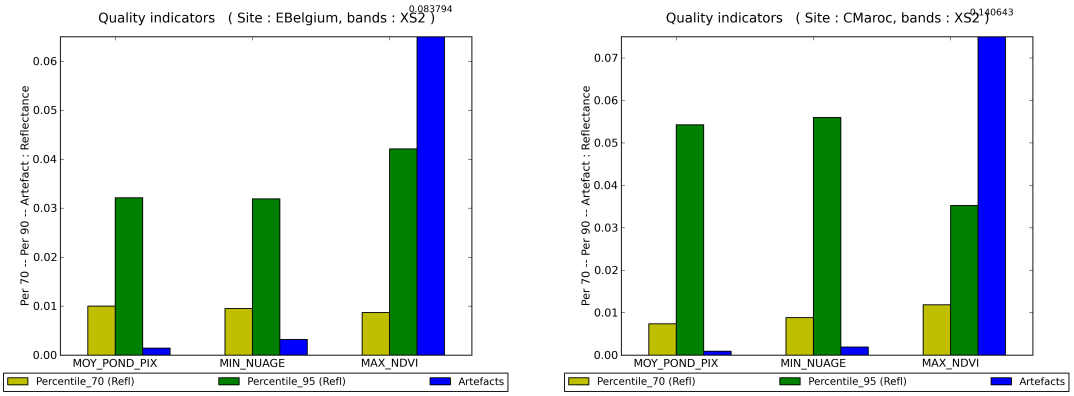

Figure 9: Compilation of quality indices for the 3 selected methods (from left to right, Weighted average, Min cloud, and NDVI MVC) for 2 other sites (Belgium,and Morocco), for the red band of SPOT 4.

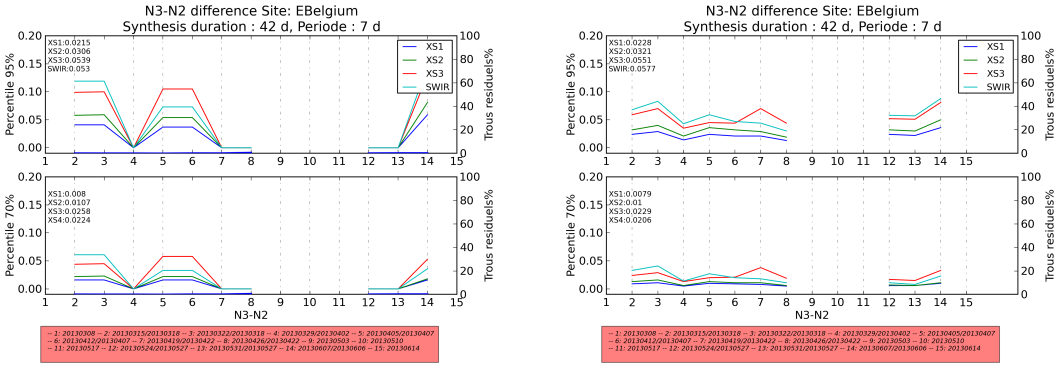

Figure 10: Performances for fidelity criterion, as a function of the synthesis date, left for Min Cloud, and right for weighted average. For some dates, the performance is not evaluated because a cloud free L2A image is not available as reference. For the Min cloud, the fidelity is sometimes perfect, when the whole L2A reference image is used in the level 3A product. When the performance is constant for several successive dates, it means that the same image was used in the various Min Cloud composites.

the image used as reference for the fidelity criterion is largely covered by snow or not.

Regarding the comparison of performances of the Weighted Average or the Min Cloud, the obtained results are much closer, but the artifacts criterion gives a large advantage (a factor 2) to the Weighted average. The fidelity criterion look close, but, as we will see below it is the result of an artifact in the fidelity criterion evaluation, and in fact, here again, the weighted average has a clear advantage.

Looking at the values of the fidelity criterion as a function of the composite date (Figure 10), one can see that the weighted average has usually a better performance than the ESA min cloud, except on a few days when the fidelity criterion is equal to zero. This phenomenon appears when the selected date for the ESA min cloud is in fact the reference date to compute the fidelity criterion. As a result, these values near zero are artifacts of the way the quality criterion is computed, and should not be taken into account.

To confirm this artifact is really the cause of the similarity of results between Weighted Average and Min Cloud, we computed the fidelity criterion excluding the images which are selected as reference image for fidelity criterion from the composite. This work was done on South Africa site (Figure 11), because we needed to use a site with a larger amount of images, in order to be able to remove the best one from each composite.

Regarding the artifacts (Figure 12), there is no doubt that the Weighted Average produces far better results than the ESA min cloud, even if the Min cloud is designed to minimize the artifacts, it still cannot compete with the weighted average, as the averaging reduces the steps at the borders of regions with an homogeneous set of dates.

4.3 Conclusion of the first phase of Benchmarking

This first phase of benchmarking aimed at selecting a baseline method for the second phase, in which the baseline will be tested with different options and sets of parameters. This first phase shows the superiority of

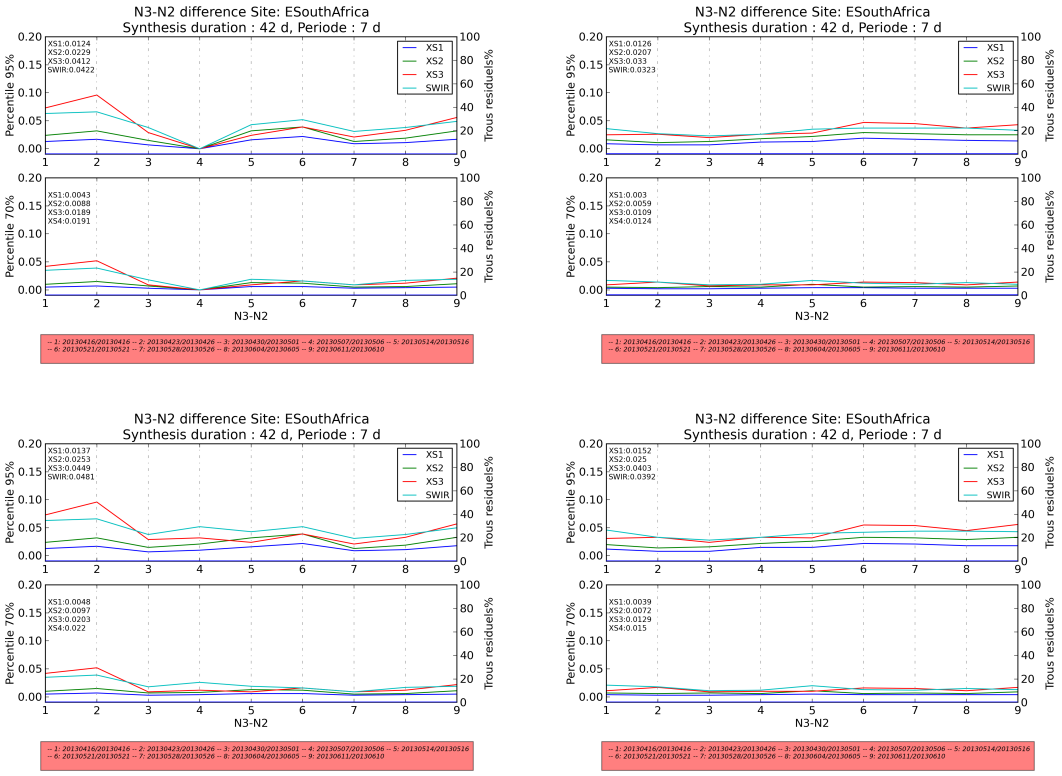

Figure 11: Results of Min Cloud (Left) and Weighted Average (right), for South Africa test data set, Top using the reference image used to estimate the fidelity criterion in the composites, and bottom after discarding the reference image. The better accuracy obtained by the Weighted average is confirmed

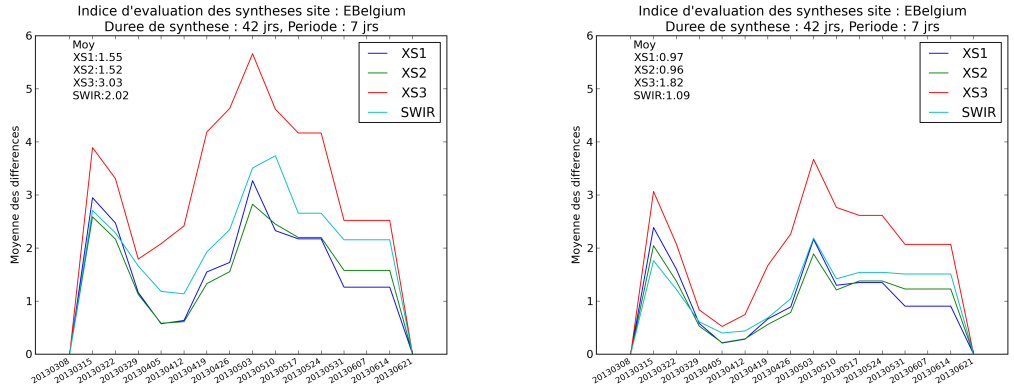

Figure 12: Artifact performance as a function of date, for the min cloud, left, and the weighted average, right

the weighted average method for all criteria with regard to the two selected "best pixel methods". The difference is obvious relatively to the historical NDVI-MVC method, which exhibits a large noise, even in conditions where the directional effects are minimized. But the weighted average also performs better tha the Min Cloud, in terms of artifacts and fidelity to the central date.

As a result, the weighted average is used as the baseline for the second benchmarking phase.

5 Benchmarking

5.1 Selected methods

For the benchmarking phase, based on the litetrature study and on the results of the preliminary phase, we finally selected six variants based on the baseline method chosen after the preliminary phase : the weighted average method.

The first 4 variants correspond to the test of various parameters of the Weighted Average Method, on all the Sen2Agri test data sets based on SPOT4 (Take5) data. The fifth variant adds a directional correction before the composite generation, while the sixth fills the remaining data gaps, via a temporal interpolation.

1. weighted average composite with a composite duration of 30 days
2. weighted average composite with a composite duration of 40 days
3. weighted average composite with a composite duration of 50 days
4. weighted average composite with a composite duration of 50 days, but based on images acquired every 10 days, to simulate the available time series of Sentinel-2 with only one satellite. The LANDSAT 8 data available during the period (after mid April) were integrated to the data set, to really simulate the situation with only one Sentinel-2A and Landsat 8, in 2015-2016.
5. weighted average composite with a compositing period of 40 days, preceded by a directional correction in order to stitch the images that come from different swaths, with different viewing angles. This variant can only be tested on two sites observed from two different swaths of SPOT4 (Take5) : Sudmipy and Maricopa.
6. weighted average composite with a compositing duration of 30 days, with a gapfilling of the residual clouds.

All these methods (except method 5) are evaluated for the 9 Sen2Agri sites which were observed with SPOT4 (Take5) : Madagascar, South Africa, Argentina, Belgium, Morocco, USA-Maricopa, Ukraine, China, France Sudmipy. However, the method 5 which includes a directional effect correction can only be applied to sites acquired from two viewing direction, and its validation was limited to France and USA.

5.2 Results for Weighted Average Method

5.2.1 Configuration for benchmarking

In order to compare the different configurations of the weighted average method, it is necessary to use them on the same dates. The first composite is generated 25 days (50/2) after the start of SPOT4 (Take5) experiment, and the last one 25 days before the last day of SPOT4 Take5. To increase the number of cases, the composites were computed every seventh day.

5.2.2 Madagascar

Madagascar site did not show large variations of vegetation cover during the SPOT4 (Take5) experiment. As a result (Figure 13), the fidelity criterion is stable with the compositing period, while less gaps and artifacts are observed with a compositing period of 50 days. With only one satellite, the rate of gaps is quite large with an average value of 6%. Globally, the best results are observed with a compositing period of 50 days.

Tested in the conditions of the first year of Sentinel-2, with a 50 days compositing window, the observed performances happen to be quite bad with a large number of residual data gaps, reaching 22% in March.

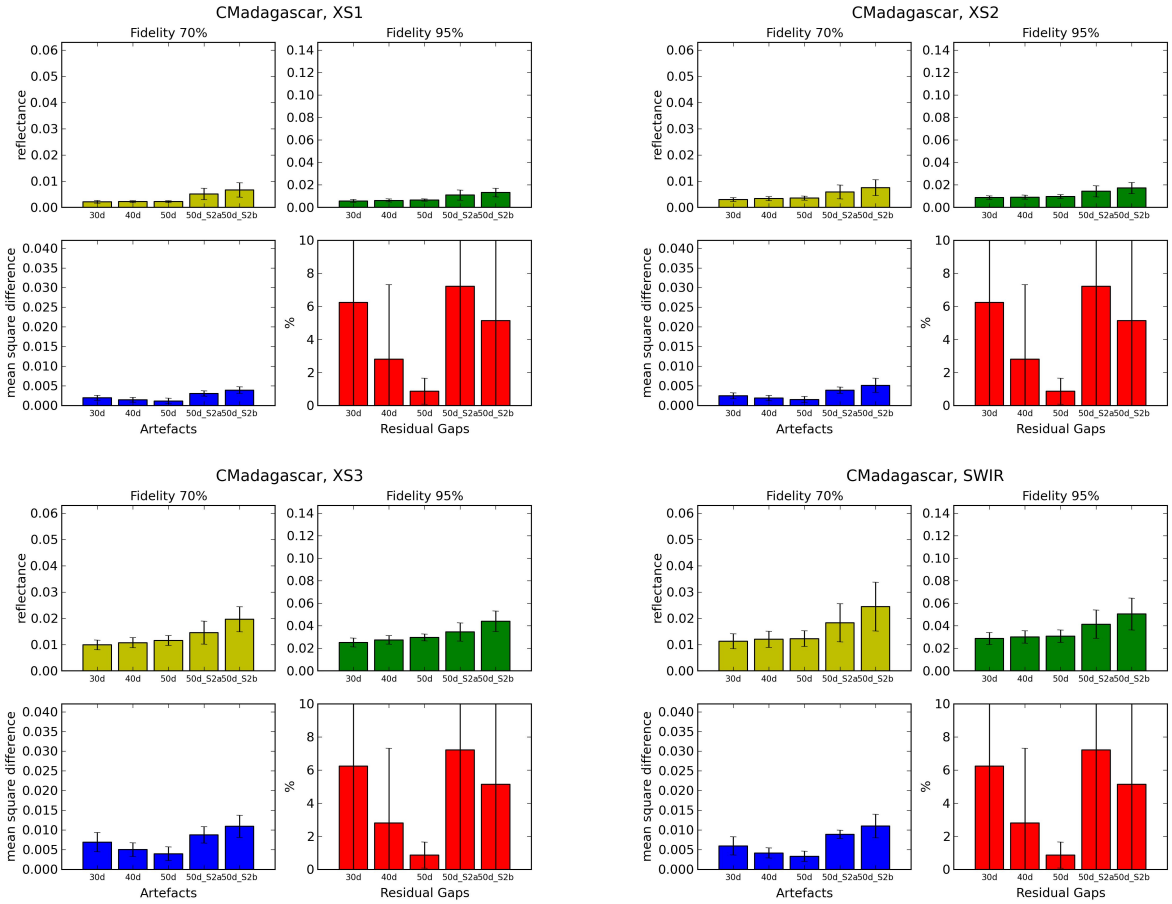

Figure 13: Summary of Benchmarking results for Madagascar site

5.2.3 South Africa

The South African site was observed in the senescence period, with some variation of vegetation cover, and the images cloudiness was relatively low. As a result (Figure 15), when the compositing period increases, only a moderate increase in terms of infidelity is observed, as well as little variations in terms of artifacts, while the proportion of data gaps is already low with a compositing period of 30 days. Globally, the best results are observed with a compositing period of 30 days.

5.2.4 Argentina

Conditions in Argentina were similar to those of South Africa. The site was observed in the senescence period, with some variation of vegetation cover, and the images cloudiness was relatively a little highe compared to South Africa. As a result (Figure 16), when the compositing period increases, some increase in terms of infidelity is observed, while little variations in terms of artifacts, while the proportion of data gaps is already low with a compositing period of 30 days. Globally again, the best results are observed with a compositing period of 30 days.

5.2.5 Belgium

Contrarily to the two previous sites, the weather conditions on this site were terrible, while the observation window happened in the period when vegetation changes mostly. As a result (Figure 17), when the compositing period increases, a large increase is observed in terms of infidelity, while a surprisingly low gain is observed in terms of artifacts. This is due to several causes (see 19) :

- the first composite with 30 days is not produced due to the absence of input images which were completely cloudy, while it is produced with 40 and 50 days periods, but with bad performances. This degrades the average performances of 40 and 50 days compared to 30 days.
- even with a long period of time, having two or more clear images in the compositing window is quite rare. It is sometimes not even possible to compute the artefact criterion with 30 days syntheses.

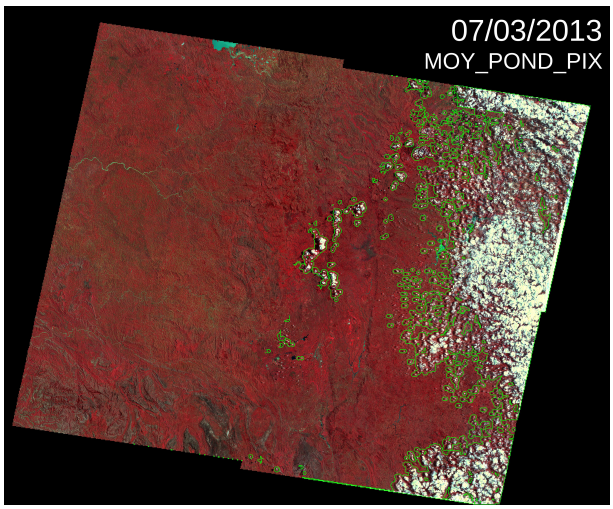

Figure 14: Weighted average synthesis obtained on Madagascar with a compositing epriod of 50 days, in the case of only one satellite.

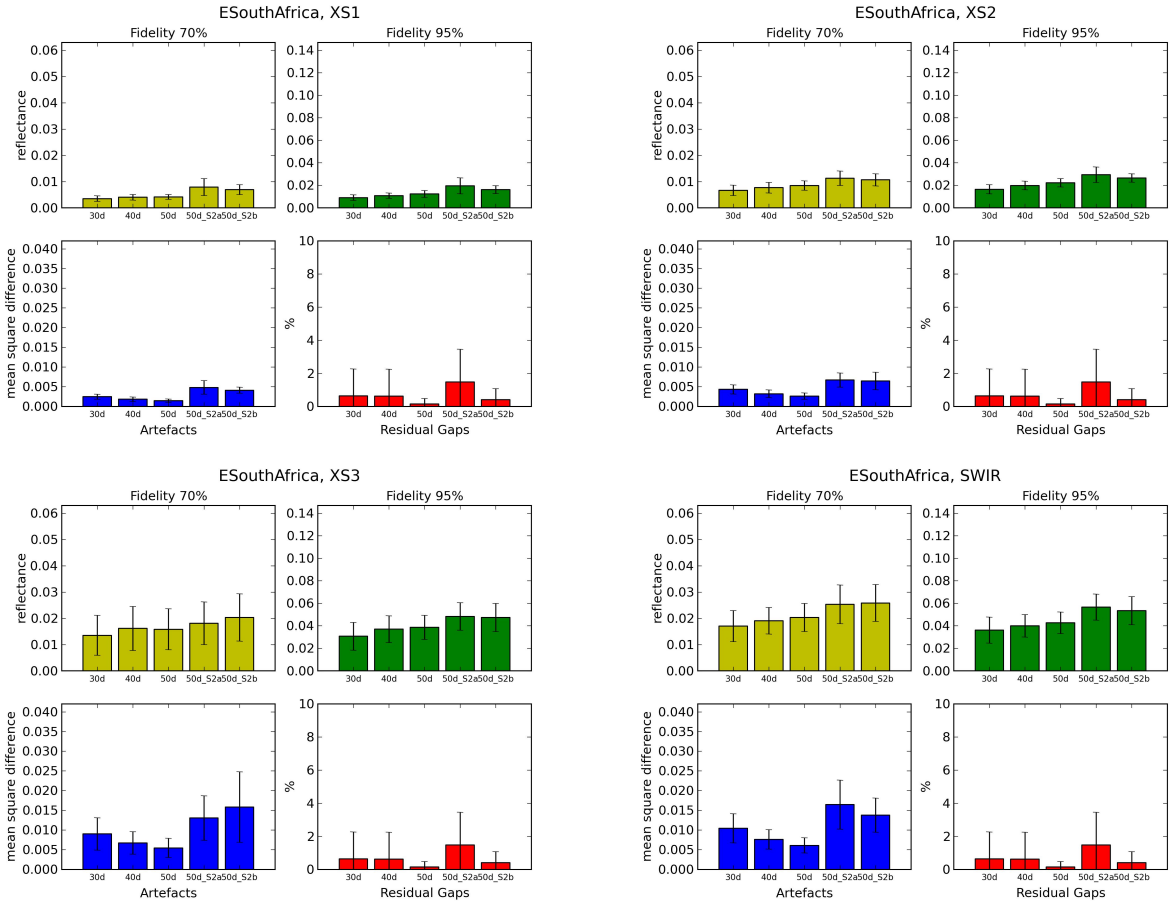

Figure 15: Summary of Benchmarking results for South Africa site

However, conversely to the previous sites, there is a large reduction of data gaps between 30 and 40 days of compositing period. For this reason, 40 days seems to be the best compromise.

Of course with such a low number of cloud free days, the performances of the syntheses obtained with a repetitivity of 10 days are poor, especially in terms of data gaps.

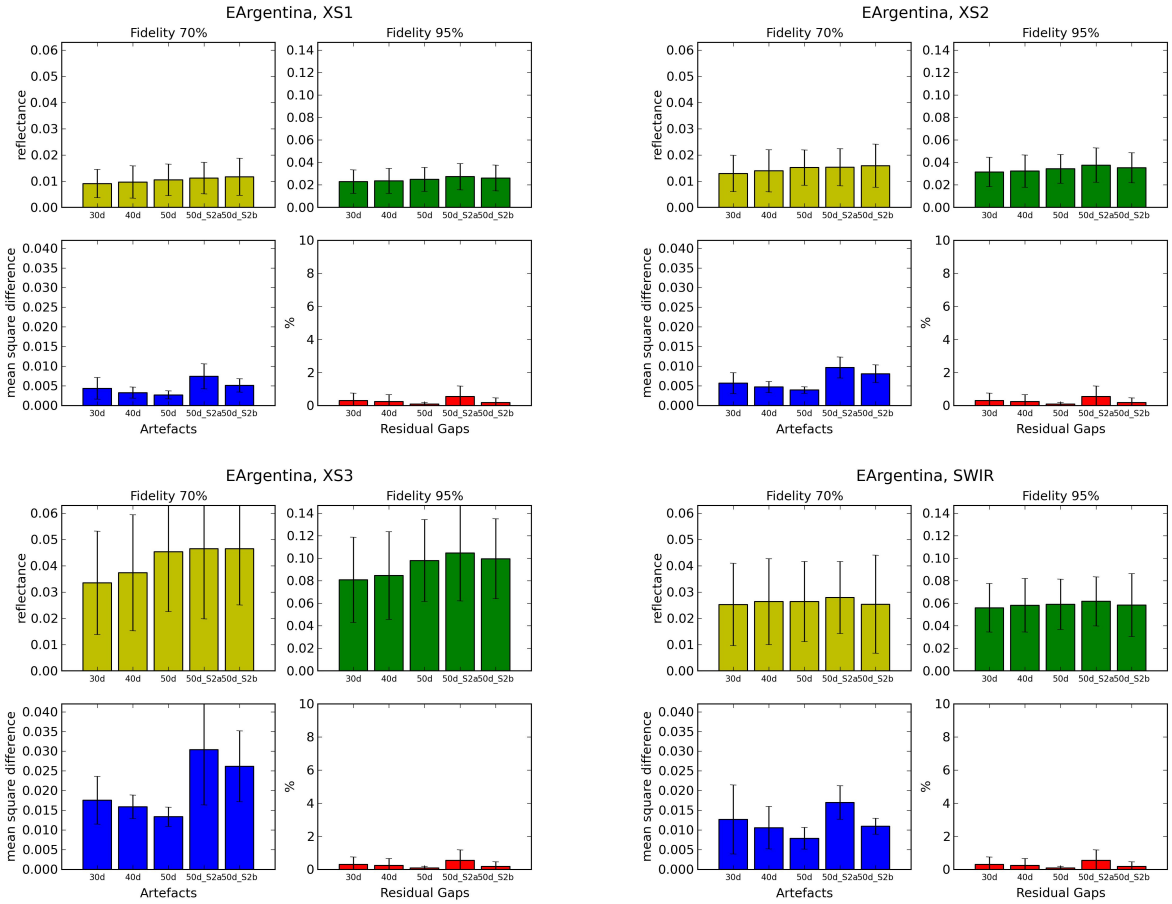

Figure 16: Summary of Benchmarking results for Argentina site

5.2.6 Morocco

The Morocco site was observed at the end of the growing period, with some variation of vegetation cover, and the images cloudiness was relatively low. When the compositing period increases (Figure 21), an insignificant increase in terms of infidelity is observed, while the quantity of artifacts is decreased, but is already very low with 30 days. The proportion of data gaps is already very low with a compositing period of 30 days. As a result, a compositing period of 30 days is enough for this site.

Even with a repetitivity of 10 days, the quality of composites would have been correct.

5.2.7 Maricopa

Maricopa is again a site with a very good weather, for which a compositing period of 30 days is enough (Figure **??**).

5.2.8 China

The China site was unexpectedly completely covered by snow at the beginning of February, resulting in unusable saturated images. Then, until April, all the images were affected by a huge amount of aerosols, resulting in very bad estimations of surface reflectances, and ugly syntheses. Starting in April the syntheses look correct. As a result (Figure 22), the level of artifacts is quite high on that site.

Using longer compositing periods reduces the level of artifacts, and the best results are obtained for compositing periods of 50 days.

It has to be noted that a very high amount of data gaps is observed in one of the cases with only 10 days repetitivity.

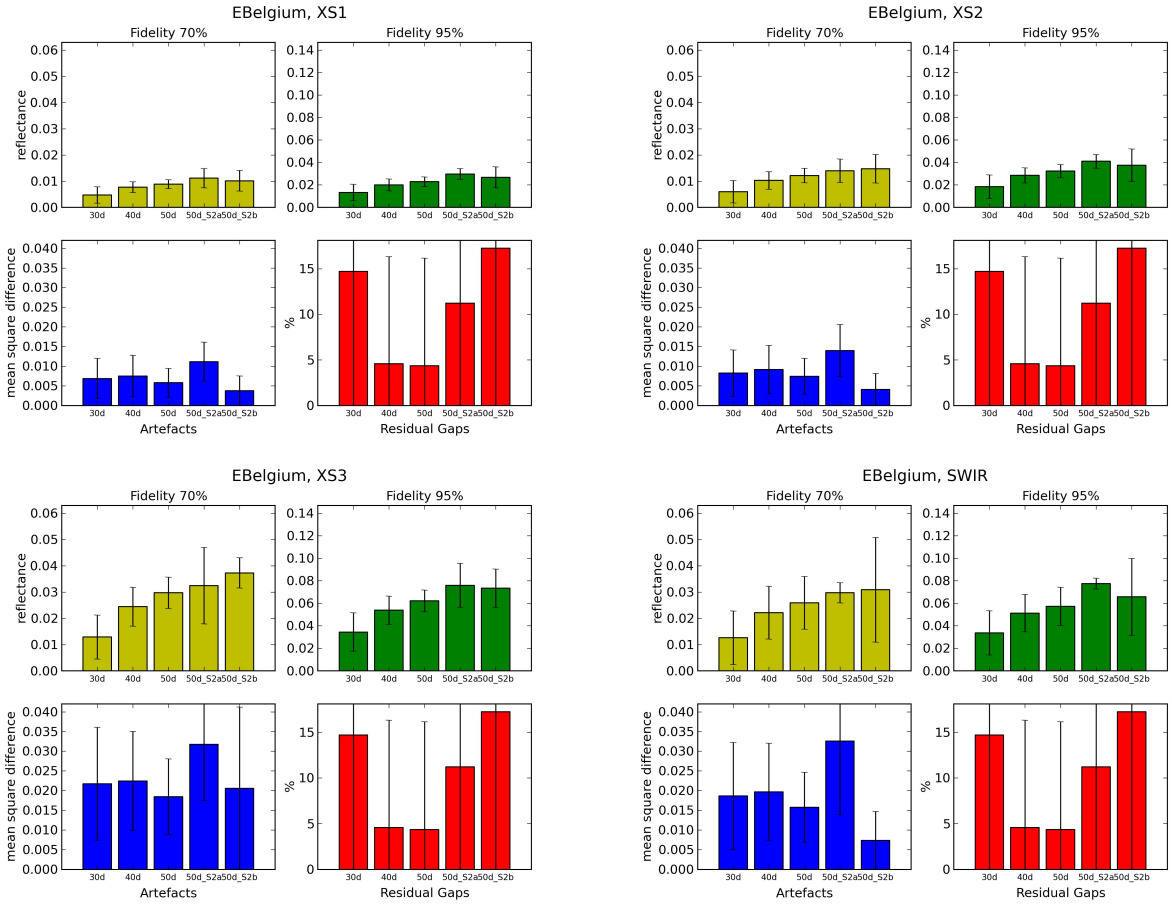

Figure 17: Summary of Benchmarking results for Belgium site

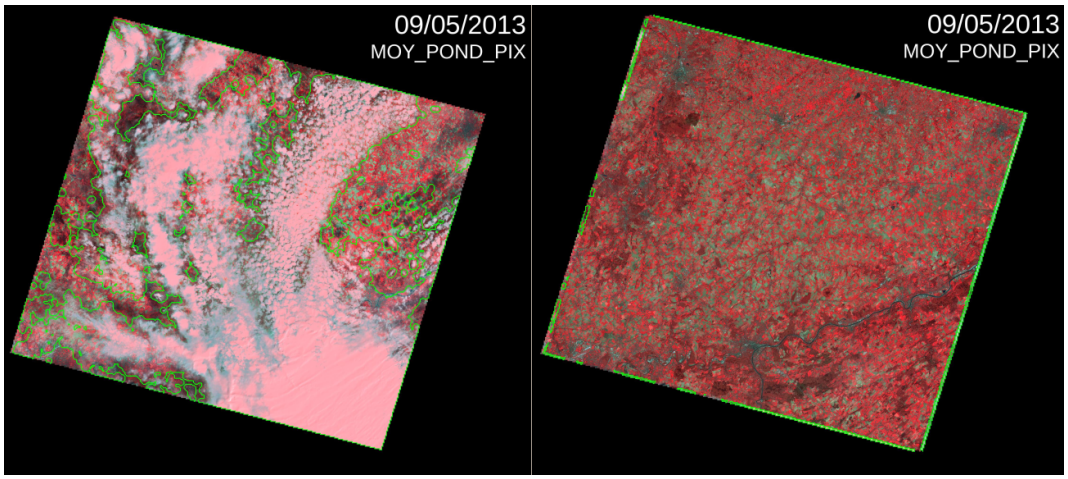

Figure 18: Comparison of two syntheses for the same date over Belgium, with a composite duration of 30 days (left), and 50 days (right)

5.2.9 Ukraine

The Ukraine site was covered by snow until April. After that, a sufficient amount of cloud free images were obtained providing good results. The best result for composite duration is 40 days (Figure 24), but this is due to the presence of snow during a large part of SPOT4 (Take5) experiment, that degrades more the 50 days composite, as snow is used in the composite for a longer time period, and because the criterion can be evaluated on the first synthesis while it is not evaluated for the 30 and 40 days syntheses, because nearly all the pixels are covered by snow . Because of snow melt, very steep variations of reflectances are observed resulting in high

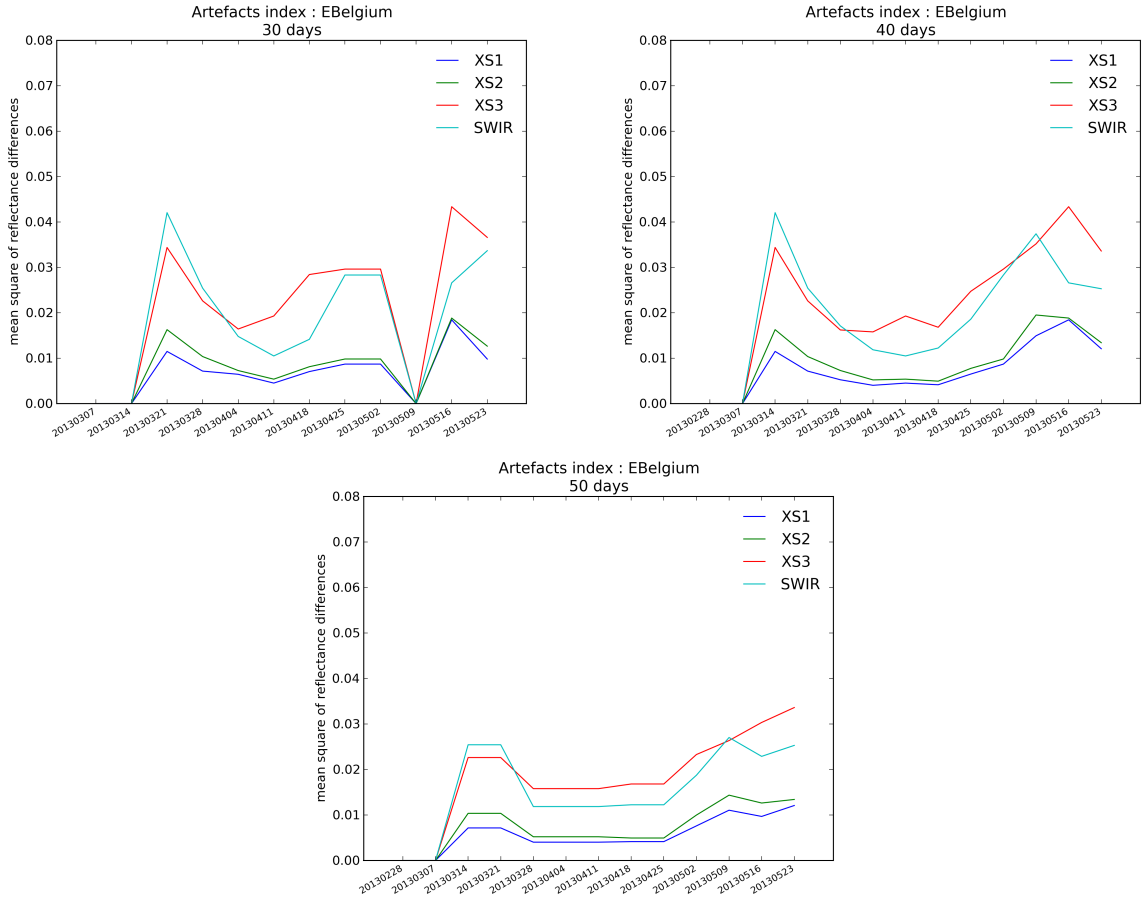

Figure 19: Summary of artefact criterion for Belgium site, for 30, 40 and 50 days respectively. The first 30 days synthesis appears one week later (March 21st instead of March 14th) and the 30 days synthesis on May 9th is made of only one image, and is not accounted in the artefact criterion. As a result, using 40 or 50 days syntheses indeed enhances the results

differences between the reference image and the composite. Snow is not accounted in the statistics, but some undetected snow within forests still can cause large errors and artifacts.

With a 10 days repetitivity, the quality of the syntheses is often largely degraded, and the amount of data gaps could be largely increased.

5.2.10 Sudmipy

A large decrease of artifacts and data gaps is observed with the length of compositing period, but a large decrease of infidelity, in a season marked by an exceptional bad weather in south west France and very rapid changes in vegetation cover. However, the balance would indicate a preference to use a compositing period of 50 days(Figure 26).

The performances observed with a repetitiviy of 50 days are also degraded with some of the syntheses having more than 30% or residual data gaps.

5.3 Nominal method with directional correction

This method is only applicable to the SPOT4 (Take5) sites which were observed from two different viewing angles, i.e. Maricopa and Sudmipy Three results are compared :

- the composite using only one viewing direction, the most vertical one, labelled “1 view”
- the composite using both viewing directions, without directional correction, labelled “2 views”
- the composite using both viewing directions, with directional correction, labelled “2 $\text{views}_{\text{corr}}$”

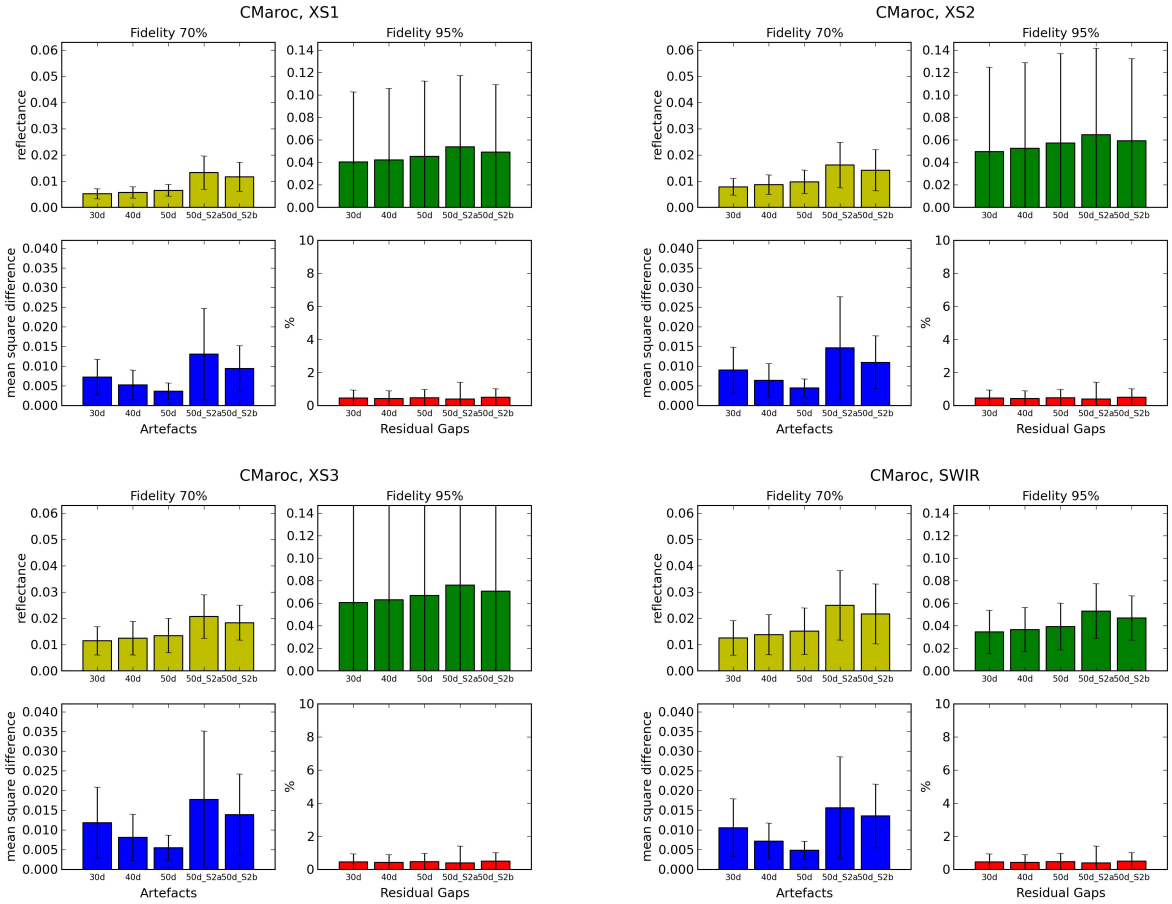

Figure 20: Summary of Benchmarking results for Morocco site

For the fidelity criterion, the reference images are picked only in the most vertical viewing direction. In the third case, with directional correction, the reference image is also corrected for directional effects.

5.3.1 Maricopa

The results (Figure 27) show that the use of directional correction improves the performances both in temps of fidelity and artifacts. However, the fidelity obtained after directional correction is not as good as the one obtained with one view. This is not surprising as the images used as input in the cases with tow viewing angles are much more different from the reference image. In fact, the composite products are already quite good with only one viewing direction.

On this site, with a lot of bare soils and therefore low directional effects, syntheses without directional correction are already extremely good, as they average around 15 dates : see 5.3.2. Difficult to do better, but the directional correction still moderately enhances the results.

5.3.2 Sudmipy

Due to the large size of the site, compared to the small size of the intersection of footprints, comparing the performances with and without directional correction is difficult. Given that the syntheses with two viewing directions can only be compared to a reference from one viewing direction, the proportion of the intersection zone is quite small. We could compare the performances only for the intersection zone, but this zone moves from one image to the other, as the footprints of SPOT images are not constant. It will be tested later on.

This difficulty results (Figure 29) in an unsignificant impact of directional correction with regard to the fidelity criterion. A small improvement is noted in terms of artifacts. However, the improvement is largely visible in the quicklooks, as it may be seen if figure 5.3.2

5.4 Nominal method with gap filling

TBD

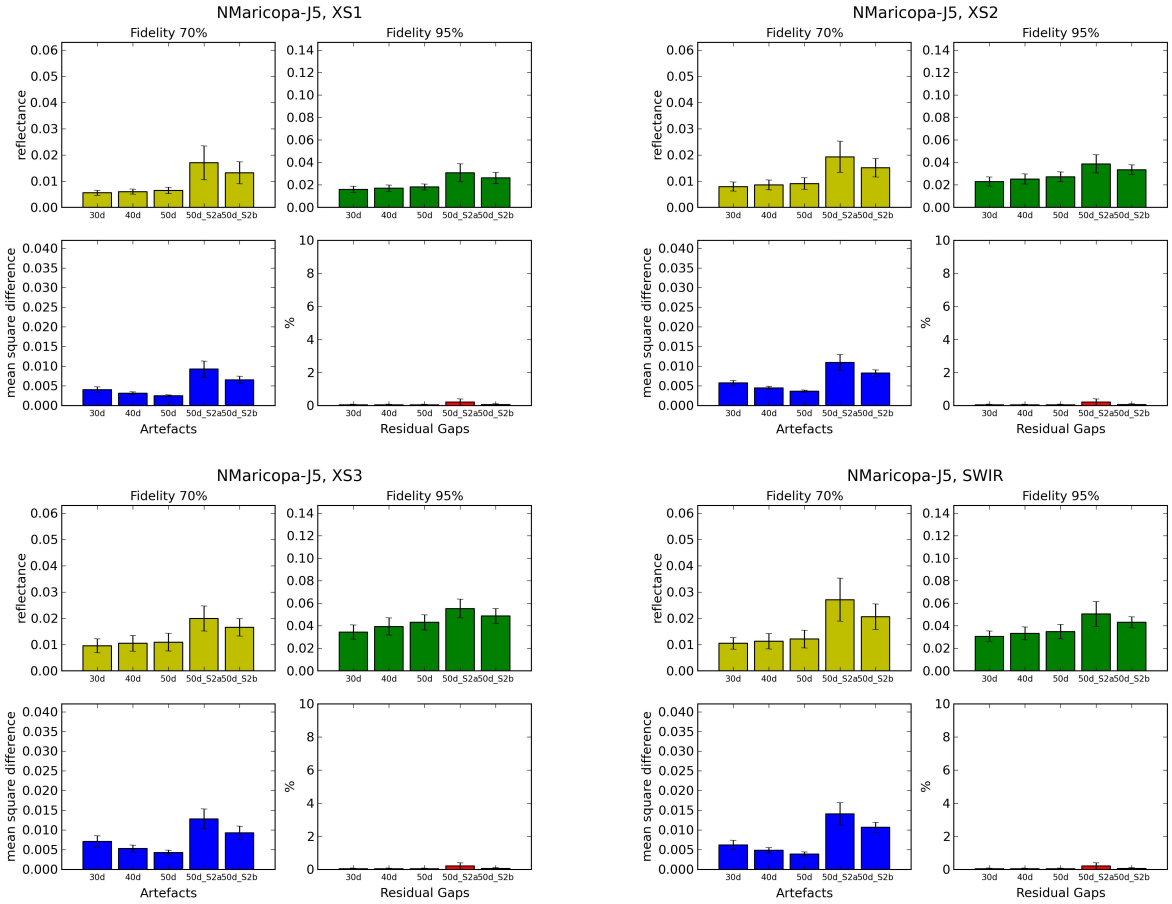

Figure 21: Summary of Benchmarking results for Maricopa site

6 Synthesis

After a preliminary benchmarking phase that showed the better performances of Weighted Average method compared to Best Pixel methods. As a result, the benchmarking concentrated on various variants of the Weighted Average method.

The following table summarizes the best compositing period obtained for each site. Great variations are observed as a function of the site and the weather observed during the period of the experiment. A value of 40 could be a good compromise, however, as a tuning is always possible depending of the country processed, it advisable to start with a baseline value of 30 days, but a possibility to tune it.

Site	Best compositing period	Perf with 1 sat
Madagascar	50	-
South Africa	30	++
Argentina	30	+
Belgium	40	–
Morocco	30	+
USA-Maricopa	30	++
Ukraine	(snow) 40	-
China	50	–
France Sudmipy	50	-

We have also tested performances of the method with only one Sentinel-2 satellite, with the addition of LANDSAT 8 in April, based on 50 days composite. It was found out that the sites noted negatively in the table are providing bad quality composites for some dates. LANDSAT images can be included in the syntheses.

The directional correction, tested on two sites, enhances the performances. Although this will have to be tested further, and other directional correction models could be tested, it is confirmed that we can rely on a constant model which does not depend on the site. It is therefore proposed to implement this in the Sen2Agri project. Level 2A products will be preprocessed for directional correction, to be used as input of the level 3A processor.

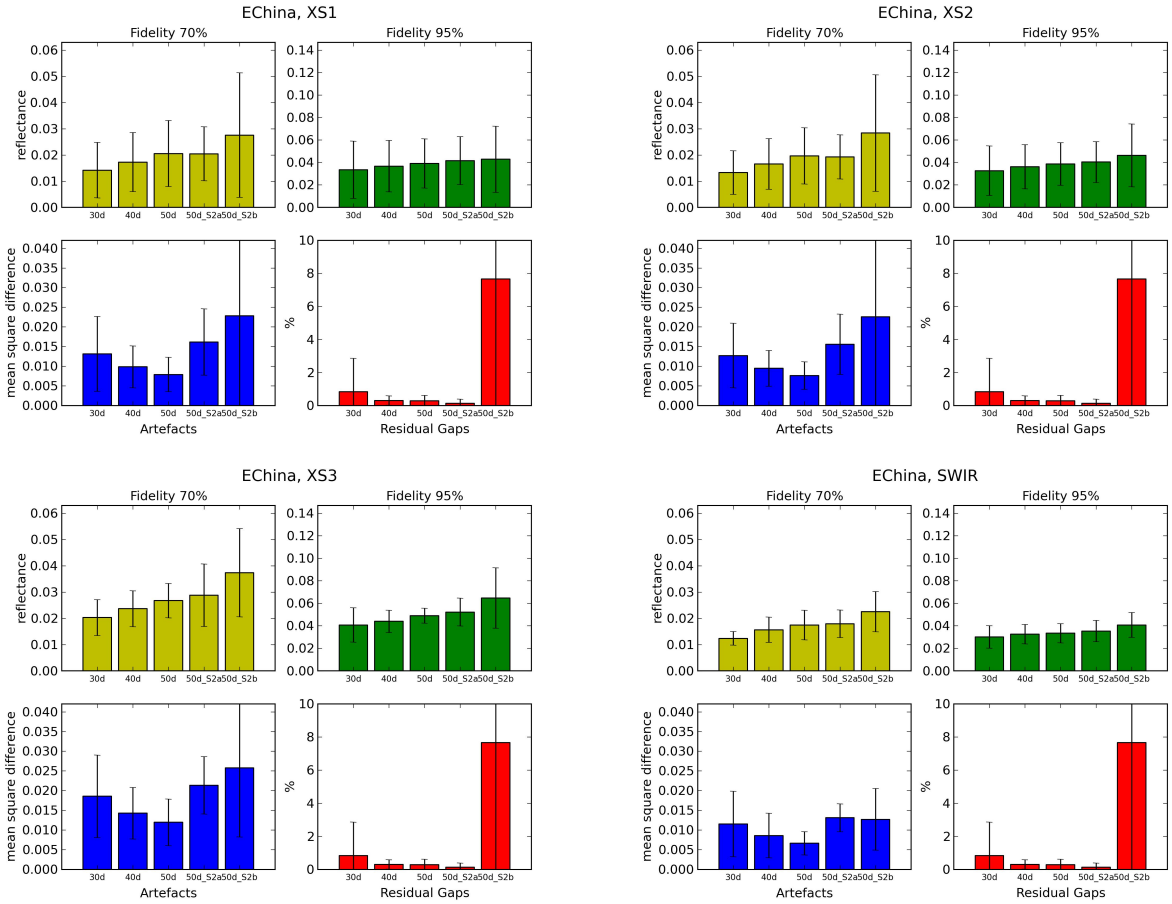

Figure 22: Summary of Benchmarking results for China site

The gap filling can be useful although its results are not perfect. But its implementation in real time is not possible as it needs an image after the data gap. It is proposed to implement it as a post processing after the level 3A processor in the toolbox, and to let the users decide to use it or not.

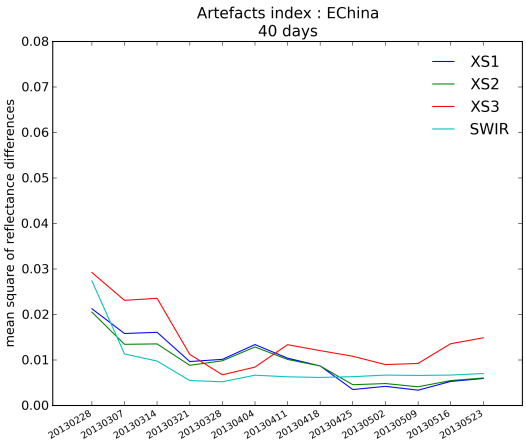

Figure 23: Artefact criterion for 40 days syntheses as a function of time, showing the degraded performances for the first syntheses

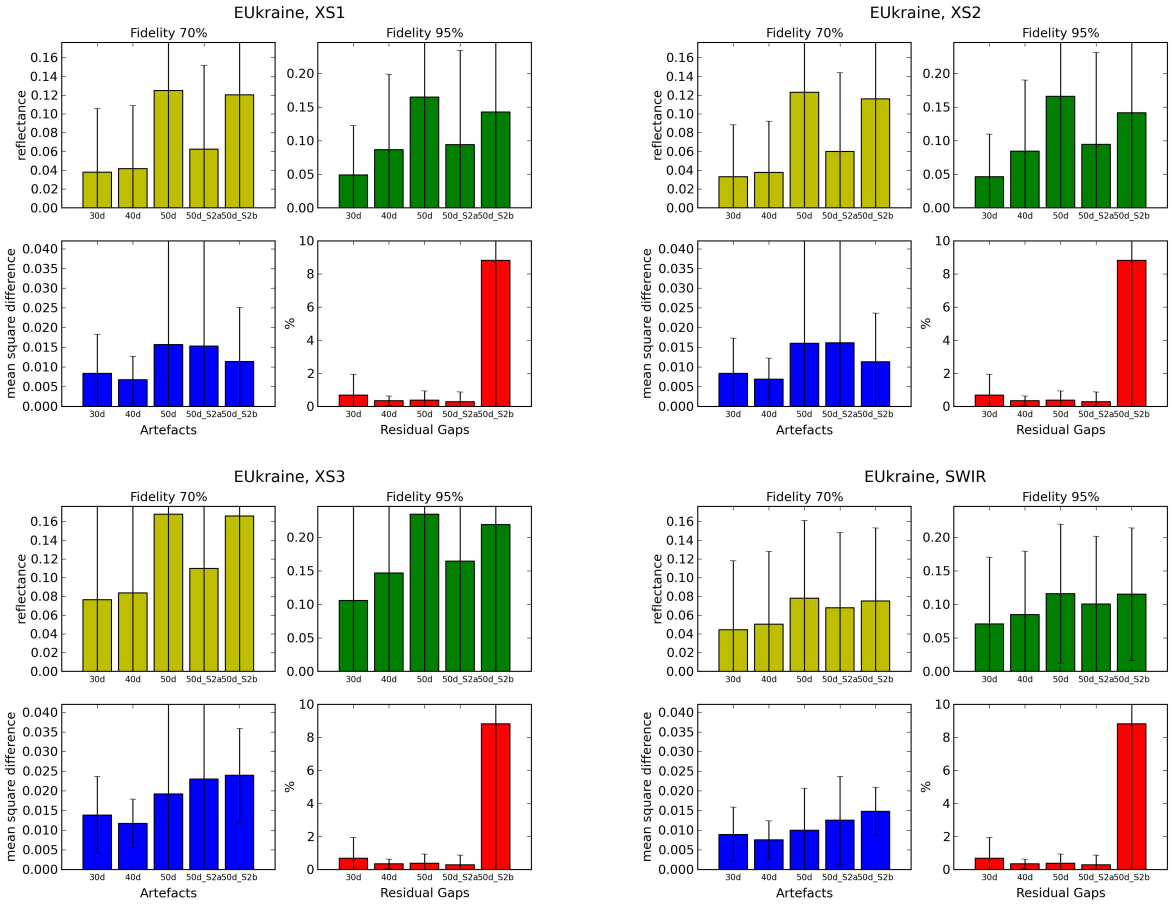

Figure 24: Summary of Benchmarking results for Ukraine site

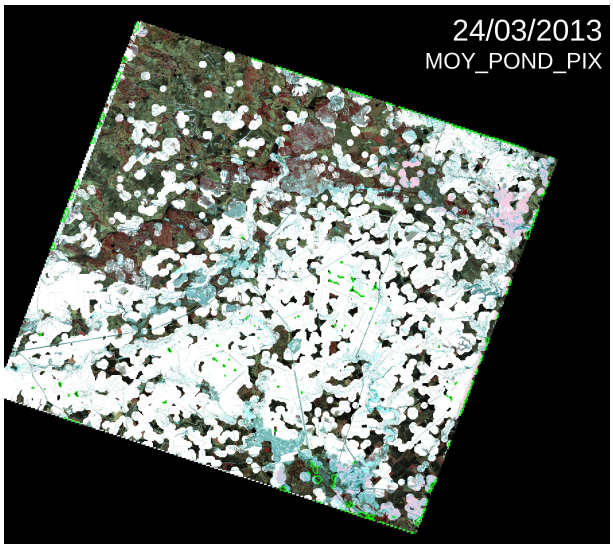

Figure 25: synthesis during snow melt period

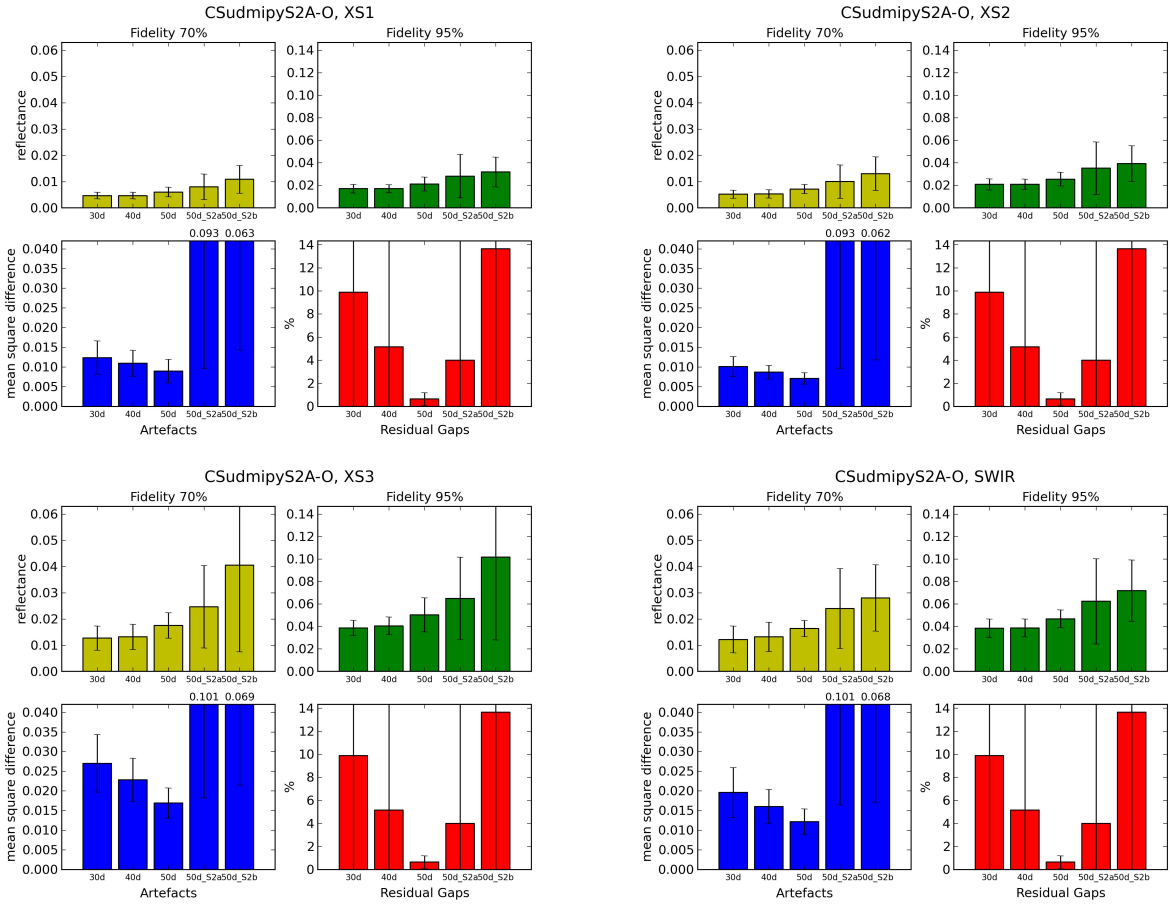

Figure 26: Summary of Benchmarking results for Sudmipy site

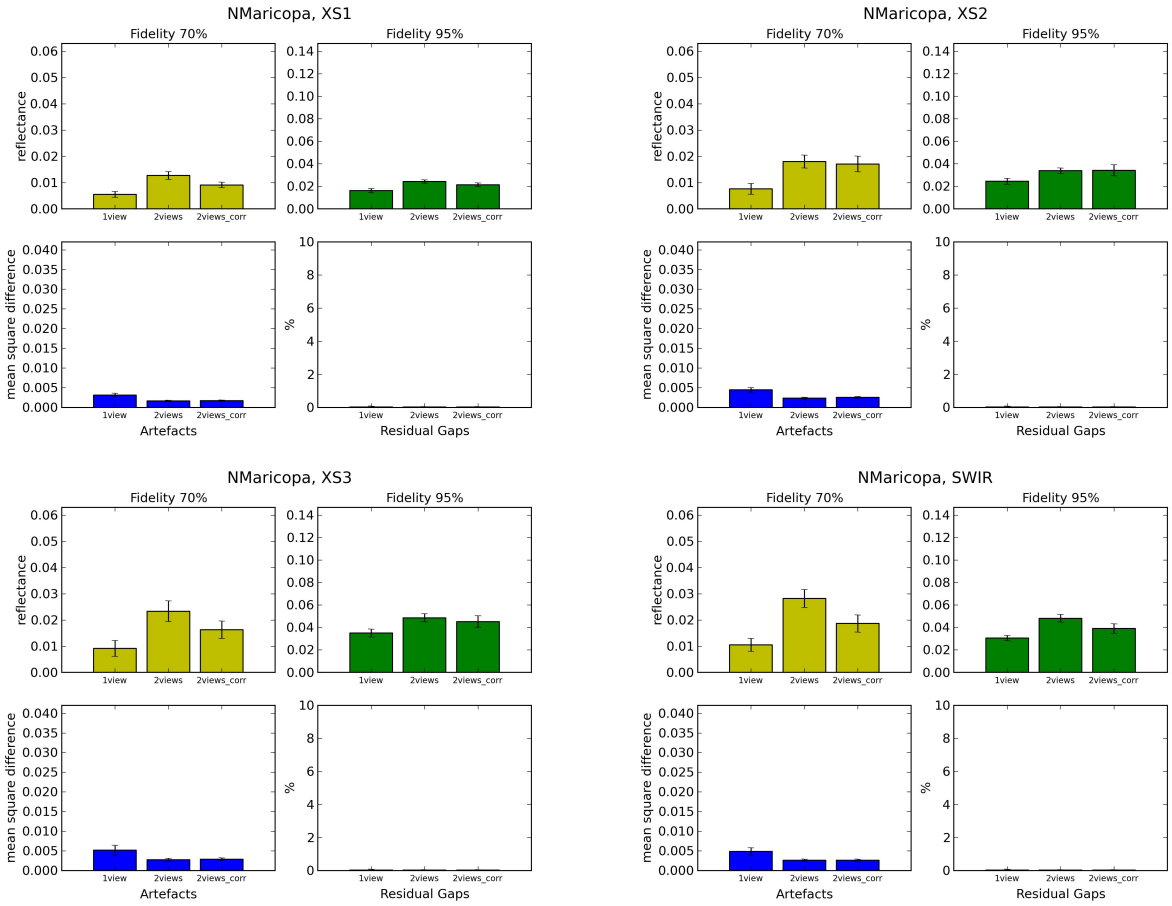

Figure 27: Summary of Benchmarking results with and without directional correction for Maricopa site

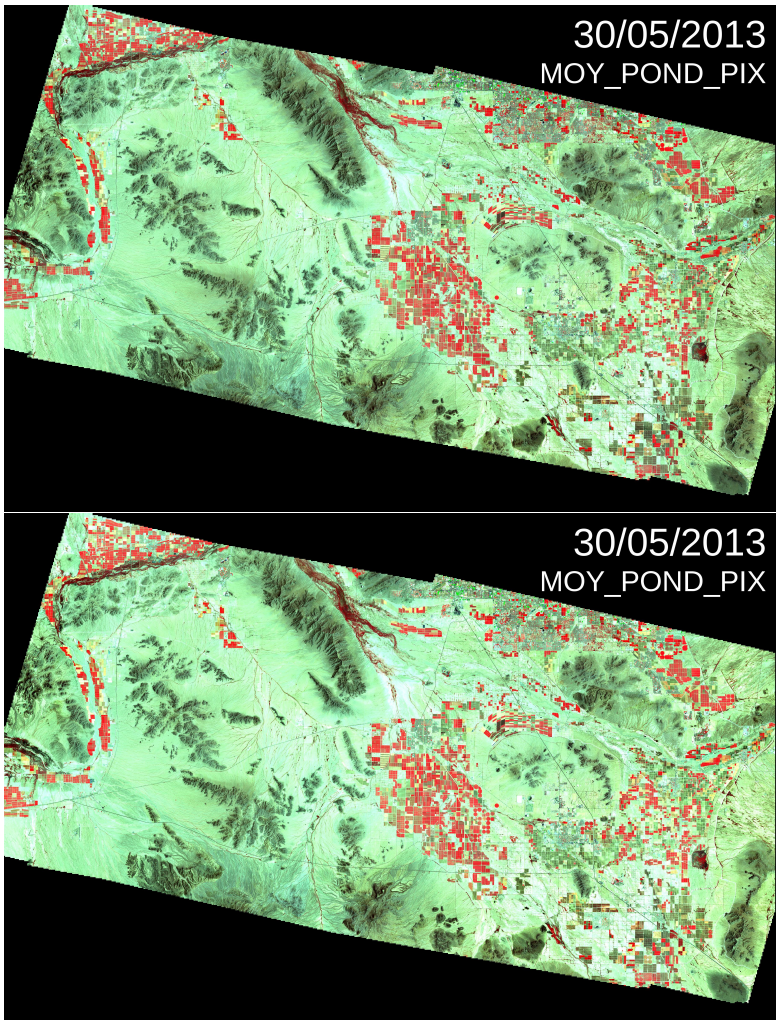

Figure 28: Maricopa. Top : without directional correction, Bottom : with directional correction

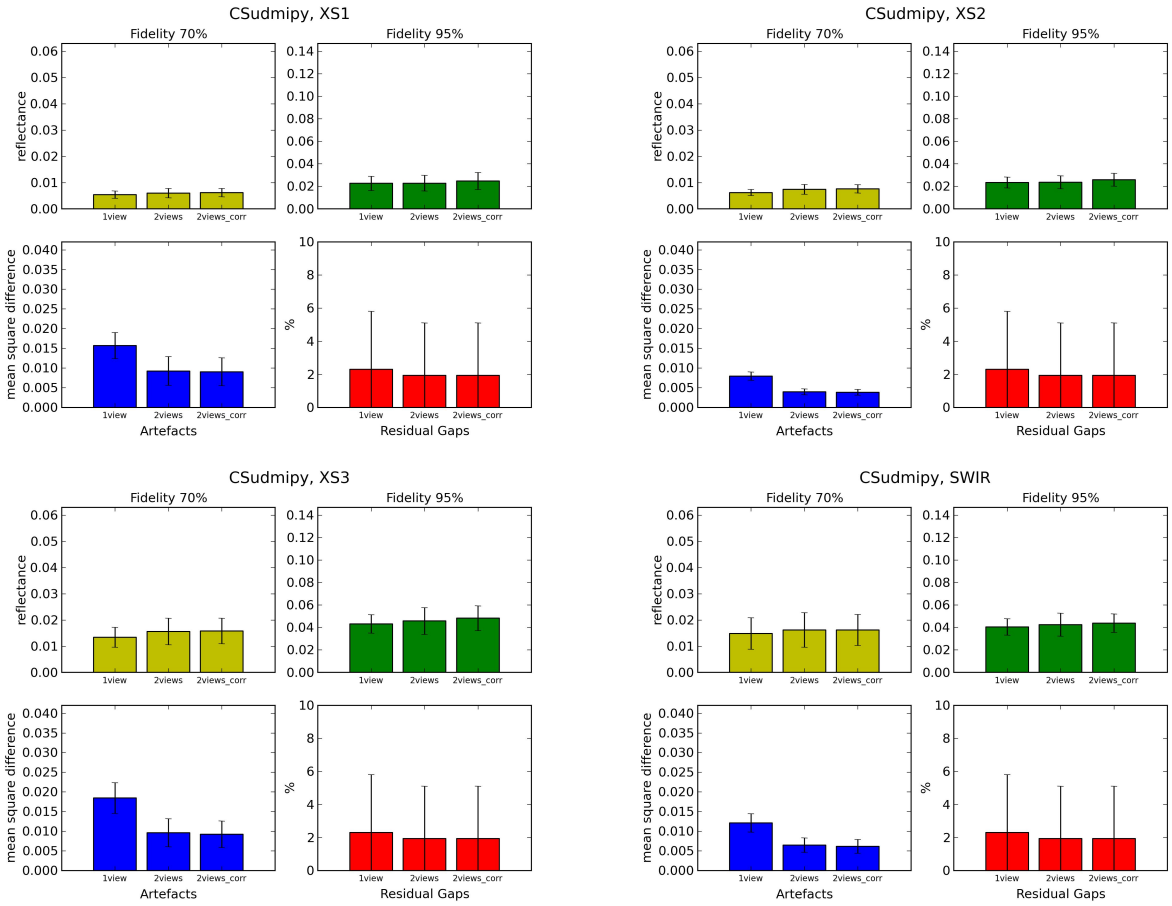

Figure 29: Summary of Benchmarking results with and without directional correction for Sudmipy site

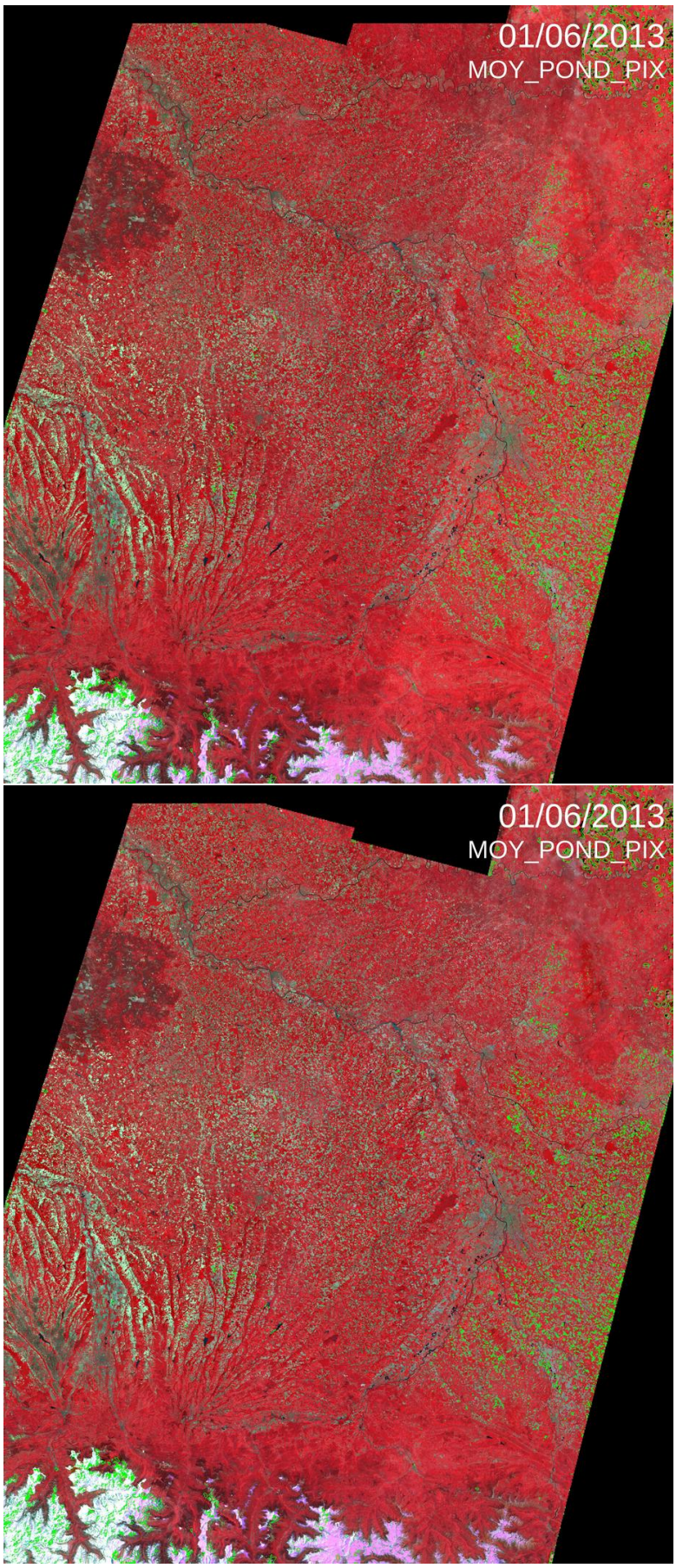

Figure 30: Sudmipy. Top : without directional correction, Bottom : with directional correction