Multitemporal Fusion for the Detection of Static Spatial Patterns in Multispectral Satellite Images--with Application to Archaeological Survey

Multitemporal fusion for the detection of static spatial patterns in multispectral satellite images – with application to archaeological survey

Bjoern H. Menze[1,2], Jason A. Ur[2]

[1] Institute for Advanced Study and Department of Computer Science, Technische Universität München, Munich, Germany
[2] Department of Anthropology, Harvard University, Cambridge MA, USA

Abstract—We evaluate and further develop a multitemporal fusion strategy that we use to detect the location of ancient settlement sites in the Near East and to map their distribution, a spatial pattern that remains static over time. For each ASTER images that has been acquired in our survey area in north-eastern Syria, we use a pattern classification strategy to map locations with a multispectral signal similar to the one from (few) known archaeological sites nearby. We obtain maps indicating the presence of anthrosol – soils that formed in the location of ancient settlements and that have a distinct spectral pattern under certain environmental conditions – and find that pooling the probability maps from all available time points reduces the variance of the spatial anthrosol pattern significantly. Removing biased classification maps – i.e. those that rank last when comparing the probability maps with the (limited) ground truth we have – reduces the overall prediction error even further, and we estimate optimal weights for each image using a non-negative least squares regression strategy. The ranking and pooling strategy approach we propose in this study shows a significant improvement over the plain averaging of anthrosol probability maps that we used in an earlier attempt to map archaeological sites in a 20 000 km^2 area in northern Mesopotamia, and we expect it to work well in other surveying tasks that aim at mapping static surface patterns with limited ground truth in long series of multispectral images.

Index Terms—Archaeological remote sensing, anthrosols, random forest, ensemble classification

1 Introduction

The analysis of spatio-temporal surface patterns is central to many applications in satellite remote sensing. In land use monitoring, for example, algorithms deal with the detection of specific changes of the land cover or the accurate quantification of urban growth [1]. Other algorithms measure, for example, growth of wild fires or shrinkage or ice sheets [2]. In all those cases the relevant information is the change of the image signal itself.

In the present paper we deal with the *opposite* case for an application in archaeological survey: we are interested in mapping surface patterns that remain *static* over time and want to tone down the effect of short-term variation which is – literally – covering the multispectral signal of the underlying soils. We also address the problem of *what* image to choose for the depth analysis of an archaeological landscape – a prominent issue in archaeological remote sensing where images from past satellite missions may show the structure of interest better than recently acquired imagery – and further develop a multitemporal fusion strategy which avoids this problem by jointly analysing *all* relevant multispectral images that are available for an area under study [3].

While our primary motivation to further develop our multitemporal fusion strategy from [3] is to improve settlement survey in the Near East, it may also be relevant for other detection tasks when some partial ground truth for training image-specific classifiers is available. These may be other archaeological applications [4], [5], [6], [7], [8] when the structures of interest are directly visible in multispectral images – and their reflectance or radiance differs from their natural surrounding – or when structures have at least an indirect imprint on the spectral signal – and the archaeological structures in the ground or underneath plant cover impact positively or negatively on the vegetation on top. Beyond archaeology, our approach may also be of interest for other detection task where the spatial distributions of the structures of interest are, over the observational period, essentially static and where eliminating spurious variability – such as seasonal changes in crop-cover, time of overflight and solar altitude, impact of short-term meteorological events like rain or snow-cover – will improve the underlying signal. This may be relevant, for example, in geological prospection when mapping mineral deposits [9], [10] or in environmental research when characterizing the surface cover for wildfire [11], [12] or resource management [13].

Accepted for publication in IEEE JSTARS 2014 (IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing); http://dx.doi.org/10.1109/JSTARS.2014.2332492

1.1 Prior work

The past years have seen a raised interest for using spectral images in archaeological survey that have a resolution of decimeters to meters and dozens to hundreds of spectral channels [14], [15]. When trading spatial against spectral resolution, however, most detection and mapping tasks in archaeological remote sensing still opt for high spatial detail and a low number of spectral bands. Many structures of archaeological interest are in the sub-meter range which is at the expensive end of highly resolved multi- and hyper-spectral sensors [4], [5], [6]. At the same time, it is often difficult to know in advance whether the archaeological matrix of interest will have a multispectral signal that is distinct from its surrounding, i.e., not before images have been acquired, processed, and analysed [4], [5], [16], [17], [8]. Here, the use of satellite images with high spatial resolution and wide coverage (but very few spectral channels) is less risky. Such image may be accessed for free on the internet [18] or can be bought at rather low costs from standard commercial satellite imaging services. Unfortunately, they come with the problem that environmental conditions – such as vegetational period, crop cover, soil moisture – may affect the visibility of the desired structures and, ideally, images from multiple acquisition time points should be studied. To address this problem we proposed in [3] to gather and process *all* relevant images from research satellite missions that are available for a certain area of interest, using – for example – Landsat or ASTER data that have a long observation record. Fusing the probabilistic maps generated from ASTER images of multiple time points showed to significantly improve the detection results in our survey task.

Multitemporal image fusion is by itself a longstanding topic in satellite remote sensing. Information from multiple images covering the same scene can be fused at the feature level, but calibrating intensities of images acquired at different time points is difficult [19], [20], as different noise processes overlap [21]. It is computationally expensive [22], [23] and requires, ideally, some knowledge about the sensor [20]. Thus, a typical approach is to extract the information of interest in a first step, for example following a pattern classification approach, and then to fuse the information across observations in a second step, for example, by averaging the probabilistic maps or by assigning the vote of the majority of the observations [24], [25], [26], [27], [28]. This post-classification is more robust as every image can be processed with its individually adapted classifier, and variation between images might be removed at the classifier level [3], [29]. As not every image necessarily contributes the same information, it may desirable to rate the quality of the images, for example by learning weights that are assigned to the individual observation when averaging all votes. This replaces the basic voting, or averaging, by another level of (linear) pattern classification. Such hierarchical models can use arbitrary fusion schemes at the second level, for example, learning neural networks, linear models, or even non-linear classifiers [30], [1], [26], [31]. These approaches, however, require that for every pixel the same set of image observations are available: they are not applicable in situations where different parts of the regions of interest are covered by different numbers of satellite images. As a consequence, they cannot be applied to large areas with irregular coverage. Here, a simple voting scheme that can use an arbitrary number of image observations as input remains the preferable fusion approach.

1.2 Contribution of this paper

In this paper we further develop multitemporal fusion schemes that can deal with arbitrary numbers of multispectral images when limited ground control is available. We will address the task of mapping anthropogenic soils in fluvial landscapes of the Near East, using limited ground truth from archaeological survey and the visual interpretation of mono-chrome high resolution satellite images. In this we build on prior work from [3], improving fusion statistics and introducing a sampling approach to identify locally optimal subsets for fusion.

In the following we describe in image and survey data, detection task, and the specific application in archaeological remote sensing (Sec. 2.1), recalling some results from [3], and propose new fusion strategies. Then, we will perform three experiments: 1) to identify the optimal fusion statistic for multiple observation in the given classification task (Sec. 3.1), 2) analyze bias and variance of the multi-temporal fusion process in order to to understand how pooling affects the quality of the fused data product (Sec. 3.2), and 3) we will provide an approach to choose optimal image subsets for pooling (Sec. 3.3), before we discuss properties of optimal fusion approaches and implications for archaeological remote sensing (Sec. 4).

2 Data and methods

An overview of the general processing pipeline with data set generation, image classification, and fusion of the probabilistic maps is given in Table 1.

2.1 Classification task and data sets

Detecting anthrosols: Our pattern classification task is the identification of "anthrosols" within the in-situ soils of an alluvial plain in Northern Mesopotamia that are visible in multi-spectral imagery [33], [16], [3]. These anthrosols are anthopogenic soils that developed over millennia from the eroding debris of human settlements, the remains of mud-brick based architecture. The spatial distribution of these sites provides insights into 9000 years of settlement history [34]. Limited information about the presence of ancient settlements is available for most regions in which anthrosols might be expected: Major sites are known from archaeological survey, larger

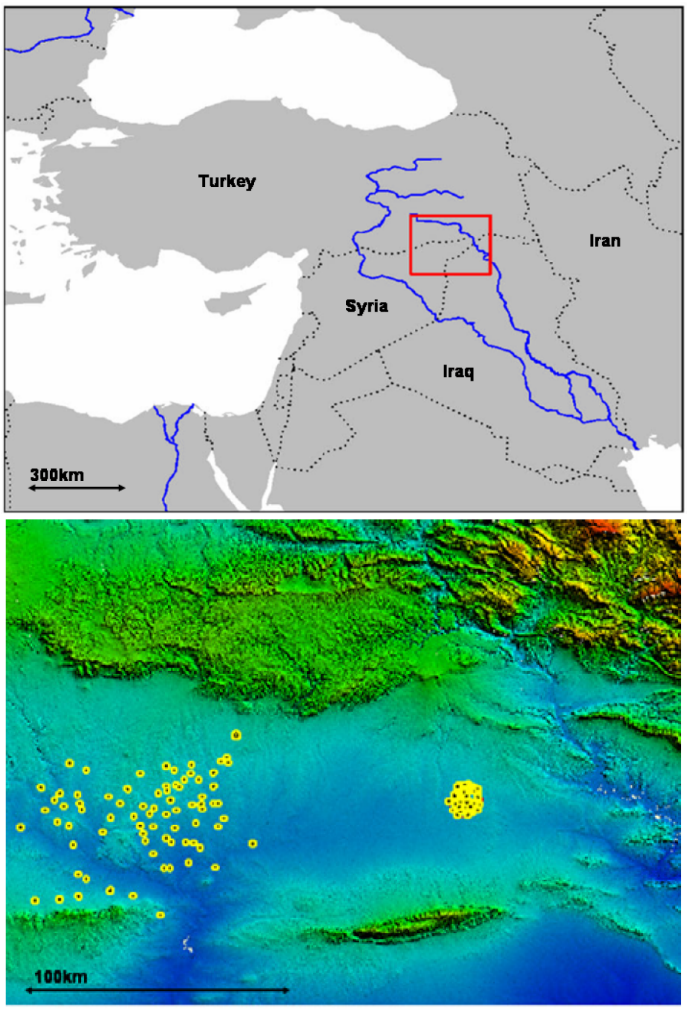

Fig. 1. Region under study. The Upper Khabur basin is situated in northern Mesopotamia, in the northeastern part of Syria (left image). Two areas in the central and the eastern parts of the Khabur plain served as testing grounds for the development of the fusion algorithms (right image). Ground truth was available from the analysis of declassified CORONA imagery (rectangular area, west; archaeological sites indicated yellow), and from an archaeological ground survey (circular area, east).

settlement "mounds" can even be localized in digital elevation models [35], and low-mounded sites can be identified from their lighter soils in an inspection of high resolution images (CORONA, Ikonos) [32], [36].

As a consequence, the primary objective for the mapping of anthrosols is the *extrapolation* of information from small ground-truthed area over the whole of a region of interest, or the *interpolation* of similar information using recordings of few clearly visible mounds from the top of the settlement hierarchy in order to locate smaller sites in between. We can address the detection task by transforming it into to a binary pattern classification task – anthrosol against all other surface classes (Sec. 2.2). However, both "anthrosols" and non-anthrosol "background" locations may regularly be covered by vegetation and crops during the course of the year, and agricultural field work and soil moisture may lead to

TABLE 1
Overview over the processing pipeline as proposed in [3] and in the present study, and as described in Section 2.

1. **Generate "partial ground truth".** Interpret CORONA images as in [32] and do:
   - register CORONA images with SPOT base map using landmarks,
   - localize likely settlement sites in CORONA images, include information from previous surveys or other image sources where available.
2. **Generate anthrosol probability maps.** For each ASTER image do:
   - register ASTER image with SPOT base map using landmarks (Sec. 2.1),
   - learn a classifiers individually for each ASTER image, as follows (Sec. 2.2):
     - extract spectral features,
     - extract training samples for foreground ("settlement sites") and background class (everything else) from the partial ground truth that is within the field of view of the ASTER image,
     - apply the classifier in a leave-one-out cross-validation that iterates over spatial blocks of about 6km*6km to obtain test errors for the anthrosol probability map.
3. **Fuse anthrosol probability maps.** For each region of interest, or each spatial block in the SPOT image, e.g. of size 5km*5km, do (Sec. 2.3):
   - identify anthrosol probability maps with (partial) overlap,
   - measure how well foreground and background are separated in each map from the partial ground truth and calculate, for example, the AUC ROC as quality score,
   - determine the weights that are associated with the different quality scores (Fig. 10),
   - sum over the weighted anthrosol probability maps.

significant changes in the spectral reflectance of both classes [3]. As multispectral images for sensors such as ASTER and Landsat have become available for long observational periods, this leads to the additional task of finding those images that have been acquired under favourable environmental conditions.

Multi-temporal data sets: We use data from two different areas in the Khabur plain in northeastern Syrian (Fig. 1). For the first western region (about 60*60 km^2 in area), a total of 243 ancient settlement sites (Fig. 3) were recognized from high-resolution CORONA satellite imagery [37], [36][1]. One may expect that ground survey would recover additional sites in this area. A total of 71 multispectral ASTER images from the time of 2002-2007, acquired during all seasons of the year, have partial overlap with this region (Fig. 3), with 32 to 47 observations for each of the 243 archaeological sites. The second region, situated some 150 km east of the first one, is significantly smaller in area – 125 km^2 – and has 60 known settlement sites [37]. For this

1. Available from http://hdl.handle.net/1902.1/14011.

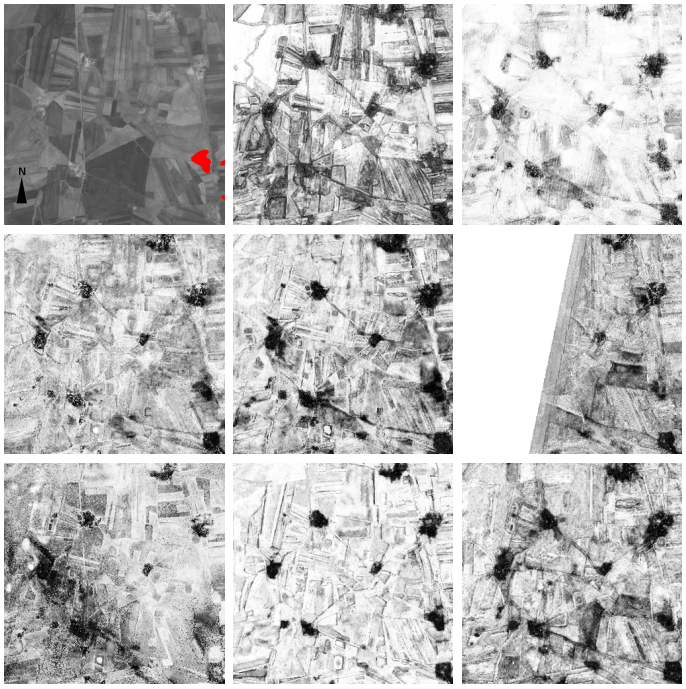

Fig. 2. Segment of the SPOT base map that was used for registering all ASTER images (top left), together with archaeological sites identified in CORONA images (red) that are used to re-train the classifier for each ASTER image. All other images show intensities from the visible red ASTER channel (ASTER band 2) that has been used for registering SPOT and ASTER. Image intensities in this channel varying significantly in between observations, one image has only partial coverage. (The shown segment is 5km*5km in size, its northwestern corner is at 36.9348 latitude, 41.24572 longitude.)

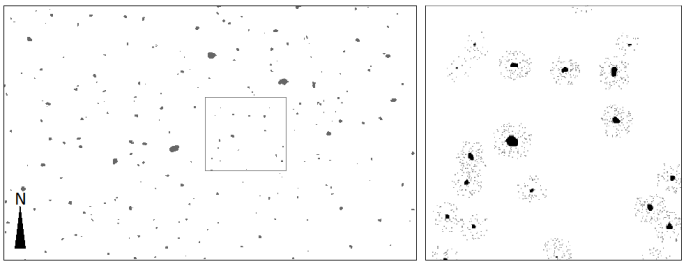

Fig. 3. The nearly random distribution of the 243 test sites in the western test area. Archaeological sites were previously identified from high resolution imagery. For each the "anthrosol" pixels of a site, an equal number of "'background" pixels were sampled from its direct vicinity (gray dots, right image). Outline of left image approx. 20*30km, right image 6*6km (compare gray box in left image).

region, however, ground truth was obtained from an archaeological field survey. A total of 19 ASTER scenes with 15-17 observations for each of the 60 sites was available for this eastern region.

For both regions anthrosol sites range between 0.5ha to 100ha in size, with most sites being in between 2 and 10ha (25% and 75% quantile). The spatial resolution of the 13 spectral channels of the ASTER sensor varies between 15m (for visible and near-infrared) and 90m (for thermal infrared channels) and individual sites are covered by > 20 pixels. We register all ASTER images to a common SPOT base image with 10m*10m resolution that was previously also used for co-registering the CORONA images [32] (Fig. 2). We did this by manually identifying landmarks that appeared in both the SPOT image and in the *red* channel of the ASTER image – typically using road crossings or boundaries between crop fields, occasionally surface features, such as clearly visible settlement mounds or wadis. We determined 10-20 such points for each ASTER image and used a cubic transformation model to estimate a dense displacement field. We interpolated all channels of the ASTER image to the resolution of the base image using a nearest-neighbour approach. This overall procedure led to a registration error between ASTER images that was below four pixels, i.e., 40m.

We then sample locations for both "anthrosol" and "background" sites in the first region to build our training data sets (Fig. 3). We sample up to 100 random locations, or pixels, from each "anthrosol" site of the first region, and same amount of pixels from the direct vicinity of the site to represent the "non-anthrosol" or "background" class. Between pixels of both classes we keep a distance that is somewhat larger than the registration error (4 pixels) and the extensions of a TIR pixel (90m, 6 pixels). Overall, this results in 35494 pixels (17747/17747) for the 243 + 60 sites of the two regions. These test locations were kept fixed in all evaluations.

About 155 ASTER images have partial overlap with our test regions. Using the observations from our test locations, we train a probabilistic model for each image and apply it to the full image in a spatially blocked cross validation similar to our approach in [3], as described in the next section.

2.2 Classification of individual ASTER images

Features.: Random forest is able to cope with a high number of features. So, we use the original spectral reflectances without further normalization, a total of 13 features for each pixel. Reflectances are subject to intensity variations depending on surface cover, but also depending on viewing angle between camera and sun that either add random offsets to the whole spectral signal, or random multiplicative factors, or both (Fig. 2). We try to remove the influence of these global noise processes – both within an image and in between images – by calculating "vegetation indices" that represent differently normalized intensities. Using ASTER band 2 as visible red (red) and band 6 for near infrared (NIR), we calculate for each pixel the "Difference Vegetation Index" $DVI = NIR - Red$, the "Ratio Vegetation Index" $RVI = NIR/Red$, and the "Normalized Difference Vegetation Index" $NDVI = (NIR - Red)/(NIR + Red)$. Finally, we correlate reflectances with template spectra from the JPL ASTER SpecLib [2] that we generate by

2. http://speclib.jpl.nasa.gov/

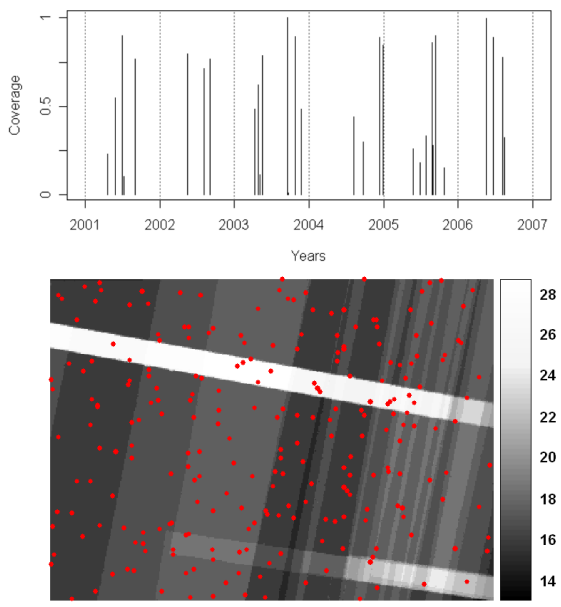

Fig. 4. Temporal and spatial coverage. A total of 155 ASTER images were available for the years 2001-2008 (top), mostly acquired during the dry-season. Single areas of the basin were covered by 4-43 images, with a maximum in the western test region (bottom, compare Fig. 1). Positions of archaeological sites (red) had been recorded from the analysis of declassified CORONA imagery or in archaeological field studies.

subsampling the signals to the 13 spectral bands of the ASTER sensor and by grouping different multiple signals of the "manmade", "minerals", "rocks", "soil", "vegetation", and "water" class. This leads to six correlation coefficients that indicate whether one of the six classes is dominating the observed spectral pattern. In total, we obtain 23 features for each pixel (Fig. 5).

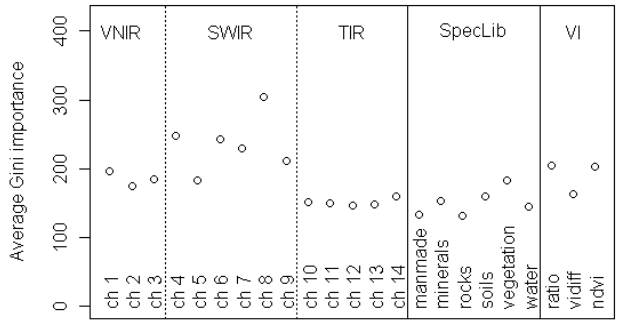

Fig. 5. Feature importance as calculated from the random forest. Most of the information is in the short-wave infrared (SWIR) and visible/near IR (VNIR). Long-wave IR (TIR), ASTER SpecLib correlation features, and vegetation indices (VI) only contribute to a lesser extend to the classification.

Random forest classifier: In first comparisons with linear classifiers (regularized linear discriminant analysis) we had observed an advantage of non-linear classification methods, and a comparison of the random forest classifier with RBF-kernel support vector machines had indicated a slight advantage of the first [38], [39]. So we model the posterior probability using random forests, a non-parametric ensemble classifier that relies on randomized decision trees as base learners. This classifier averages the decisions of many unbiased but highly variable decision trees that have been generated by using different subset of the data during training ("bootstrapping") and by randomizing the feature subspace when searching for the best split at every node of the tree ("random subspaces"). In general, the random forest classifier is capable of dealing with few samples in high dimensional spaces, a property that makes it an attractive learning algorithm in the classification of spectral data [40]. In fact, it was adopted early for spectral classification tasks also in remote sensing [41] and in surface classification [42], [43]. The random forest algorithm generates feature relevance scores that can be used for visualizing relevant features – we show the relevance of the features used here in Fig. 5 – or for feature selection and dimension reduction [44], [31]. A random forest ensemble has few parameters to be optimized, so it can be trained very fast and, hence, is well suited for repeated classification of an image in cross-validation experiments. It takes discrete class labels as input and returns continuous probabilities, as the averaged votes of the decision tree ensemble model the posterior probability of the input classes. We make use of this property when mapping the anthrosol probabilities. In the present study we chose random forests with univariate split functions in the node. Initial tests with "oblique random forests"[3] [45] indicate that using random forests with multivariate split models have advantages – as these node models are better capable of dealing with correlation between features, a property of most spectral data sets [40] – but we leave the optimization of this aspect open to further studies.

We use Breiman and Cutler's original Fortran implementation as available for R[4]. We grow trees to full depth and have to set two parameters of the classifier: the number of trees N_{tree} in the ensemble, and the dimensionality of the random subspaces M_{try}. We test different parameterizations for M_{try} on a small subset of the training data evaluating the out-of-bag test error. While M_{try} is often the only model parameter parameter of the random forest algorithm, we here find that $M_{try} = 3$ performs well – a value close to the default recommendation (that is the square root of the number of features). In the evaluation of M_{try}, we generated ensembles with 300 trees and find that out-of-bag error typically converged for 100-150 trees. We keep $M_{try} = 3$ and $N_{tree} = 300$ fixed for all further experiments. As the number of training samples depends on the number of sites present in each ASTER image, the number of samples used for a classification ranged between less than 1000 and 45000, balanced for both classes. In case of very few training sites for the anthrosol class, additional

3. cran.r-project.org/package=obliqueRF
4. cran.r-project.org/package=randomForest

samples were drawn from other ASTER images acquired on the same day (if possible).

Anthrosol probability maps: The training data may have a significant number of false labels: Sites identified on the ground or in high resolution imagery may not always show the characteristic spectral pattern of soils transformed by anthropogenic activity, they may be geological or artificial surface features that are misinterpreted as settlement mounds or anthrosols. Similarly, sites identified in images acquired in the 1960s-70s, may have been destroyed in recent agricultural transformation of the landscape ("bulldozing" and "deep-plowing").

The random forest classifier is able to cope with a small amount of false labels. However, to prevent these deficiencies in the annotation from being propagated to the classification results and to have unbiased probabilistic maps as input to the fusion step, we choose a blocked spatial cross-validation strategy to apply the classifier [46]. Each ASTER image is separated into a grid of 36 subregions where each region has a side length of approx. 6km – well above the average correlation length of most structures of interest on the ground. Then, the random forest is trained using data from training locations of 35 subregions, and applied to all pixels of the one hold-out region. We iterate this leave-one-block-out classification over all 36 regions and obtain a dense probability map indicating the most likely location of anthrosols within the given ASTER image (e.g., Fig. 12, right column).

2.3 Multitemporal fusion for the detection of static spatial patterns

Averaging probability maps: As variable numbers of observations are available for every pixel, we use parametric noise models for summarizing the observed anthrosol probabilities. Parameters of the distributions can then be evaluated for their ability to separate "anthrosol" and background pixels. This approach can be considered to follow a generative modeling strategy: Given the unknown label that indicates presence or absence of anthrosols we have an observational model that – with different sets of model parameters for either "anthrosol" or "background" – is capable of generating an arbitrary number of observations. This generative "forward" model matches our averaging approach from [3] under the assumption of a Gaussian observation model, i.e., assuming a normal distribution for the noise in both classes.

Some multispectral images are acquired under ideal conditions, while others may contribute little more than noise. At best, averaging over such low quality maps will average out if noise is uncorrelated and many image are available, for example from local cloud cover or artifacts of the camera system that are unrelated to surface features. At worst, they will significantly bias results towards systematic errors which are in our application, for example, modern sites or geological features with light soils. So we may want to identify subsets of the available images that have the optimal contrast between anthrosol and surrounding, i.e., that have the least bias. At the same time, we want to pool over the maximum number of ASTER images that are available to remove image-specific noise for uncovering the static surface pattern we are interested in. This requires strategies for finding optimal subsets of probability maps that we want to average over.

In general, our image classification approach requires a retraining of the classifier for every multispectral image in order to cope with changing environmental conditions and changing spectral signature [3]. To this end, a limited number of archaeological sites have to be present in every image we want to use in our analysis. This, in turn, also allows us to measures how well individual satellite images reveal the surface features we want to map. We can order the images accordingly and test how well averages over different top ranking subsets perform and choosing the subset that minimizes, for example, the local least-squares fit error. Our experiments suggest, however, that such a crisp selection may lead to a selection of very few images (Section 3.2). In both test areas no more than 10% of the locally available images are combined, returning results that are very noisy and do not make full use of the available data. Also, we have to keep in mind that the same probability map may be have good contrast for one location while being less than optimal for other locations nearby, and we may not want to rely on approaches that follow a very aggressive selection strategy.

Averaging weighted subsets of the probability maps: As an alternative to the crisp threshold we may weight observations predict anthrosol locations from a weighted sum of all available probability maps. With K available probability maps from the ASTER images I^k ($k = 1 \dots K$), the predictions p_i^k of a fix subset of N pixels i with available labels ϑ_i, we seek for the optimal weights w^k that are obtained by minimizing a least squares criterion:

$$arg\min_{w} \sum_{i=1}^{N} (\vartheta_i - \frac{1}{K} \sum_{k=1}^{K} w^k \; p_i^k)^2. \tag{1}$$

To enforce that we do not *subtract* probabilities, we introduce nonnegativity constraints for image weights w_k

$$w^k \geq 0 \; \forall k = 1 \dots K, \tag{2}$$

leading to a standard non-negative least squares regression. Unfortunately, we cannot apply this standard approach directly, as most probabilistic maps only have *partial* overlap with our region of interest, and the set of the K observations varies locally (Fig. 3).

To this end, we make our weights w not dependent on the individual image, but on a more general image "quality" score q_i that we calculate for each image regardless of its localization. An example of this quality

score can be the average site-wise area under the curve of the receiver-operator-characteristic (AUC ROC), a non-parametric ranking measure of class separation that has a value of 0.5 in case of a random mixture of both classes and 1 for perfect separation. We substitute the map-specific w_k from Eq. 1 by function $W(q^j) = c_j$, with $j = 1 \dots D$ equally spaced intervals in the range q, and coefficients c_j that represents one common weight for all images with a quality score that match q_j. For a given prediction p_i^k from an image I^k with data quality $q(I^k) = c_j$ we obtain

$$W(q(I^k)) = \sum_{j=1}^{D} c_j \delta_{jk}, \tag{3}$$

where δ is the Kronecker delta δ_{jk}, that is equal to one (and equal to zero otherwise) when image I^k has has the global quality score q_j and, hence, is assigned weight c_j. We can look at the second term in eq. 1

$$\frac{1}{K}\sum_{k=1}^{K} W(q)\ p_i^k = \frac{1}{K}\sum_{k=1}^{K}\sum_{j=1}^{D} c_j \delta_{jk}\ p_i^k \tag{4}$$

$$= \sum_{j=1}^{D} c_j \frac{1}{K}\sum_{k=1}^{K} \delta_{jk}\ p_i^k. \tag{5}$$

and reorder it with respect to the different values of c_j. With $p'_i = \sum_{i=1}^{K} \delta_{jk}\ p_i^k$ being the average probability over the K observations available at pixel i we obtain, similar to Eq. 1, for c_j

$$arg\min_{c} \sum_{i=1}^{N} (\vartheta_i - \sum_{j=1}^{D} c_j\ p'_i)^2 \tag{6}$$

now subject to

$$c_j \geq 0\ \forall j = 1 \dots D, \tag{7}$$

where each weight c_j determines how much an observation with data quality q_j should be considered for explaining the ground truth labels ϑ. To obtain a smoother distribution of W, we repeat the estimation of the c_i with different temporal subsets K ("bootstrapping") and average the resulting estimates (Fig. 10).

3 EXPERIMENTS

In a first experiment we test alternative fusion statics that may be as well or better suited for fusing probability maps than a plain averaging. In a second experiment we test how the fusion of ranked subsets differs from pooling all probabilistic maps, and in a third we evaluate the proposed method weighting images for fusion.

3.1 Testing generative fusion models

Noise and variation in the given data result from changes in vegetation cover, type of crop and land use, differences in contrasts on the ground after precipitation – due to humidity of the different soils – or variation in incoming or reflected radiance due to cloud coverage or aerosols. So we want to test whether averaging, i.e., assuming a "normal" or at least symmetric distribution performs sufficiently well by comparing it against a number of different noise models.

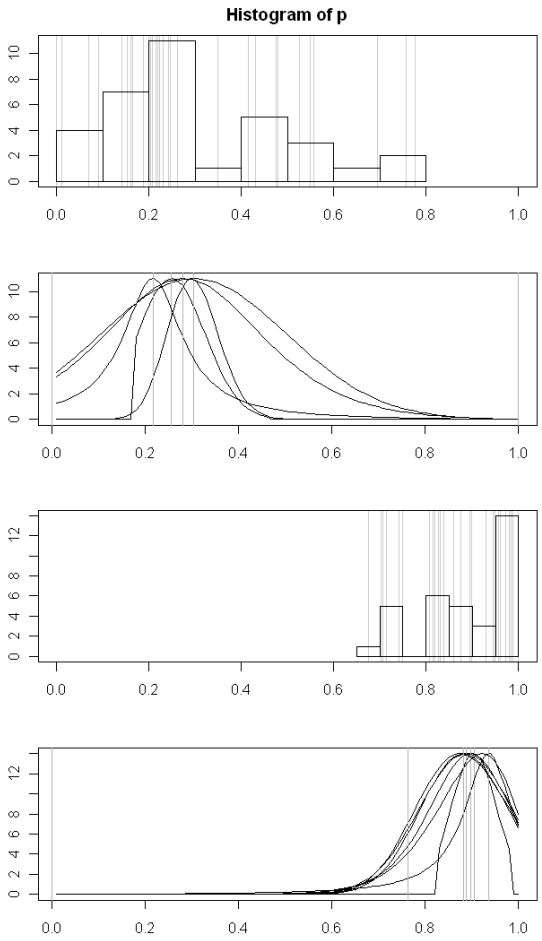

Fig. 6. Parametric distributions fit to the observations of a pixel in the anthrosol class (first row) and the background (third row). The gray vertical lines show the observed probabilities for the different time points, also summarized by the histograms. The boxes below (second and forth row) show the different distribution models that have been fit to the same data; here the gray vertical lines indicate the corresponding distribution parameters used to summarize the observations of the given pixel.

Experiment: We test a number of distributional noise models – "Cauchy", "chi-squared", "exponential", "log-normal", "logistic", "negative binomial", "normal", "Poisson", "t-distribution" and "Weibull" distributions – focusing on a comparison with our previous averaging approach [3] that is assuming a normal distribution of the noise. In addition, we test median and trimmed mean (using the inner 50% quantile) as robust variants to the normal distribution, and the posterior error. We evaluate the quality of the class separation in terms of the area under the curve of the receiver operator characteristic (AUC ROC) that measures class overlap.

For every pixel in our first test set we calculate the

parameters of these models using the time series of anthrosol probabilities observed at that pixel. Figure 6 shows the different distribution for the probabilities observed at a "anthrosol" pixel, and a "background" pixel, as well as an estimate of the model parameters used to summarize the observations in either case. For Normal, log-Normal, exponential and Poisson distributions we do this using analytical solutions, for all others we estimate the parameters using a direct optimization of the log-likelihood (Nelder-Mead downhill simplex method). For each archaeological site, we obtain parameters of the anthrosol area, and parameters from its direct surrounding, and we can calculate the ROC AUC to quantify how well the parameters of the given distribution model distinguish the signature of the anthrosols from their surroundings. Fig. 7 summarizes the resulting 243 ROC AUC scores for each fusion strategy.

We find that most summary statistics perform equally well, including the exponential, poisson, logistic and normal distribution (as determined in a Cox-Wilkinson test between the respective distributions for differences at 5% significance level; indicated gray). The distribution models also outperform their robust counter-parts (median, trimmed mean). The posterior, producing nearly binary maps (Fig. 8), ranks last due to its sensitivity to extreme probabilities (i.e., a single zero due to complete coverage of a site for a single observation will lead to an overall assignment to class 0). As several fusion statistics perform equally well, we focus on the normal distribution – as the most basic noise model – in the following.

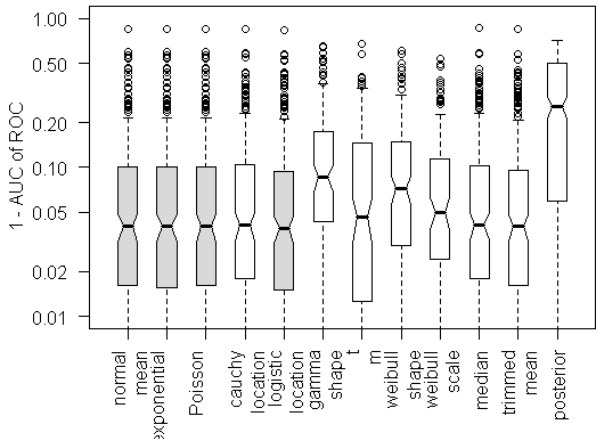

Fig. 7. Performance of different fusion statistics and approaches. Box-and-whisker plot show results for the 243 AUCs of the site-wise ROCs for the first data set. Boxes represent quartiles of these distributions, whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box. Non-overlapping notches indicate strong evidence that two medians differ. Normal performs best in terms of average site-wise prediction error, statistically indistinguishable from exponential, Poisson, and logistic (gray boxes).

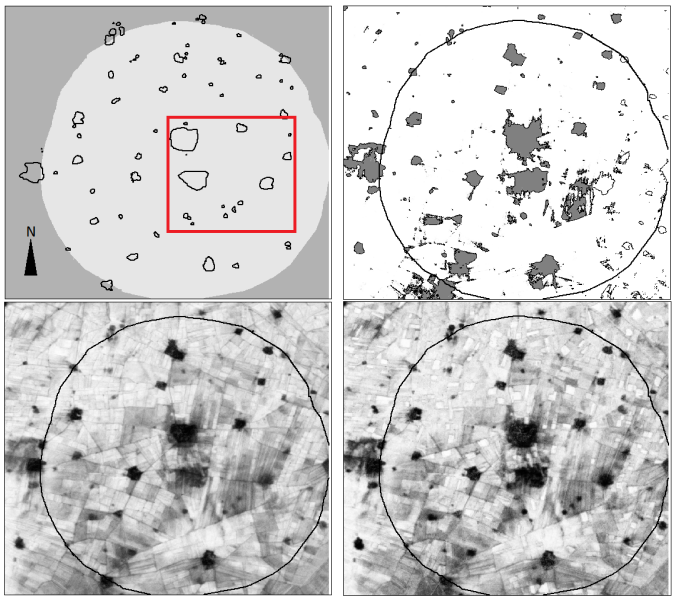

Fig. 8. Eastern test area with survey region (light gray) and sites mapped on the ground (outline black). The posterior probability (top right) does not show much details and is nearly identical with .5 threshold from average (outlined black). The averaged probability maps show significantly more details (average of all images: bottom left; average of optimal subset: bottom right). The image is centered around 36.8126 latitude / 41.9561 longitude, the extension of the area is about 15km in both directions. The red box corresponds to the area shown in Fig. 12.

3.2 On the benefit of ranking and pooling

We calculate two ranking criteria. One considers the difference between class means; the other considers the difference between the full distributions: 1) We calculate the average probability of all "anthrosol" locations within the image and the average probability of all "background" pixels. We then use the difference of both values, which indicates global class separation, to rank the images covering our two test areas. 2) We test another ranking criterion that also considers class overlap. For each site within a given scene, we calculate the AUC ROC between "anthrosol" and nearby "background" pixels. We then calculate the median AUC ROC value of all sites that are visible in the image.

Experiment: We evaluate the benefit of pooling using the ground truth for the western area. Here all 243 sites are at least covered by 30 observations (Fig. 4). Figure 9 presents results from pooling the top $n = 1, \ldots, 30$ images, and from pooling random subsets of probabilistic maps of the same size. We evaluate least squares error, variance, bias (being inverse proportional to class separation as used for the ranking), and AUC ROC. The fusion without ranking reduces the least squares error by reducing the variance while keeping the bias unchanged. The class-separation is maximal for the maximal number of 30 images that can be fused. At this point the unranked fusion coincides with the two ranked approaches, reducing the RMSE by 3%, a value

that corresponds to the level of the best single images. The two ranked fusion approaches return very similar results. Pooling reduces variance as well, although not as fast as for the unranked baseline method, with the best results obtained from pooling a large number of images. At the same time, the bias is smallest – i.e., the average class separation is maximal – when only a few "high quality" images are fused. As a consequence, both the mean error (which is composed of bias and variance) and the separation of the distributions (AUC ROC, which depends on class separation, but also the dispersion or noise of each distribution) has a minimum in between these two extremes. Pooling prediction maps by averaging (black lines) reduces the AUC error (i.e., 1-ROC AUC) from 0.14 (average performance of individual images) and 0.08 (best image) to 0.035. Ranking the images according to one of the two measures (red and green lines) reduced the AUC error even further, to 0.025 at best, corresponding to 20% and 5% of the results obtained for the single image. While results are very robust with respect to the ranking measures used – returning similar results for both average class difference and median site-wise AUC ROC – other parametric or non-parametric quality scores that measure differences between univariate distributions, such as entropy, Gini impurity, or Fisher's ratio, may be used.

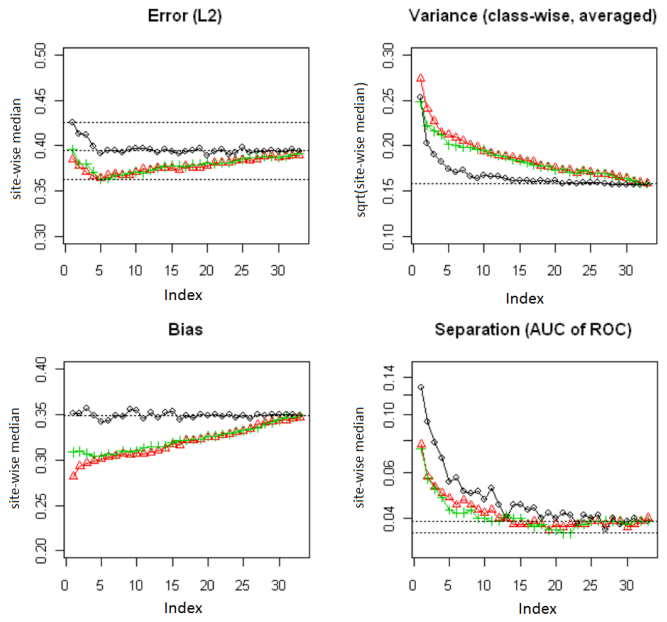

Fig. 9. Ranking classification maps according to their quality (using the average site-wise AUC - green crosses, or using average site-wise class difference - red triangles), reduces the L2 error and AUC noticeably in comparison to a fusion of all available images (black circles). Decrease in L2 error is due to a lower bias when discarding low quality data, and and only fusing the anthrosol maps that have the best best separation between the two classes.

3.3 Learning fusion weights using NNLS regression

Ordering observations using the cross-validated test error and pooling only a subset may improve results. Searching for the subset that optimizes the least squares criterion in a cross-validation as in Fig. 9, however, leads to a selection of very few images. In both test areas no more than 10% of the locally available images are combined, returning results that are very noisy and do not make full use of the available data. As an alternative to the crisp threshold we may weight observations and predict anthrosol locations from a weighted sum of all available probability maps.

Experiment: We apply the NNLS resampling strategy to the probabilistic maps of both test areas. Figure 10 shows the resulting coefficients c_i as a function $W(q_i)$ of quality score AUC ROC. The approximate cut-off at $W = 1$, where images with corresponding AUC ROC are either upweighted $W > 1$ or down-weighted $W < 1$, is in between the cutoff that has been found for AUC ROC and least squares criterion in the cross-validation (Fig. 9).

Fig. 12 shows results for the eastern test region when fusing (i) all images that are available, (ii) all images above the AUC ROC cutoff from the cross-validation, and (iii) all images weighted by $W(q)$. The probabilistic maps from the fused subsets show more detail, including a string of small sites that are only visible in one or two of the individual images. Setting a – somewhat arbitrary – threshold of 50% probability, we can calculate the average probability of the foreground pixels for each of the 76 site as well as the average probability of nearby background pixels. For the simple average over all anthrosol map, we find that 45 sites with a foreground probability surpassing the 50% threshold, while at the same time 10 sites have an average background probability that is also above 50%. Calculating the same for the map that we obtain by averaging the image subset with the least fit error, the two numbers change to 52 and 7, and to 54 and 12 when calculating them with the weighted averaging. These results indicate that the subset selection strategies perform better than our previous fusion approach from [3] that was averaging over all available images.

A more comprehensive comparison for all possible thresholds is possible by evaluating precision and recall for all pixels of the test area, as shown in Fig. 11. Again, the plain averaging of all anthrosol probability maps ("fusion all") is outperformed by the two more restrictive approaches. Comparing results from the subset that has the least fit error ("fusion XV") with the results from the weighted averaging ("fusion NNLS"), we find that both method perform nearly equally well – here, with a slight edge for the weighted averaging.

4 Discussion

4.1 Multitemporal fusion and bias-variance tradeoff

We tested several distributional models to cope with a variable number of observations, and different approaches improved the performance significantly. Fusing multiple decisions is a relatively common concept in

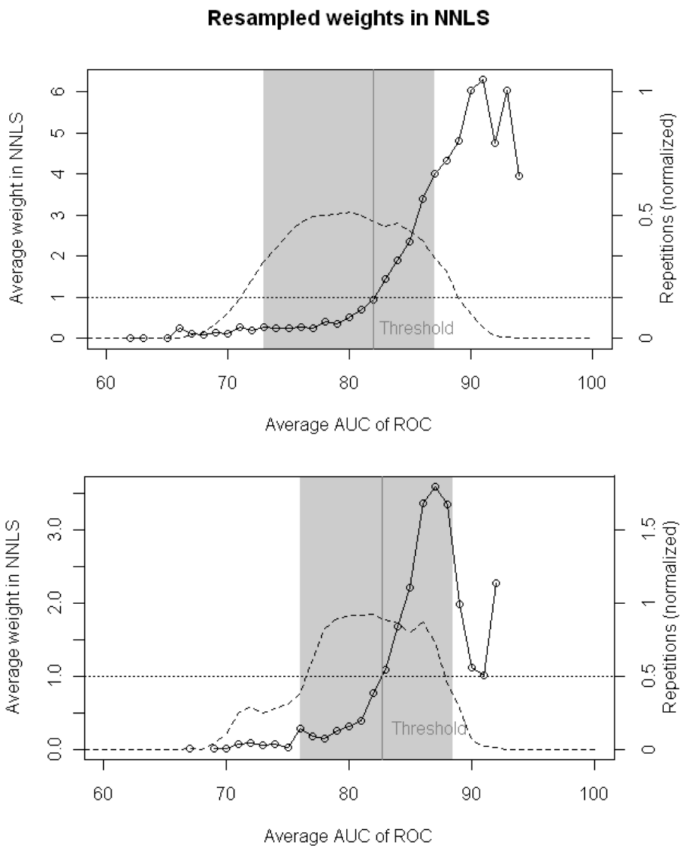

Fig. 10. Learning the relation between weight and data quality for the western (top) and eastern (bottom) test area. Shown is the weight function used when averaging the probability maps, the distribution of the available observations with inner 75% quantile (dark gray).

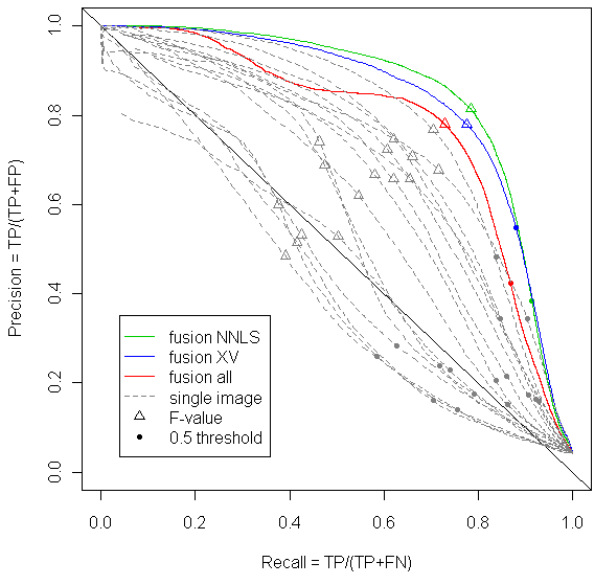

Fig. 11. Precision and recall for the eastern test area with pixel-wise ground truth, using 49 7091px. for background and 21 959px. for the anthrosol "foreground". Precision-recall curves along the negative diagonal (indicated black) can be considered a random classification. The F-measure, i.e., the harmonic mean of the observations, is indicated by triangles. Points indicate precision and recall for a classification with threshold 0.5.

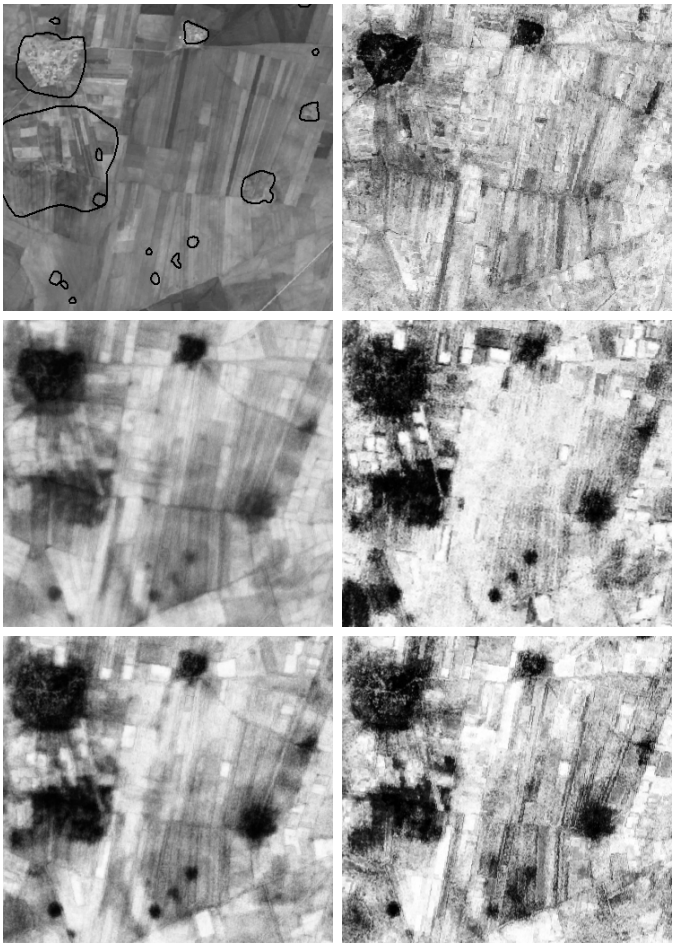

Fig. 12. Ground truth, averaged, averaged subset (left column), and individual images of different quality (right column), for the south eastern part of the second area. Dark areas indicate high probability. Pooled images background shows less variance, while sites are still visible. True outlines of the large area in central left part of the image are still subject to debate. Also see Fig. 8 to compare with ground truth. The extensions of the area shown are about 3.5km in both directions.

machine learning and different concepts of how to generate slightly different predictions from the same training data have been proposed: using different subsets of the training data for training every single decision tree is an essential element of the random forest classifier [47], as well as the use of several different classifiers trained on the same training data, but with slightly different predictions [25], [48], [49]. By fusing slightly different observations of the same spatial scene – each of which generates a slightly differen spatial patterns – we follow a similar approach. As pointed out earlier, e.g., by Bachman et al: "In remote sensing applications, variance reduction also can be achieved by using multi-sensor or multi-temporal data to produce a pool of classifiers with decorrelated error distributions" [50]. However, while some recommend very generally to pool predictions [51] and that "inaccurate classifiers should not be excluded …since they may have the potential to improve the overall combined accuracies" [48], we find

that pooling does not always increase the performance of the classification, but that the benefit will crucially depend on the trade-off in between variance and bias. In our experiments pooling only reduces variance, i.e., the "decorrelated" error distribution. Errors that come from *bias* of the individual image, i.e., the average mismatch between labels and predictions, are not removed by averaging and – when including biased images – pooled results may even be worse than those of the best single prediction. As selecting un-biased predictors (or probabilistic maps) reduces the overall prediction error, we benefit from the NLLS weighting strategy.

4.2 Application to archaeological survey

In this application we did not perform any kind of spatial regularization or smoothing as this might potentially remove the signal of some of the smaller archaeological sites; the pixel sizes of ASTER and Landsat images is already beyond the limit of what is useful for many archaeological survey problems. Eventually, combined spectral-spatial classification approaches [52] may help in detecting and delineating sites or in updating training labels when repeating the whole classification procedure.

To retrain the classifier for each image, we need some ground truth locations in the given image. For our anthrosol detection task local ground truth is often available, as settlement mounds at the upper end of the settlement hierarchy are visible for example in digital elevation models [35], and many of the smaller sites have been mapped in field survey for many regions in the Near East [37]. For other detection tasks in archaeological survey, however, this requirement of having some prior knowledge about the locations of interest is the strongest limitation of the multi-temporal fusion approach. Still, learning one classifier for each image from local ground truth, evaluating the benefit of the individual map and ranking it, and fusing the best subset, provides a fairly general approach to combine different sources of information: Landsat images can be combined straightforwardly with ASTER images, probabilistic maps generated from hyper-spectral imagery – available for some areas – can be considered as well. Overall, our local retraining and fusion approach may be well tailored to the needs of many applications in archaeological remote sensing. While in the present study our focus has been on anthrosols, we would expect that mapping the spectral signal of other archaeological structures – with distinct spectral signature and of appropriate size – would also benefit from our pooling approach.

5 Summary and Conclusions

Averaging probabilistic maps from multiple observations, as we did in [3], is optimal for our application. We can show that fusing several maps improves the result by reducing the variance that can be high when evaluating individual observations only. Ranking the observations according to some local quality measure, for example the AUC ROC, and considering only those probabilistic maps that separate classes well reduces the overall bias and, hence, reduces the overall prediction error even further. In order to identify and suppress those samples that do not help in separating classes, we propose a weight function that can be estimated for every local test region in an automated fashion using a non-linear least squares regression strategy. Different from a standard linear fusion model, this subset selection approach is capable of dealing with variable numbers of observations.

Using our weight function improves the results also in locations where where soil properties have a visible imprint on vegetation only during a short time of the year. To this end we expect a significant improvement over our previous results when applying our classification strategy to landscapes in the Near East with less favourable environmental conditions. While we focus on one particular application in archaeological survey, we expect that our ranking and fusion strategy may also have significant impact in other related tasks. It may be generally applicable to the mapping of spatially static surface properties that are subject to strong seasonal variation and that have some limited ground truth available, for example, in the characterization of in situ soils or minerals.

In a next step we will use the subset selection algorithm to improve the probabilistic map of the 20 000 km^2 we studied in [3] [5]. The subset selection would have to be adapted locally, and further tests may be required to study how it might be applied to large areas, and how large local blocks should be. Partial ground truth from CORONA images can also be generated for other regions of the Near East, for example, in a crowd-sourcing effort using the CORONA Atlas of the Middle East[6] [53]. Current surveys in Northern Iraq will allow us to further test the generalization behaviour of the algorithm, and to evaluate it on the ground in further prospective studies [54].

Acknowledgements This research was supported by the Technische Universität München - Institute for Advanced Study (funded by the German Excellence Initiative and the European Union Seventh Framework Programme under grant agreement n 291763, the Marie Curie COFUND program of the the European Union), and by a fellowship of the Fritz-Thyssen-Stiftung to BHM. The ASTER L1B data product was obtained through the online Data Pool at the NASA Land Processes Distributed Active Archive Center (LP DAAC), USGS/Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota (https://lpdaac.usgs.gov/data_access).

5. Fused probabilistic map and other data products available from the Harvard Dataverse Network via http://hdl.handle.net/1902.1/17731.

6. corona.cast.uark.edu/

REFERENCES

[1] M Fauvel, J Chanussot, and JA Benediktsson. Decision fusion for the classification of urban remote sensing images. *IEEE Trans Geosc Rem Sens*, 44:2828–2838, 2006.

[2] Y Tarabalka, G Charpiat, L Brucker, and B H Menze. Enforcing monotonous shape growth or shrinkage in video segmentation. In *Proc BMCV (British Machine Vision Conference)*, 2013.

[3] B H Menze and J A Ur. Mapping patterns of long-term settlement in Northern Mesopotamia at a large scale. *Proc Nat Acad Sci USA*, 109:E778–E787, 2012.

[4] V De Laet, E Paulissen, and M Waelkens. Methods for the extraction of archaeological features from very high-resoluition Ikonos-2 remote sensing imagery, Hisar (southwest Turkey). *J Archaeol Science*, 34:830–841, 2007.

[5] A Beck, G Philip, M Abdulkarim, and D Donoghue. Evaluation of Corona and Ikonos high resolution satellite imagery for archaeological prospection in western Syria. *Antiquity*, 81:161–175, 2007.

[6] D C Comer and M J Harrower. *Mapping Archaeological Landscapes from Space*. Springer, New York, 2013.

[7] R.M. Cavalli, G.A. Licciardi, and J. Chanussot. Detection of anomalies produced by buried archaeological structures using nonlinear principal component analysis applied to airborne hyperspectral image. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6(2):659–669, 2013.

[8] Athos Agapioua, Diofantos G Hadjimitsisa, and Dimitrios D. Alexakisa. Development of an image-based method for the detection of archaeological buried relics using multi-temporal satellite imagery. *International Journal of Remote Sensing*, 34:5979–5996, 2013.

[9] L C Rowan and J C Mars. Lithologic mapping in the Mountain Pass, California area using advanced spaceborne thermal emission and reflection radiometer (ASTER) data. *Remote sensing of Environment*, 84:350366, 2003.

[10] Yoshiki Ninomiyaa, Bihong Fu, and Thomas J. Cudahy. Detecting lithology with Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) multispectral thermal infrared radiance-at-sensor data. *Rem Sens Environm*, 99:127–139, 2005.

[11] M. J. Falkowski, P. E. Gessler, P. Morgan, A. T. Hudak, and A Smith. Characterizing and mapping forest fire fuels using ASTER imagery and gradient modeling. *Forest Ecology and Management*, 217:129–146, 2005.

[12] G Mallinis, I. D. Mitsopoulos, A. P. Dimitrakopoulos, I.Z. Gitas, and M. Karteris. Local-scale fuel-type mapping and fire behavior prediction by employing high-resolution satellite imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 1(4):230–239, 2008.

[13] K. E. Sawaya, L. G. Olmanson, N. J. Heinert, P. L. Brezonik, and M. E Bauer. Extending satellite remote sensing to local scales: land and water resource monitoring using high-resolution imagery. *Rem Sens Environm*, 88:144–156, 2003.

[14] S H Parcak. *Satellite remote sensing for archaeology*. Routledge, London, 2009.

[15] M J Abrams and D C Comer. Multispectral and hyperspectral technology and archaeological applications. In *Mapping Archaeological Landscapes from Space*, pages 57–71. Springer, New York, 2013.

[16] M Altaweel. The use of ASTER satellite imagery in archaeological contexts. *Archaeol Prospection*, 12:151–166, 2005.

[17] Stephen H. Savage, Thomas E. Levy, and Ian W. Jones. Prospects and problems in the use of hyperspectral imagery for archaeological remote sensing: a case study from the Faynan copper mining district, Jordan. *Journal of Archaeological Science*, 39:407–420, 2012.

[18] A Beck. Google Earth and World Wind: remote sensing for the masses. *Antiquity*, 80, 2006.

[19] R R Colditz, T Wehrmann, M Bachmann, K Steinnocher, M Schmidt, G Strunz, and S Dech. Influence of image fusion approaches on classification accuracy: a case study. *Intern J Rem Sens*, 27:3311–3335, 2006.

[20] H Aanaes, J R Sveinsson, A A Nielsen, T Bovith, and J A Benediktsson. Model-based satellite image fusion. *IEEE Trans Geosc Rem Sens*, 46:1336–1346, 2008.

[21] C H Song and C E Woodcock. Monitoring forest succession with multitemporal Landsat images: Factors of uncertainty. *IEEE Trans Geosc Rem Sens*, 41:2557–2567, 2003.

[22] Z J Wang, D Ziou, C Armenakis, D Li, and Q Q Li. A comparative analysis of image fusion methods. *IEEE Trans Geosc Rem Sens*, 43:1391–1402, 2005.

[23] K A Kalpoma and J I Kudoh. Image fusion processing for IKONOS 1-m color imagery. *IEEE Trans Geosc Rem Sens*, 45:3075 – 3086, 2007.

[24] G M Foody, D S Boyd, and C Sanchez-Hernandez. Mapping a specific class with an ensemble of classifiers. *Intern J Remote Sens*, 28:1733–1746, 2007.

[25] G J Briem, J A Benediktsson, and J R Sveinsson. Multiple classifiers applied to multisource remote sensing data. *IEEE Trans Geosc Rem Sens*, 40:2002–2291, 2002.

[26] B Waske and J A Benediktsson. Fusion of support vector machines for classification of multisensor data. *IEEE Trans Geosc Rem Sens*, 45:3858–3866, 2007.

[27] B Waske and S van der Linden. Classifying multilevel imagery from SAR and optical sensors by decision fusion. *IEEE Trans Geosc Rem Sens*, 46:1457–1466, 2008.

[28] H T X Doan and G M Foody. Increasing soft classification accuracy through the use of an ensemble of classifiers. *Intern J Remote Sens*, 28:4609–4623, 2007.

[29] B H Menze and J A Ur. Multi-temporal classification of multispectral images for settlement survey in northeastern Syria. In *Mapping Archaeological Landscapes from Space*, pages 219–228. Springer, New York, 2013.

[30] JA Benediktsson and I Kanellopoulos. Classification of multisource and hyperspectral data based on decision fusion. *IEEE Trans Geosc Rem Sens*, 37:1367–1377, 1999.

[31] B H Menze, W Petrich, and F A Hamprecht. Multivariate feature selection and hierarchical classification for infrared spectroscopy: serum-based detection of bovine spongiform encephalopathy. *Anal Bioanal Chem*, 387:801–1807, 2007.

[32] J Ur. Corona satellite photography and ancient road networks: A northern mesopotamian case study. *Antiquity*, 77:102–115, 2003.

[33] K N Wilkinson, A R Beck, and G Philip. Satellite imagery as a resource in the prospection for archaeological sites in central Syria. *Geoarchaeol*, 21:735–750, 2006.

[34] T J Wilkinson. *Archaeological landscapes of the Near East*. University of Arizona Press, Tuscon, 2003.

[35] B H Menze, J A Ur, and A G Sherratt. Detection of ancient settlement mounds – Archaeological survey based on the SRTM terrain model. *Photogr Eng Rem Sens*, 72:321–327, 2006.

[36] Jason Ur. Spying on the past: Declassified intelligence satellite photographs and near eastern landscapes. *Near Eastern Archaeology*, 76(1):28–36, 2013.

[37] J A Ur. *Urbanism and cultural landscapes in northeastern Syria: the Tell Hamoukar Survey, 1999-2001*. Oriental Institute Publications, Chicago, 2010.

[38] B H Menze and J A Ur. Classification of multispectral ASTER imagery in the archaeological survey for settlement sites of the Near East. In *Proc ISPMSRS (International Symposium on Physical Measurements and Signatures in Remote Sensing)*, pages 1–12, 2007.

[39] B H Menze, S Mühl, and J A Ur. Surveying and mapping Near Eastern settlement mounds from space. In *Proc CAA (Conference on Computer Applications and Quantitative Methods in Archaeology)*, 2007.

[40] B H Menze, B M Kelm, R Masuch, U Himmelreich, W Petrich, and F A Hamprecht. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*, 10:213, 2009.

[41] J Ham, Y C Chen, M M Crawford, and J Ghosh. Investigation of the random forest framework for classification of hyperspectral data. *IEEE Trans Geosc Rem Sens*, 43:492–501, 2005.

[42] P O Gislason, J A Benediktsson, and J R Sveinsson. Random forests for land cover classification. *Patt Recogn Lett*, 27:294–300, 2006.

[43] JCW Chan and D Paelinck. Evaluation of random forest and adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Rem Sens Environm*, 112:2999–3011, 2008.

[44] B Waske, S van der Linden, J A Benediktsson, A Rabe, and P Hostert. Sensitivity of support vector machines to random feature selection in classification of hyperspectral data. *IEEE Trans Geosc Rem Sens*, 48:2880–2889, 2010.

[45] B H Menze, B M Kelm, D N Splitthoff, U Koethe, and F A Hamprecht. On oblique random forests. In *Proc ECML (European Conference on Machine Learning)*, 2011.
[46] S N Lahiri. *Resampling Methods for Dependent Data*. Springer, New York, 2003.
[47] L Breiman. Random forests. *Mach Learn J*, 45:5–32, 2001.
[48] M Petrakos, JA Benediktsson, and I Kanellopoulos. The effect of classifier agreement on the accuracy of the combined classifier in decision level fusion. *IEEE Trans Geosc Rem Sens*, 39:2539–2546, 2001.
[49] Z Huang and B G Lees. Combining non-parametric models for multisource predictive forest mapping. *Photogr Eng Rem Sens*, 70:415–425, 2004.
[50] C M Bachmann, M H Bettenhausen, R A Fusina, T F Donato, A L Russ, J W Burke, G M Lamela, W J Rhea, B R Truitt, and J H Porter. A credit assignment approach to fusing classifiers of multiseason hyperspectral imagery. *IEEE Trans Geosc Rem Sens*, 41:2488–2499, 2003.
[51] S Prasad and L M Bruce. Decision fusion with confidence-based weight assignment for hyperspectral target recognition. *IEEE Trans Geosc Rem Sens*, 46:1448–1456, 2008.
[52] Y Tarabalka, J A Benediktsson, J Chanussot, and J C Tilton. Multiple spectral–spatial classification approach for hyperspectral data. *IEEE Trans Geosc Rem Sens*, 48:4122–4132, 2010.
[53] Jesse Casana and Jackson Cothren. Stereo analysis, DEM extraction and orthorectification of CORONA satellite imagery: archaeological applications from the Near East. *Antiquity*, 82:732–749, 2008.
[54] J. A. Ur, L. de Jong, J Giraud, J. F. Osborne, and J. MacGinnis. Ancient cities and landscapes in the kurdistan region of iraq: The erbil plain archaeological survey 2012 season. *Iraq*, 75:89–114, 2013.

Bjoern Menze is Assistant Professor in Computer Science at the Technische Universität München (TUM) in Munich, Germany, holding a Rudolf Moessbauer tenure-track professorship of the TUM Institute for Advanced Study. He studied physics and computer science in Uppsala, Sweden, and Heidelberg, Germany, and was postdoc at Harvard University, Harvard Medical School, and the MIT, as well as a senior researcher and lecturer at the INRIA Sophia-Antipolis and the ETH Zürich. His research is in medical image computing – exploring topics at the interface of medical computer vision, image-based modeling and computational physiology – also seeking for transfers of related computational methods towards applications in Near Eastern Archaeology.

Jason Ur is Professor of Anthropology in the Department of Anthropology at Harvard University. He specializes in early urbanism, landscape archaeology, and remote sensing, particularly the use of declassified US intelligence imagery. He has directed field surveys in Syria, Iraq, Turkey, and Iran. Since 2012, he has directed the Erbil Plain Archaeological Survey, an archaeological survey in the Kurdistan Region of northern Iraq. He is also preparing a history of Mesopotamian cities.