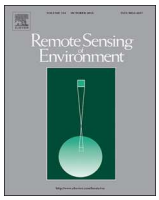
Remote Sensing of Environment 245 (2020) 111741

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse



Deep learning on edge: Extracting field boundaries from satellite images with a convolutional neural network

François Waldner[a,*], Foivos I. Diakogiannis[b]

[a] CSIRO Agriculture & Food, 306 Carmody Road, St Lucia, Queensland, Australia
[b] CSIRO Data61, Analytics, 147 Underwood Avenue, Floreat, Western Australia, Australia

ARTICLE INFO

Edited by Marie Weiss

Keywords:
Agriculture
Field boundaries
Sentinel-2
Semantic segmentation
Instance segmentation
Multitasking
Computer vision
Generalisation

ABSTRACT

Applications of digital agricultural services often require either farmers or their advisers to provide digital records of their field boundaries. Automatic extraction of field boundaries from satellite imagery would reduce the reliance on manual input of these records, which is time consuming, and would underpin the provision of remote products and services. The lack of current field boundary data sets seems to indicate low uptake of existing methods, presumably because of expensive image preprocessing requirements and local, often arbitrary, tuning. In this paper, we propose a data-driven, robust and general method to facilitate field boundary extraction from satellite images. We formulated this task as a multi-task semantic segmentation problem. We used ResUNet-a, a deep convolutional neural network with a fully connected UNet backbone that features dilated convolutions and conditioned inference to identify: 1) the extent of fields; 2) the field boundaries; and 3) the distance to the closest boundary. By asking the algorithm to reconstruct three correlated outputs, the model's performance and its ability to generalise greatly improve. Segmentation of individual fields was then achieved by post-processing the three model outputs, *e.g.*, *via* thresholding or watershed segmentation. Using a single monthly composite image from Sentinel-2 as input, our model was highly accurate in mapping field extent, field boundaries and, consequently, individual fields. Replacing the monthly composite with a single-date image close to the compositing period marginally decreased accuracy. We then showed in a series of experiments that, without recalibration, the same model generalised well across resolutions (10 m to 30 m), sensors (Sentinel-2 to Landsat-8), space and time. Building consensus by averaging model predictions from at least four images acquired across the season is paramount to reducing the temporal variations of accuracy. Our convolutional neural network is capable of learning complex hierarchical contextual features from the image to accurately detect field boundaries and discard irrelevant boundaries, thereby outperforming conventional edge filters. By minimising over-fitting and image preprocessing requirements, and by replacing local arbitrary decisions by data-driven ones, our approach is expected to facilitate the extraction of individual crop fields at scale.

1. Introduction

Many of the promises of digital agriculture centre on assisting farmers to monitor their fields throughout the growing season. Having precise field boundaries has become a prerequisite for field-level assessment and often, when farmers are being signed up by service providers, they are asked for precise digital records of their boundaries. Unfortunately, this process remains largely manual and time-consuming, which creates disincentives. There are also increasing applications whereby earth observation is used for estimating areas of crop planted, for forecasting crop yields and for monitoring food security (De Wit and Clevers, 2004; Blaes et al., 2005; Matton et al., 2015). These applications would greatly benefit from repeated extraction of field boundaries across large areas. Automating this process not only facilitates bringing farmers on board, and hence fostering wider adoption of digital agricultural services, but also improves products and services to be provided using remote sensing.

Several approaches have been devised to extract field boundaries from satellite imagery. These approaches generally fall into three categories: edge-based methods (Mueller et al., 2004; Shrivakshan and Chandrasekar, 2012; Turker and Kok, 2013; Yan and Roy, 2014; Graesser and Ramankutty, 2017); region-based methods (Evans et al., 2002; Mueller et al., 2004; Salman, 2006; Garcia-Pedrero et al., 2017); and hybrid methods that seek to address shortcomings of one category with components from the other (Rydberg and Borgefors, 2001). Edge-based methods rely on filters to identify discontinuities in images where

* Corresponding author.
E-mail address: franz.waldner@csiro.au (F. Waldner).

https://doi.org/10.1016/j.rse.2020.111741
Received 25 October 2019; Received in revised form 21 February 2020; Accepted 26 February 2020
Available online 19 May 2020
0034-4257/

pixel values change rapidly. Filters define specific kernels (the Scharr, Sobel, and Canny fil are common examples), which are then convolved with the input image to emphasise edges. A number of issues arise when working with edge operators: their sensitivity to high-frequency noise often creates false edges; and their parameterisation is arbitrary, relevant to specific contexts, and a single parameterisation may lead to incomplete boundary extraction. Post-processing and locally-adapted thresholds may address these issues and lead to better-defined, closed boundaries. Region-based algorithms, such as the watershed segmentation (Soille and Ansoult, 1990), group neighbouring pixels into objects based on some homogeneity criterion. Finding the optimal segmentation parameters for region-based algorithms remains a trial and error process that likely yields sub-optimal results. For instance, with poor parameterisation, objects may stop growing before reaching the actual boundaries (Chen et al., 2015a), creating sliver polygons and shifting the extracted boundaries inwards. Region-based methods also tend to over-segment fields with high internal variability and under-segment small adjacent fields (Belgiu and Csillik, 2018). Some of these adverse effects might be mitigated by purposefully over-segmenting images and deciding whether adjacent objects should be merged with machine learning (see Garcia-Pedrero et al., 2017, for instance). Despite the availability of edge-based and region-based methods, there seems to be a low uptake of these methods by the user community, suggesting a lack of fitness for purpose. For instance, the only global map of field size was obtained from crowdsourced, manually-digitised polygons (Lesiv et al., 2019).

This low uptake can be explained by the expense of image preprocessing, local, often arbitrary, tuning which does not generalise to other locations, and requires ancillary data, such as cropland or crop type maps (*e.g.*, Yan and Roy, 2014; Graesser and Ramankutty, 2017). For instance, multitemporal features are commonly used as input data for field boundary detection. While multitemporal information is likely to improve accuracy especially in those highly-dynamic systems, humans can draw field boundaries in many cases from well-targeted single-date images. Besides, persistent cloud coverage makes it difficult to generate consistent time series in areas such as the tropics. Therefore, we think that new methods with lower preprocessing requirements and data-driven parameter tuning are needed to facilitate large-scale extraction of field boundaries and improve their uptake.

Deep neural networks open up new avenues for field boundary extraction because they do not require hand-crafted features and their architectures are highly adaptive to new problems (Agravat and Raval, 2018). Deep convolutional neural networks, a group of deep neural networks which use convolution operations, are increasingly used in image analysis because they can exploit hierarchical (local to global) features in images. While filters are hand-engineered in edge-based methods, convolutional neural networks can learn these filters. Initially devised for natural images, these networks have been revisited and adapted to tackle semantic segmentation problems (*i.e.*, the process of linking each pixel in an image to a class label) in remote sensing, such as road extraction (Cheng et al., 2017), cloud detection (Chai et al., 2019), crop identification (Ji et al., 2018), river and water body extraction (Chen et al., 2018b; Isikdogan et al., 2018), and urban mapping (Diakogiannis et al., 2020). As such, convolutional neural networks seem particularly well-suited to extract field boundaries at scale, but this has yet to be empirically proven.

Our overarching aim is to develop and evaluate a method to routinely extract field boundaries across large areas by replacing context-specific arbitrary decisions and by minimising the image preprocessing workload. We formulated this task as a multi-task semantic segmentation problem for a fully convolutional neural network where each pixel is simultaneously annotated with three labels: 1) the probability of belonging to a field (extent mask); 2) the probability of belonging to a boundary (boundary mask); and 3) the distance to the closest boundary (distance mask). Here, we used ResUNet-a (a deep convolutional neural network with a fully connected UNet architecture which features dilated convolution and conditioned inference; Diakogiannis et al., 2020) to learn to perform these tasks.[1] As neural networks might produce discontinuous boundaries, we also introduce two post-processing methods that leverage its outputs to generate closed boundaries and retrieve individual fields. In other words, the post-processing methods exploit the semantic segmentation outputs to achieve instance segmentation.

This paper intends to test the performance and limitations of the proposed approach and to lay the blueprint for national to global field boundary extraction using deep learning. As this paper will show, our method extracted field boundaries with high thematic and geometric accuracy when using a 10-m monthly composite image of Sentinel-2 covering our main site in South Africa. We then conducted a series of experiments which demonstrate that over-fitting was minimised, allowing our convolutional neural network to be applied across a range of conditions without recalibration. Specifically, it generalised well 1) to a 10-m single-date Sentinel-2 image close to the compositing period, 2) to a 30-m Landsat image, and 3) to other locations (secondary sites in Argentina, Australia, Canada, Russia, Ukraine) and acquisition dates. By learning spectral and contextual information, our deep convolutional neural network discards edges that are not part of field boundaries and emphasises those that are, providing a clear advantage over conventional methods.

This paper is organised as follows. In Section 2, we describe our method to extract field boundaries (the neural network, the data augmentation procedure and the strategy for training and inference) and to post-process its outputs to retrieve individual fields (instance segmentation). Section 3 presents the data collected for the main and the secondary sites and Section 4 details the experimental design. Results are presented in Section 5 and discussed in Section 6, where we propose evidence-based recommendations for large-scale field boundary extraction with deep learning.

2. Extracting field boundaries with multi-task semantic segmentation

Conceptually, extracting field boundaries consists of labelling each pixel of a multi-spectral image with one of two classes: “boundary” or “not boundary”. While predicting only the classes of interest (single-tasking) can achieve acceptable performance (see Persello et al., 2019, for instance), it ignores training signals of related learning tasks that might help improve the accuracy of the initial task (Ruder, 2017). By sharing representations between related tasks *i.e.*, multi-task learning, a model can learn to generalise better on the initial task. Instead of predicting each related task simultaneously and independently, these can also be combined in the last layer of the architecture to further constrain inference of the initial task (conditioned multitasking), which was found to reduce the variance, to stabilise the gradient updates and thus to improve model performance (Diakogiannis et al., 2020).

Therefore, we formulate the extraction of field boundaries as a semantic segmentation problem where the goal is to predict multiple class labels. We trained a convolutional neural network called ResUNet-a to perform the four correlated tasks: to map the extent of fields, to identify field boundaries, to estimate the distance to the nearest boundary and reconstruct input images (Fig. 1b; Section 2.1). We refer to Brodrick et al. (2019) for a thorough introduction to convolutional neural networks and to Diakogiannis et al., 2020 for more details about the ResUNet-a framework.

Our overarching goal is to generate individual fields (instance segmentation) to enable per-field analytics, which requires that boundaries are closed contours. However, it is likely that a convolutional neural network provides open contours, especially if it is trained to predict only boundaries. Compared to single-tasking, multitasking is likely to provide improved boundaries because it learns to predict multiple highly correlated outputs (boundary, distance, extent) and it provides

[1] See https://github.com/feevos/resuneta for a python implementation of the ResUNet-a model.

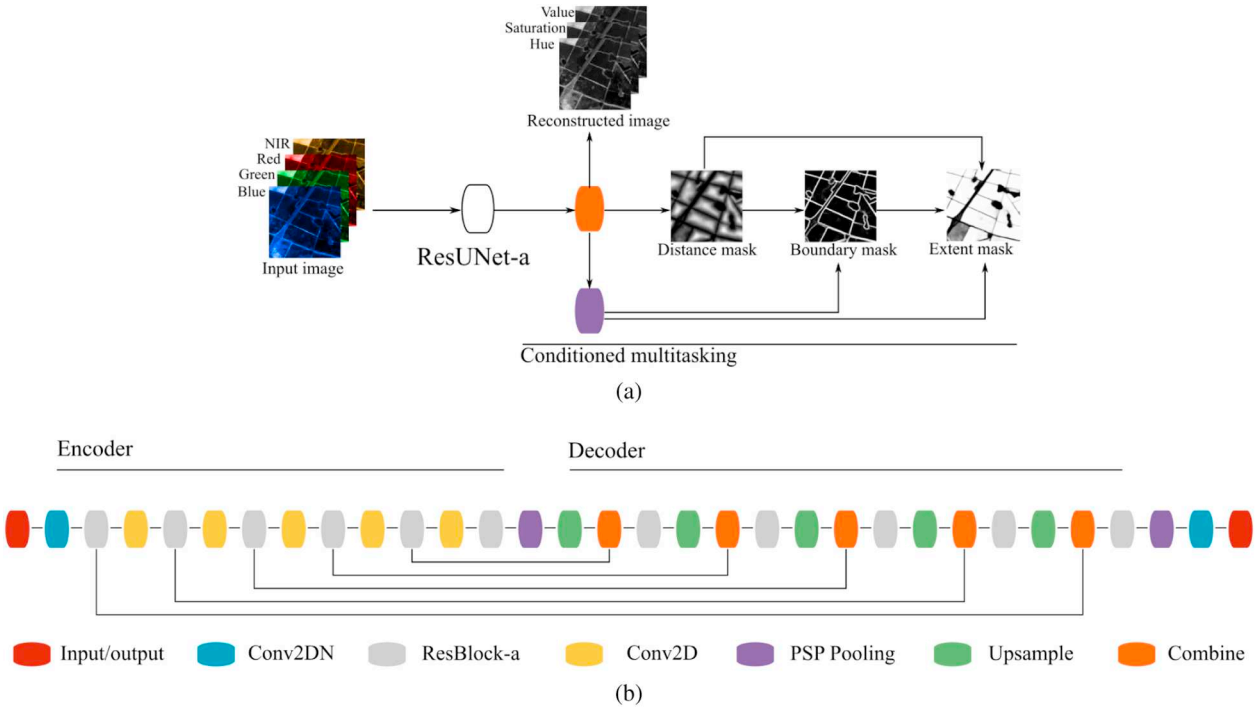


Fig. 1. Field boundary extraction formulated as multiple semantic segmentation tasks. (a) Overview of the ResUNet-a's graph for multitasked inference. The distance mask is first generated from the feature map, then is combined to the feature map to predict the boundary mask. Both masks are used to predict the extent mask. An independent branch reconstructs the input image. (b) Architecture of ResUNet-a D6.

more information that can help remedy the problem of open contours. We introduce two data-driven post-processing methods to generate closed field boundaries and, therefore, to extract individual fields (Section 2.2). One is based on thresholding and the other on watershed segmentation; both leverage multiple outputs of ResUNet-a.

2.1. Boundary detection with a deep convolutional neural network

2.1.1. Model architecture

The ResUNet-a model is a deep convolutional neural network, which has been previously shown to outperform state-of-the-art architectures for semantic segmentation (Diakogiannis et al., 2020). ResUNet-a features the following components (Fig. 1b): 1) a UNet architecture, which consists of a contracting path that captures context and a symmetric expanding path that localise objects precisely (Ronneberger et al., 2015); 2) residual blocks, which helps alleviate the problem of vanishing and exploding gradients (He et al., 2016a); 3) atrous convolutions (with a range of dilation rates) that increase the receptive field (Chen et al., 2017, 2018a); 4) pyramid scene parsing pooling to include contextual information (Zhao et al., 2017); and 5) conditioned multitasking. ResUNet-a produces four output layers: the extent mask, the boundaries, the distance mask (Borgefors, 1986), and full reconstruction of the input image in the Hue-Saturation-Value colour space (Fig. 1a). The colour space transformation provides additional information for the correlation between colour variations and object extent.

The UNet backbone architecture, also known as encoder-decoder, consists of two parts: the contraction part or encoder, and the symmetric expanding path or decoder. The encoder compresses the information content of an arbitrarily high-dimensional image. The decoder gradually upscales the encoded features back to the original resolution and precisely localises the classes of interest. The building blocks of the encoder and decoder in the ResUNet-a architecture consist of residual units with multiple parallel branches of atrous convolutions, each with a different dilation rate.

We implemented two basic architectures that differ in depth (number of layers in the encoder-decoder). ResUNet-a D6 has six residual building blocks in the encoder followed by a PSPPooling layer (Fig. 1b), ResUNet-a D7 has seven building blocks in the encoder.

The different types of layers of ResUNet-a have the following functions:

Input layer - This layer stores the input image, which is a 256 × 256 raster with four spectral bands (blue, green, red, near-infrared).

Conv2D layer - The Conv2D layer is a standard 2D convolution layer (kernel size = 3, padding = 1).

Conv2DN layer - This is a 2D convolution layer (kernel size = 3 and padding = 1) followed by a batch normalisation layer. By normalising the output of a previous activation layer by subtracting the batch mean and dividing by the batch standard deviation, batch normalisation is an efficient technique to combat the internal covariate shift problem and thus improves the speed, performance, and stability of artificial neural networks (Ioffe and Szegedy, 2015).

ResBlock-a layer - This layer follows the philosophy of the residual units (He et al., 2016b), *i.e.*, units that skip connections between layers to improve the flow of information between the first and the last layers. Instead of having a single residual branch that consists of two successive convolution layers, there are up to four parallel branches with increasingly larger dilation rates so that the input is simultaneously processed at multiple fields of view. The size of the input feature map dictates the dilation rates.

PSPPooling layer - The pyramid scene parsing pooling layer emphasises contextual information by applying maximum pooling at four different scales (Zhao et al., 2017). The first scale is a global max pooling. At the second scale, the feature map is divided into four equal areas and max pooling is performed in each of these areas, and so on for the next two scales. As a result, this layer encapsulates information about the dominant contextual features across scales.

Upsample layer - This layer consists of an initial upsampling of the feature map (interpolation) that is then followed by a Conv2DN layer. The size of the feature map is doubled and the number of filters in the feature map is halved.

Combine layer - This layer receives two inputs with the same number of filters in each feature map. It concatenates them and produces an output at the same scale, with the same number of filters as each of the input feature maps.

Output layer - This is a multitasked layer with conditioned inference that produces a total of four output layers (Fig. 1a): the extent mask, the boundary mask, the distance mask, and the reconstructed image of the input image in the Hue-Saturation-Value space. A network graph is constructed so as to take advantage of the inference results for the previous layers. The process starts by combining the first and the last layers of the UNet. The distance mask is the first output to be generated without the use of PSPPooling. Then, to produce the boundary mask, the distance transform is conflated with the last UNet layer using PSPPooling. The distance mask and the boundary mask are combined with the output features to generate the extent mask. A parallel fourth layer reconstructs the initial input image which also helps reduce the model variance (Diakogiannis et al., 2020).

2.1.2. Data augmentation

While convolutional neural networks and specifically UNets can integrate spatial information, they are not equivariant to transformations such as scale and rotation (Goodfellow et al., 2016). Data augmentation increases the variance of training data, which confers invariance on the network for certain transformations and boosts its ability to generalise. To this end, we flipped the original images (horizontal and vertical reflections; Fig. 2) and randomly modified their brightness. As the size of the training data set was sufficiently large to cover significant variance, we avoided random rotations and zoom in/out augmentations with reflect-padding because these may break field symmetry, *e.g.*, for fields under pivot irrigation.

2.1.3. Training

Most traditional deep learning methods commonly employ cross-entropy as the loss function for segmentation (Ronneberger et al., 2015). However, empirical evidence showed that the Dice loss can outperform cross-entropy for semantic segmentation problems (Milletari et al., 2016; Novikov et al., 2017). Here, we relied on the Tanimoto distance with complements (a variant of the Dice loss). The Tanimoto distance achieves faster training convergence than other loss functions irrespective of the weights initialisation and keeps a good balance across multiple tasks (Diakogiannis et al., 2020). Our loss function was defined as the average of the Tanimoto distance $T(\mathbf{p}, \mathbf{l})$ and its complement $T(\mathbf{1} - \mathbf{p}, \mathbf{1} - \mathbf{l})$:

$$\widetilde{T}\left(\mathbf{1}-\mathbf{p}, \mathbf{1}-\mathbf{l}\right) = \frac{T(\mathbf{p}, \mathbf{l}) + T(\mathbf{1}-\mathbf{p}, \mathbf{1}-\mathbf{l})}{2}. \tag{1}$$

with

$$T\left(\mathbf{p}, \mathbf{l}\right) = \frac{\sum_i p_i l_i}{\sum_i (p_i^2 + l_i^2) - \sum_i (p_i l_i)} \tag{2}$$

where $\mathbf{p} \equiv p_i \in [0,1]$, represents the vector of probabilities for the ith pixel, and l_i are the corresponding ground truth labels. For multiple segmentation tasks, the complete loss function was defined as the average of the loss of all tasks:

$$\widetilde{T}_{\mathrm{MTL}}(\mathbf{p}, \mathbf{l}) = 0.25[\widetilde{T}_{\mathrm{extent}}(\mathbf{p},\mathbf{l}) + \widetilde{T}_{\mathrm{boundary}}(\mathbf{p},\mathbf{l}) + \widetilde{T}_{\mathrm{distance}}(\mathbf{p},\mathbf{l}) + \widetilde{T}_{\mathrm{reconstruction}}(\mathbf{p},\mathbf{l})] \tag{3}$$

2.1.4. Inference

Once trained, the model can be applied to any a 256 × 256 image by a simple forward pass through the network. As convolutional neural networks use contextual information for prediction, it is likely that their classification accuracy varies depending on the location of objects in the input image. We expect that objects at the centre of input images are well classified because their context is evenly described in every direction. Objects near the edge of input images, however, are missing parts of their context and are thus likely to be misclassified. To mitigate this effect, we created 16 sets of input images by using a moving window of size 256 with stride 64 in every direction, effectively varying the position of image objects in the input images. Inference was performed on all sets of input images which were averaged using the arithmetic mean.

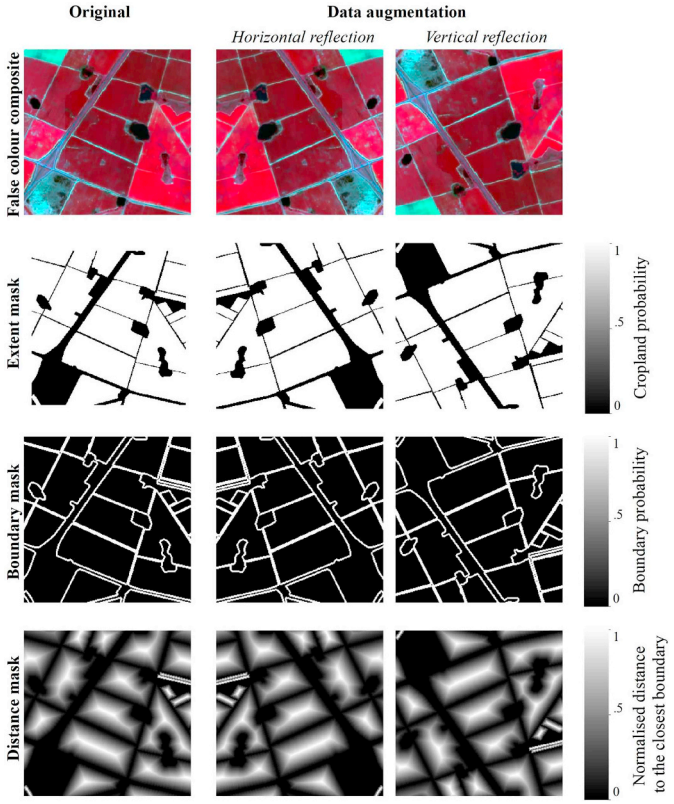


Fig. 2. Example of data augmentation for an input image of South Africa (256 × 256 pixels; 27°16′S, 27°28′E). The four rows represent the four inputs to the deep neural network: the satellite image, the extent mask, the boundary mask, and the distance mask. Feeding the network with random transformations of the original data is expected to improve its ability to generalise.

2.2. Extraction of individual fields

While ResUNet-a can learn to identify strictly field boundaries, it does not necessarily generate closed boundaries. We propose two post-processing methods, one based on thresholding (hereafter the *cutoff* method) and the other based on watershed segmentation (hereafter the *watershed* method), to retrieve closed boundaries and thus, to achieve instance segmentation and retrieve individual fields. Both leverage multiple semantic segmentation outputs and are data-driven. Thus, they can be automatically optimised using a small sample of reference data.

2.2.1. Post-processing methods for instance segmentation

The *cutoff* method delineates individual fields by thresholding the extent mask and the boundary mask (Fig. 3). A threshold on the extent map defines the extent of fields and, by extension, boundaries between "field" and "non-field" objects. A threshold on the boundary mask helps further define boundaries between adjacent fields. The symmetric difference of these two binary layers (the set of pixels which are in either of the layers but not in their intersection) produces individual fields with closed boundaries.

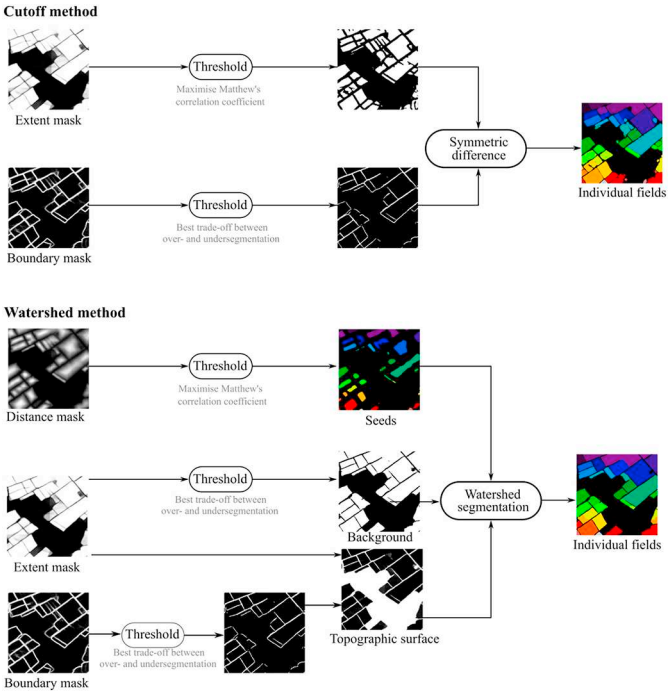


Fig. 3. Post-processing methods to generate closed boundaries and extract individual fields. The *cutoff* method generates individual fields by thresholding and combining the boundary mask and the extent mask. The *watershed* method extracts fields by all segmentation masks in a watershed algorithm. Seeds for the watershed algorithm are extracted from the distance mask.

The *watershed* method delineates individual fields by applying a seeded watershed segmentation algorithm on the three output masks. Seeded watershed segmentation considers an input image as a topographic surface (where high pixel values mean high elevation). It then simulates the flooding the topographic surface from specific seed points (Meyer and Beucher, 1990), thus generating different growing catchment basins. Catchment basins stop growing when they hit other catchment basins; the lines that separate adjacent catchment basins are their boundaries. By analogy, we can identify the centres of each field (seeds) and grow them until they hit other field boundaries or reach pixels that do not belong to fields (background). Seeds are defined by thresholding the distance mask and the topographic surface is defined by thresholding the extent mask (Fig. 3). The background, *i.e.*, part of the image that does not belong to any field, is defined by thresholding the boundary mask. Unlike the cutoff method that requires two thresholds, three thresholds must be set for the watershed method, one per segmentation mask.

2.2.2. Threshold optimisation

The thresholds of the post-processing methods are instantiated using a two-step optimising process: the first step optimises the extent of fields that are extracted (threshold on the extent mask); the second step optimises the fields' shape and size (thresholds on the boundary and distance masks). This procedure thus requires that reference data are available to identify to optimal threshold values (see Section 3.3).

First, the threshold for the extent mask is defined by maximising the Matthew's correlation coefficient (MCC; Matthews, 1975) between the thresholded and a reference map of the fields' extent (cropland map). The MCC is computed as follows:

$$\mathrm{MCC} = \frac{\mathrm{TP}\,\mathrm{TN} - \mathrm{FP}\,\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FN})(\mathrm{TP}+\mathrm{FP})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \tag{4}$$

where TP, TN, FP and FN are the true positive rate, the true negative rate, the false positive rate and the false negative rate. The MCC varies between -1 (perfect disagreement) and $+1$ (perfect agreement) while 0 indicates an accuracy no different from chance. As MCC uses all four cells of a binary confusion matrix, it is a robust indicator when data are unbalanced (Boughorbel et al., 2017), which is typically the case for boundary detection.

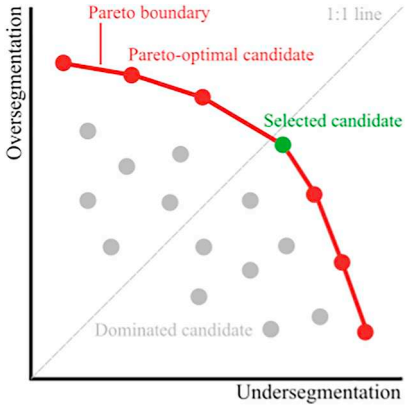


Fig. 4. Identification of the optimal threshold among the Pareto-optimal candidates for instance segmentation. The optimal threshold provides the best balance between undersegmentation and oversegmentation.

Second, the thresholds of the boundary and distance masks are jointly optimised to minimise incorrect subdivision of larger objects into smaller ones (oversegmentation) and an incorrect consolidation of small adjacent objects into larger ones (undersegmentation). The oversegmentation rate (S_{over}) and undersegmentation rate (S_{under}) were computed from the reference fields (T) and the extracted fields (E) data using the following formulae (Persello and Bruzzone, 2010):

$$S_{\mathrm{over}} = 1 - \frac{T_i \cap E_j}{T_i} \tag{5}$$

$$S_{\mathrm{under}} = 1 - \frac{T_i \cap E_j}{E_j} \tag{6}$$

where $|\cdot|$ is an operator that calculates the area of a field from E or T and $\cap$ is the intersection operator. These metrics provide rate values ranging from 0 to 1: the closer to 1, the less prone to over- and undersegmentation. If an extracted field E_j intersects with more than one field in T, the location error, and the over- and undersegmentation rates are weighted by the corresponding intersection areas.

The average over- and undersegmentation rates are tuned using a multi-objective optimisation procedure that first identified all Pareto-optimal (Coello, 2000) candidates among those generated. The Pareto-optimal candidates are those candidates for which it is impossible to improve oversegmentation without deteriorating undersegmentation, and *vice versa*. Finally, the optimal thresholds are given by the Pareto-optimal candidate providing the best trade-off between over- and undersegmentation, *i.e.*, the closest to the 1:1 line (Fig. 4).

3. Data and study sites

3.1. Study sites

Our experiments covered one main study and five secondary sites, all part of the Joint Experiment for Crop Assessment and Monitoring (JECAM[2]) network. The purpose of JECAM is to compare data and methods for crop monitoring, with the aim of establishing best practices for different agricultural systems. By selecting JECAM sites, we sought to facilitate future method benchmarking exercises. The main study area encompasses 120,000 km^2 of South Africa's “Maize quadrangle”,

[2] www.jecam.org.

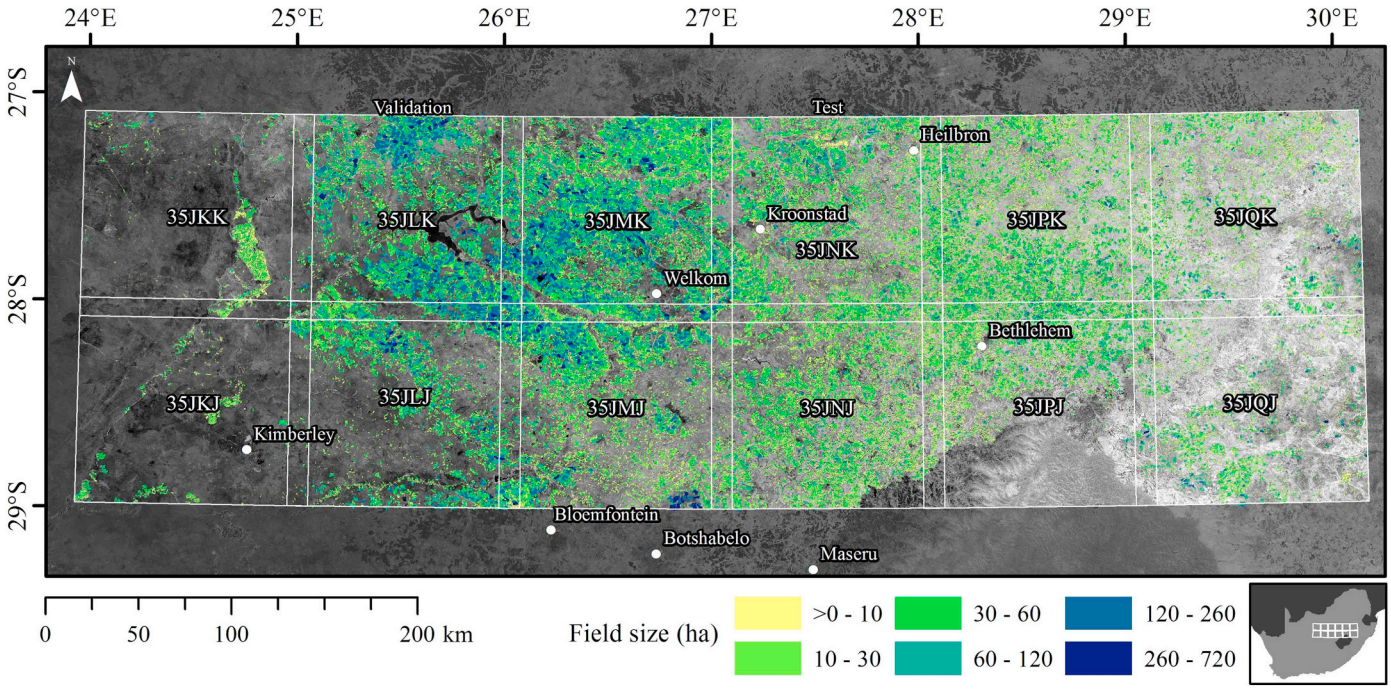


Fig. 5. Location of the 208,667 fields available for training and validation in South Africa, the main study site. The field colour indicates the size. Fields in 35JLK were used for validation, and 35JNK for testing. Fields from other tiles were used for training. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

which spans across the major maize-producing area. The field size averages 17 ha and ranges from 1 ha to 830 ha (Fig. 5). The five secondary study sites span 10,000 km^2 each and are located in Argentina, Australia, Canada, Russia, and Ukraine (Fig. S1). They were selected to sample broad-acre cropping systems covering a variety of climate regions, field size, crop types and crop calendar (Table 1).

3.2. Satellite data and preprocessing

Twelve Sentinel-2 tiles cover the main study site and each secondary site was defined by the extent of a single Sentinel-2 tile (Table 2). We chose Sentinel-2 imagery for its 5-day global revisit frequency, which increases the likelihood of cloud-free image acquisition and provides consistent multi-temporal composites, and for its 10-m resolution, which is sufficient to resolve most fields in South Africa (Waldner et al., 2018). Sentinel-2 images are made available free of charge, which facilitates large-scale applications.

Sentinel-2 images were obtained from the European Space Agency's Scientific Hub and were converted to surface reflectance using the Multisensor Atmospheric Correction and Cloud Screening (MACCS; Hagolle et al., 2010) algorithm available in the Sen2-Agri toolbox (Defourny et al., 2019). The same toolbox was used to generate a monthly cloud-free composite of March 2017, which corresponds to the middle of the growing season. Across the images available for compositing, 93% of the pixels were cloud-free. Five or six cloud-free Sentinel-2 images were processed in each secondary site (Table 2). We also sourced the closest cloud-free Landsat-8 image to the Sentinel-2 compositing window (Table 2).

There were only two preprocessing steps: subsetting the blue, green, red, and near infrared bands from the satellite images; and standardising pixel values so that each band had a mean of zero and a standard deviation equal to one. Standardisation is necessary because neural

Table 1
Description of the main and secondary sites. The main study site is 120,000 km^2 and the secondary sites are 10,000 km^2.

Main characteristics	Main crops and crop calendar
South Africa (main site) – 28° S, 27°E	
• Sub-humid to semi-arid climate • Average field size of 17 ha • Flat undulating plains to mountainous	• Maize, wheat, sunflower and soybean with some irrigation (centre pivot) • From December to June; from April to November
Argentina – 34°32′S, 59°06′W	
• Temperate humid climate • Average field size of 20 ha (with high variability) • Gentle slopes	• Soybean, maize and wheat (mostly rainfed) • From June to December and from October to March/April
Australia – 36°12′S, 143°33′E	
• Temperate climate with winter-dominant rainfalls • Fields as large as 250 ha • Gently undulating landscape	• Wheat, barley, canola, and legume crops • From April to December
Canada – 50°54′N, 97°45′W	
• Humid continental climate • Average field size of 64 ha • Flat	• Canola, wheat, soybean, maize • From April to August and May to September/October
Russia – 45°98′N, 42°99′E	
• Arid to humid continental climate • Fields ranging from 30 to 130 ha • Mostly flat	• Wheat, barley, peas, soybean, sunflower and rapeseed • From April to October; From September to July
Ukraine – 50°51′N, 29°96′E	
• Humid continental climate • Fields ranging from 30 to 250 ha • Mostly flat with some hills	• Wheat, sunflower, maize, barley, potatoes, and soybean • From September to July; from April to October

Table 2
Summary of the images used in every experiment of this study. The coordinates indicate the centroid of each site.

Sensor	Site	Tiles	Acquisition dates/window
4.1. and 4.2. Baseline and comparison with edge filters			
Sentinel-2	South Africa	T35JKJ, T35JKK, T35JLJ, T35JLK, T35JMJ, T35JMK, T35JNJ, T35JNK, T35JPJ, T35JPK, T35JQJ, and T35JQK	All images in 03/2017
4.3.1. Generalisation to single-date imagery: Can the model be applied to single-date images?			
Sentinel-2	South Africa	T35JNK	13/03/2017
4.3.2. Generalisation across resolutions and sensors: Can the model be applied to a 30-m Sentinel-2 image and to a Landsat-8 image?			
Sentinel-2	South Africa	T35JNK	13/03/2017
Landsat-8	South Africa	170–079	29/03/2017
4.3.3. Space-time generalisation: Can the model be applied to other locations and acquisition dates?			
Sentinel-2	Argentina	T21HTB	09/01, 03/02, 10/03, 20/03, 14/04, and 23/07/2018
	Australia	T54HXE	18/07, 02/08, 11/09, 26/09, 01/10, and 11/10/2018
	Canada	T14UNA	26/04, 16/05, 10/06, 15/07, 09/08, and 23/10/2018
	South Africa	T35JNK	13/11 and 03/12/2016; 02/01, 02/04, 01/06, and 11/06/2017
	Russia	T35UPR	08/04, 02/06, 22/06, 27/07, and 11/08/2018
	Ukraine	T37TGL	21/04, 11/05, 05/07, 24/08, 13/10, and 18/10/2018

Table 3
Accuracy of the two model architectures for different training modes.

Training mode	Training loss	Validation loss	Validation MCC
ResUNet-a D6			
WD = 10^{-4}	0.1212	0.1310	0.8109
WD = 10^{-5}	0.1118	0.1326	0.8099
WD = 10^{-6}	0.1095	0.1310	0.8109
Interactive	0.0947	0.1344	0.8091
ResUNet-a D7			
WD 10^{-4}	0.1119	0.1369	0.8012
WD 10^{-5}	0.1017	0.1339	0.8086
WD 10^{-6}	0.0979	0.1326	0.8099
Interactive	**0.0801**	**0.1285**	**0.8158**

The best results are emboldened.

networks are sensitive to the scale of the input variables and because gradient optimisation methods converge more rapidly with when features have zero mean and unit variance. Standardisation values were obtained from the monthly composite image only, not from images from the secondary sites.

3.3. Reference and ancillary data

For the main site, we obtained boundaries for every field in the area of interest from the South African Department of Agriculture, Forestry and Fisheries (Crop Estimates Consortium, 2017). These were created by manually digitising all fields throughout the country based on 2.5 m resolution, pan-merged SPOT imagery acquired between 2015 and 2017. Field polygons were rasterised at 10 m to match Sentinel-2's grid, providing a wall-to-wall validation data set rich of > 2 billion reference pixels. We created three reference layers: a binary layer of the extent of fields, a binary layer of the field boundaries (a 10-m buffer was applied), and a continuous layer representing, for every within-field pixel, the distance to the closest boundary. Distances were normalised per field so that the largest distance was always one. “Non-field” pixels were set to zero.

We randomly split the twelve Sentinel-2 tiles into a training (10 tiles), a validation (1 tile, 35JLK) and a test (1 tile, 35JNK) set. Sentinel-2 images were partitioned into a set of smaller images (hereafter referred to as input images) of a size of 256 × 256 pixels because it is not possible to process entire images at once due to limited GPU memory. Input images on the border of tiles, *i.e.*, those with missing values, were discarded in further analyses. This yielded an average of 6150 input images per tile. Input images in the training set were used to train the deep neural network, those in the validation set were used for parameter tuning, and those in the test set were utilised only for final accuracy assessment. Discrepancies between field boundaries and the Sentinel-2 images were not necessarily concomitantly acquired. While these discrepancies might be handled during the training phase, their impact is far greater for accuracy assessment. Therefore, we followed a random sampling procedure to visually confirm 1000 field boundaries.

For the secondary sites, we first created the layer of the fields' extent for each site by intersecting two global 30-m cropland maps: Globeland 30 (Chen et al., 2015b) and Global Food Security-support Analysis Data (GFSAD) Cropland Extent (Phalke et al., 2017; Massey et al., 2017; Zhong et al., 2017). We resampled this intersection map to 10 m and used information about water bodies, roads, and railways from OpenStreetMap to further mask out pixels wrongly labelled as cropland. Two hundred randomly-selected fields were then manually digitised, for each secondary site, based on Sentinel-2 images and Google Earth images.

4. Methods

We conducted several experiments to evaluate the ability of our method to extract field boundaries from satellite images. In a first, baseline experiment, we evaluated the accuracy of our approach in the main site using a single monthly composite of Sentinel-2. Our second experiment compared ResUNet-a's performance to identify field boundaries against that of a conventional edge detection filter. Our last set of experiments was designed to evaluate the generalisation ability of our approach. Specifically, we assessed its ability to generalise to single-date imagery, to coarser resolutions and different sensors, and to other dates and locations. With these experiments, we wish to gather the necessary evidence to propose guidelines to inform large-scale field boundary extraction with deep learning.

4.1. Baseline: ResUnet-a D6 vs. ResUnet-a D7

4.1.1. Model training

We trained a series of ResUnet-a D6 and D7 models using the Adam optimiser because it achieves faster convergence than other alternatives (Kingma and Ba, 2014). Adam computes individual adaptive learning rates for different parameters from estimates of first and second moments of the gradients. We followed the parameter settings

recommended in Kingma and Ba (2014).

Two training approaches were tested. The weight decay (WD) approach adds a weight decay parameter to the learning rate in order to decrease the neural network weight and bias values by a small amount in each training iteration when they are updated. This technique can speed up convergence and is equivalent to introducing an L2 regularisation term to the loss function. We trained three models for 100 epochs with different weight decay parameters (10^{-4}, 10^{-5}, 10^{-6}). The interactive approach involves training a neural network for 100 epochs with a given learning rate (10^{-4}) and resuming training for another 100 epochs with a reduced learning rate at the epoch that reached the highest MCC. This process was repeated twice for increasingly lower learning rates (10^{-5} and 1^{-6}). Models were trained on a single node on a high-performance computing system, consuming four NVIDIA Tesla P100 GPUs for a maximum seven days, depending on the training approach. As a result, eight models were compared (three weight decay models and an interactive model for both ResUnet-a D6 and D7).

Finally, individual fields were extracted using the outputs of the best model. We compared the two post-processing methods and assessed the significance of their differences using paired Wilcoxon tests (Wilcoxon, 1945), a nonparametric test for paired data that compares the locations of two populations to determine if one is shifted with respect to the other. For the *cutoff* method, the optimal thresholds were identified using grid search (between 0.01 and 0.99 by step of 0.01). Given the large number of possible combinations in the *watershed* method, a random search of 250 threshold combinations was used instead because it scales better (Bergstra and Bengio, 2012).

4.1.2. Accuracy assessment

Model accuracy was evaluated with two types of accuracy metrics: pixel-based metrics and object-based metrics. Object-based metrics latter were particularly important to quantify the accuracy of individual fields that were extracted.

Pixel-level accuracy was computed from two global accuracy metrics derived from the error matrix: Matthew's correlation coefficient; and the overall accuracy (OA), which provides the proportion of pixels that were correctly classified:

$$\mathrm{OA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP}} \tag{7}$$

We also computed the F-score (F) as a class-wise accuracy indicator:

$$\mathrm{F}_k = \frac{2 \times UA_k \times PA_k}{UA_k + PA_k} \tag{8}$$

where UA_k is the user's accuracy of class k ($UA = \frac{TP}{TP + FN}$) and PA_k is the producer's accuracy ($PA = \frac{TP}{TP + FP}$). For experiments other than those involving South African data, the hit rate was computed as the ratio between the number of fields in the validation data set that were successfully detected and the total number of fields in the validation data set. The hit rate was the preferred thematic accuracy metric because the cropland map in these sites, despite our best efforts, was not accurate enough to be considered as validation data.

Object-based accuracy was evaluated with four metrics that collectively captured differences in shape, size and shifts in the location of extracted fields with respect to target or reference fields. The first two (the oversegmentation rate and the undersegmentation rate) were previously introduced and express incorrect subdivision or incorrect consolidation of fields. The third one, the eccentricity factor (ε) reflects absolute differences in shape (Persello and Bruzzone, 2010):

$$\varepsilon = \|\mathrm{Eccentricity}_{T_i} - \mathrm{Eccentricity}_{E_j}\| \tag{9}$$

where the eccentricity indicates how much the shape of a field deviates from a circle (Eccentricity = 0). The fourth and last object-based metric, the location shift (L), was computed to indicate the difference between the centroid location of reference and extracted fields (Zhan et al., 2005):

$$L = \sqrt{(x_E - x_T)^2 + (y_E - y_T)^2} \tag{10}$$

where (x_E, y_E) and (x_T, y_T) are the centroids of corresponding fields E and T. Here, the location shift was expressed in pixels. If a reference field was matched with multiple extracted fields, the location shift was defined as the area-weighted average all location shifts.

We evaluated the significance of the difference in accuracy bewteen the two post-processing methods with paired Wilcoxon signed-rank tests and declared statistical significance for P values < 0.05.

4.2. Comparison with conventional edge detection

We compared the boundary mask of the model against the edges detected with a Scharr filter. The Scharr filter, which has a better rotation invariance than other oft-used filters, such as the Sobel or the Prewitt operators, identifies edge magnitudes by taking the square root of the sum of the squares of horizontal and vertical edges. Edges were detected in each spectral band then averaged. We randomly sampled 1000 boundary and interior pixels from this layer along with the corresponding pixels from the boundary mask. As the edge detection layer indicates a magnitude rather than a probability, we computed pseudoprobabilities by scaling the magnitude values between 0 and 1 based on their 0.05 and 0.95 percentiles. We then compared the pseudoprobabilities of boundary and interior pixels against the corresponding probabilities of the boundary mask with paired Wilcoxon signed-rank tests.

4.3. Generalisation experiments

The previous experiment set the baseline for field boundary extraction from a Sentinel-2 image composite. Next, we designed a series of experiments to evaluate the model ability to generalise to a cloud-free single-date image close to the compositing window (Section 4.3.1), to the composite image resampled to 30 m and to a Landsat-8 image close to the compositing window (Section 4.3.2), to Sentinel-2 images acquired across the season in the main and in the secondary sites (Section 4.3.3).

4.3.1. Generalisation to a single-date image

Monthly composites are attractive for training because, by removing pixels contaminated by clouds and shadows, they maximise the use of training data which are costly to develop. For inference, however, they are less practical than single-date images because of their additional preprocessing needs. Therefore, we tested the assumption that, for inference, composites could be replaced by single-date images without significantly reducing accuracy. We applied the best ResUNet-a to a cloud-free Sentinel-2 image covering the test area in South Africa and evaluated the accuracy obtained at the pixel and field levels.

4.3.2. Generalisation across resolutions and sensors

We evaluated the model's ability to generalise to another sensor coarser by applying it to a single-date Landsat-8 image covering the South African test site. For comparison, we resampled the monthly Sentinel-2 composite to 30-m to match the grid of the Landsat-8 image using a nearest neighbour algorithm. This algorithm is naive because it neglects the sensor spatial response (Waldner et al., 2018). Therefore, the fields extracted using the resampled image are expected to be more accurate.

4.3.3. Generalisation across space and time

Finally, we assessed the model's ability to generalise in space and time by extracting field boundaries in the main study site and in the secondary sites using cloud-free single-date images across the growing

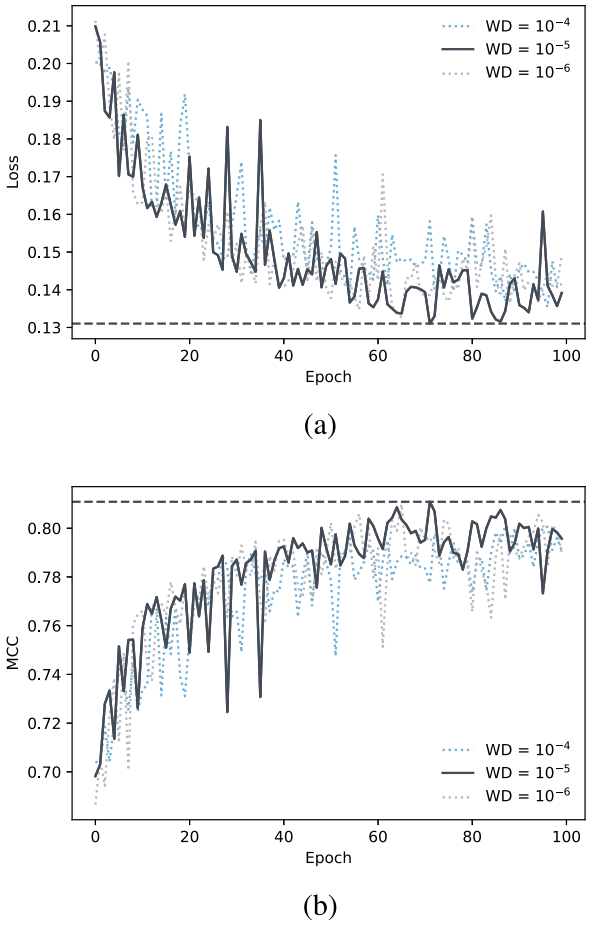


Fig. 6. Evolution of the (a) loss function and (b) Matthew's Correlation Coefficient during model training for the procedures involving weight decay (WD). Optimal training (MCC = 0.81) was achieved before 80 epochs, then earlier signs of overfitting can be seen.

season. Thematic and geometric changes in accuracy were measured over time for each single-date image.

We hypothesised that, given the dynamic nature of cropping systems, some dates would be more appropriate than others. Therefore, we also evaluated the benefit of building consensus from multiple dates as a mechanism to prevent loss of accuracy. Consensus is built by averaging the segmentation masks of multiple observation dates. To identify the number of observations that delivers consistent improvement, consensus was gradually built along the season. For example, consensus for the third acquisition date was obtained by time-averaging the segmentation mask of the first, second and third acquisition dates. Individual fields were then extracted from the time-averaged masks.

5. Results

5.1. Model selection

We trained four versions of the ResUNet-a D6 and D7 models; the first three versions were parameterised with different weight decay values and, the last version was parameterised interactively. After each epoch, the loss function and the MCC were computed for the test set (see Fig. 6 for the D7 model and the three training modes involving weight decay).

Interactive training produced the lowest training loss for both model architectures. However, when validated against the test data, the best D6 model was the one trained with highest weight decay performed best (MCC = 0.81). We assumed that training the models for > 100 epochs was unlikely to improve accuracy because the training curves started to show signs of overfitting after 80 epochs. Overall, ResUNet-a D7, the deeper model, yielded the highest accuracy (MCC = 0.82). Thus this model was selected for all subsequent analyses.

Table 4
Pixel-based assessment of the extent map of South Africa for the baseline experiment, the generalisation to single-date imagery, and to a resampled 30-m Sentinel-2 image.

Threshold	OA	MCC	F_C	F_{NC}
Baseline				
Default threshold (50%)	91.69	82.23	89.09	93.29
Optimised threshold (38%)	91.66	82.46	89.26	93.18
Generalisation to a single-date image				
Optimised threshold (38%)	0.893	0.782	0.872	0.909
Generalisation to a resampled 30-m image				
Optimised threshold (61%)	0.856	0.696	0.798	0.889

OA: Overall accuracy; MCC: Matthew's correlation coefficient.
F_C: F-score for the cropland class; F_{NC}: F-score for the non-cropland class.

5.2. Assessment of the baseline model

We assessed the accuracy of the ResUNet-a D7 model using pixel-based accuracy metrics (Table 4). The overall accuracy was 92% and the MCC reached 82%. The cropland class had a slightly lower F-score (89%) than the non-cropland class (93%). Tuning threshold of the extent map to maximise MCC led to only marginal differences (< 1%) compared to using a default threshold of 50%. Nonetheless, as we sought to achieve the classification with the most balanced accuracy for cropland and non-cropland, we retained the optimised threshold.

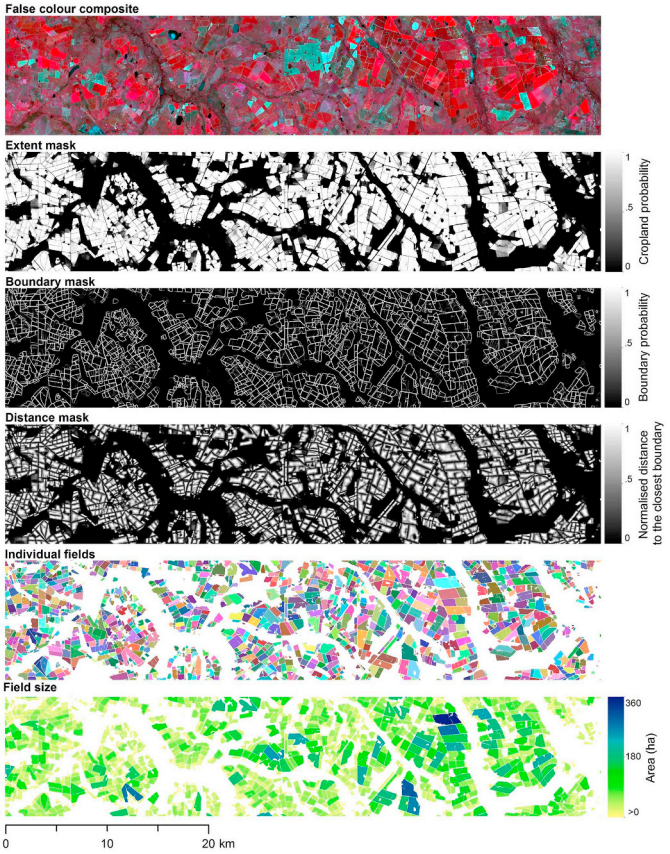


Fig. 7. Field extraction examples in South Africa. Each inset is 77 km × 40 km centred on 27°E 45″, 27°S 33″.

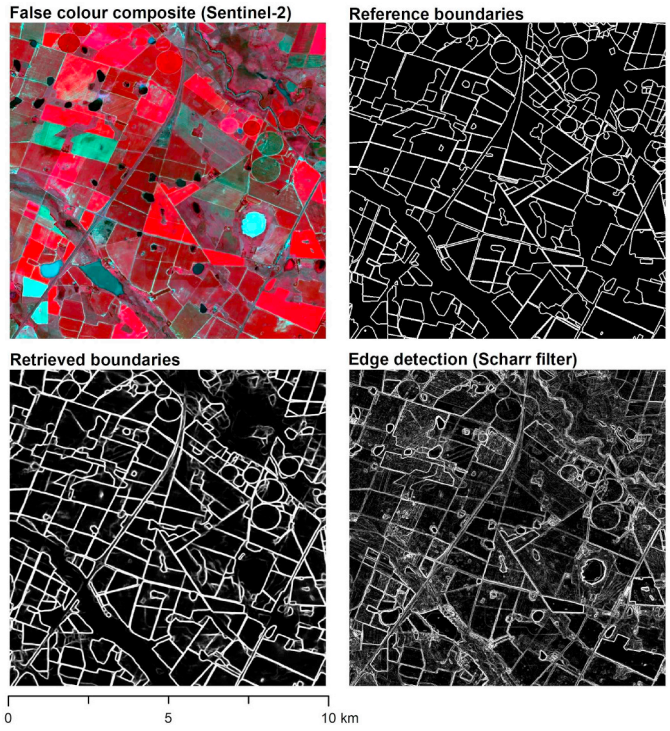


Fig. 8. Comparison of the boundaries retrieved by our model against references boundaries and edges detected with a Scharr filter for a 10-km × 10-km region in the test region (27°E 49′, 27°S 30′).

We extracted 55,720 fields ranging up to 380 ha (mean = 13 ha) from the Sentinel-2 image of the test region in South Africa (Fig. 7). Ninety-nine per cent of the reference fields were identified (Table 4) with high accuracy for both shape (all metrics > 0.85) and position (location shift of 7 pixels). Despite being simpler, the *cutoff* approach achieved similar results to the *watershed* approach. This might be explained by the number of possible combinations of parameters to optimise in the *watershed* approach. A more advanced optimisation method, such as the Bayesian optimisation, may be key to find the optimal combination. We report also no clear benefit in tuning the cutoff values compared to using a default cutoff value of 50%—the oversegmentation rate was the only metric for which the impact was > 0.05.

5.3. Comparison with conventional edge detection

We compared the boundary mask retrieved by our ResUNet-a against the edges detected by a Scharr filter (Fig. 8). The edge-based method yielded significantly weaker boundaries ($P < 0.001$) and noisier interiors ($P < 0.001$). With conventional edge detection, interior pixel values were on average higher and spread across a wider range than those obtained with ResUNet-a. As the neural network learns to be sensitive to certain types of edges, the retrieved edges become clearer.

5.4. Generalisation to single-date imagery

Feeding the model with single-date data instead of a monthly composite had little effect on its performance (Table 5). For instance, the over- and undersegmentation rates dropped by 0.02–0.05 while the offset remained unchanged. These differences were statistically significant only for the undersegmentation rate ($P < 0.001$), which suggests that extracting field boundaries from single-date images is a viable option.

5.5. Generalisation across resolutions and sensors

Resampling to 30 m reduced the hit rate by only 0.06. This loss of accuracy illustrates the impact of the sensor spatial response (which blurs the image and was neglected during resampling) on the ability to detect cropland (Waldner et al., 2018).

For individual field extraction, our results showed that the model was more sensitive to a change of resolution than to a change of sensor (Table 5). When extracting field boundaries from 30-m rather than 10-m Sentinel-2 data, only the hit rate changed significantly (0.88 to 0.79). These changes ought to be related to loss of spatial details and increased difficulty of extracting boundaries, especially in more fragmented landscapes. On average, the size of the undetected fields (15 ha) was smaller than that of detected fields, which was significant ($P < 0.001$). We concluded that the model exhibits good generalisation capabilities across scales and sensors because it achieved similar performances with Landsat-8 data and 30-m Sentinel data.

5.6. Generalisation across time and space

The acquisition date had a significant impact on the accuracy of the field extraction process in South Africa (Fig. 9). The hit rate and the location shift were particularly affected, with changes from 0.75 to 0.99 and 7 pixels to 17 pixels, respectively. Building consensus successfully reduced variability: it achieved equal, if not better, performance than single-date cases. The rate of improvement due to consensus reduced after four images.

The impact of changes in acquisition dates was even more marked in secondary sites (Fig. 9). In Canada, for instance, the hit rate ranged from 0.21 to 0.97 and the oversegmentation rate from 0.26 to 0.80. Of all the metrics, eccentricity was the least sensitive. Unlike in the main site, it was critical to optimise thresholds in secondary sites, which indicates that local parameter tuning is required for good generalisation.

While it was possible to yield high accuracy with a single image, the success of the extraction was highly variable, depending on the date of image acquisition. Building consensus was highly effective at reducing the variability in accuracy, which improved the model's ability to generalise across the board (Figs. 9 and 10). For instance, the hit rate after consensus exceeded 0.95 in every site except Australia (0.86). The retrieved fields and matched the geometry of reference fields, with under- and oversegmentation rates ranging from 0.7 to 0.86. The benefit of consensus plateaued after combining four images. Fig. 11 illustrates the boundary extraction process in all secondary sites.

6. Discussion

6.1. Deep learning for field boundary extraction

We extracted field boundaries from satellite images with a convolutional neural network called ResUNet-a which was trained to solve multiple, correlated semantic segmentation tasks. We initially trained a ResUNet-a model with a monthly Sentinel-2 composite covering South Africa and achieved high accuracy at the pixel and field levels. The same model, without retraining, performed accurately when applied to single-date imagery of the same site acquired by Sentinel-2 and Landsat-8. The ability to extract boundary from a single-date image is cost-effective and a great advantage in areas where cloud cover is persistent such as in the tropics. The model also generalised well in five different cropping systems and for a range of acquisition dates. Building consensus by time-averaging predictions from multiple single-date images further improved accuracy. Such generalisation abilities clearly indicate that our approach successfully minimised over-fitting, a prevalent problem in deep learning. It suggests that convolutional neural networks reduce the importance of spectral and temporal information by heavily exploiting multilevel contextual information.

Table 5
Object-based assessment for experiment, for the generalisation to single-date imagery, for the generalisation to other resolutions and sensors, and across space and time. The range of over- and undersegmentation, eccentricity, and hit rate is from 0 to 1, with 1. The location shift is given in pixels.

Image	Location	Post-processing	Hit rate	Oversegmentation	Undersegmentation	Eccentricity	Shift
Baseline							
Sentinel-2	South Africa	Naive	0.994	0.870	0.869	0.997	7
		Watershed	0.995	0.870	0.858	0.997	6
Generalisation to a single-date image							
Sentinel-2	South Africa	Cascade Watershed	0.990	0.849	0.847	0.998	7
Generalisation across resolutions and sensors							
Sentinel-2 30 m	South Africa	Naive	0.883	0.695	0.696	0.996	5
Landsat-8	South Africa	Naive	0.791	0.665	0.673	0.995	6
Generalisation across space and time - consensus approach							
Sentinel-2	Argentina	Naive	0.965	0.818	0.701	0.910	20
	Australia	Naive	0.886	0.764	0.806	0.788	20
	Canada	Naive	0.965	0.796	0.800	0.819	20
	Russia	Naive	0.935	0.828	0.830	0.955	22
	Ukraine	Naive	0.980	0.864	0.855	0.890	14

This study shows that addressing the problem of field boundary extraction from satellite images as multiple semantic segmentation tasks with a convolutional neural network achieves excellent performance (> 85%) both at the pixel level and the object level. We believe this is because our neural network learned to discard edges that are not part of field boundaries and to emphasise those that are using spectral and multilevel contextual features that are relevant for multiple correlated tasks. As a result, it identifies significantly sharper boundaries with significantly than a benchmark edge-based method. The extent of the cropping area in the main site was mapped with an accuracy of ~90%, which is comparable to state-of-the-art methods for cropland classification (*e.g.*, Inglada et al., 2017; Waldner et al., 2017; Oliphant et al., 2019). This is nonetheless remarkable because our method did so with a single monthly composite. Most published works argue that multitemporal features are crucial to accurately map cropland because they capture the dynamics of a pixel across the growing season. This suggests that convolutional neural networks reduce the importance of temporal information by heavily exploiting hierarchical contextual information. However, reported difficulties in distinguishing certain classes (such as cropland from grassland) from a single image indicate that temporal information still needs to be considered in certain cases.

Our convolutional neural network can detect fields and their boundaries across different resolutions and different sensors. Convolutional neural networks are designed to be scale-invariant. The features created by the model are object-based: they do not rely on individual pixel values but on all those belong to an object. This mechanism does not depend on a particular resolution. The ability of ResUNet-a to derive field boundaries from coarser images might be related to the multiple parallel atrous convolutions that identify features at different scales. However, this remains to be verified experimentally. When we applied the model to 30-m Sentinel-2 data, the hit rate and the geometric accuracy decreased by 10–15%. We think that this reduction was driven smaller fields which could not be resolved at that resolution. Differences between the resolution of the reference (2.5 m) and extracted (30 m) data might introduce an artificial bias in the accuracy assessment. Applying the model to Landsat-8 data decreased the geometric accuracy to ~70% and the hit rate to ~80%. Differences in the spectral and spatial responses of Sentinel-2 and Landsat-8 partly explain this loss in accuracy. Nonetheless, the accuracy metrics were similar to those reported using multi-temporal Landsat images across South America (Graesser and Ramankutty, 2017). The multi-modal capabilities of our approach open up the possibility to operate cross-platform in cloud prone areas. They further hint at exploiting open image archives imagery, such as the Landsat collection (Egorov et al., 2019), to reconstruct temporal patterns of field sizes. Transfer to even coarser (> 30 m) and higher (< 10-m) resolutions remains to be empirically evaluated.

The date of image acquisition matters, especially when extracting fields in previously unseen areas. In all sites, accuracy varied with acquisition date, with differences as large as 75%. Part of this sensitivity may be attributed to the training data set that consisted solely of images from a single month. Training the neural network with images spanning the whole season or covering more locations would likely reduce this effect. However, high model performance (for at least one date per site) suggests that changes of local image contrast linked to crop calendars are the main cause of these variations in accuracy. Local knowledge of crop calendars and cropping practices might dictate which image to select. However, this is fraught with uncertainty and is subject to data availability.

Building consensus by averaging predictions over time is a simple, yet effective, solution to consistently improve accuracy. Consensus halved location errors and, in most cases, yielded accuracy at least as large as using single-date images. Although it had been previously suggested that combining multiple image dates improves boundary detection (Watkins and van Niekerk, 2019), ours is the first study to clearly demonstrate the relationship between accuracy and the number of images. In particular, we showed that predictions from four images well-spread across the growing season are required to achieve most benefits. By covering changes in crop development, the likelihood of capturing an image with sharp contrast increases. Alternatives to harness the temporal dimension include replacing the input image with temporal features (for instance, see Graesser and Ramankutty, 2017) or adding a temporal dimension to the convolution filters. Compared to these alternatives, the consensus approach has lighter processing requirements since most space agencies now directly provide surface reflectance images for download.

None of our experiments included transfer learning. Transfer learning adapts a neural network to a new region by freezing the feature extraction layers and fine tuning the last layers based on a small number of labelled data (Pan and Yang, 2009, see also Penatti et al., 2015). This means that our model might have been exposed to reflectance values that it had never been previously exposed to. Implementing transfer learning would amount to updating the multi-tasking head of ResUNet-a and would likely further boost accuracy. The promising generalisation capabilities seem to indicate that requirements in terms of training set size are relatively low.

6.2. Recommendations for large-scale field boundary extraction

In this paper, we gather empirical evidence to lay the blueprints of a data-driven system that can derive field boundaries at scale. Based on our results, recommendations can be made as to how to efficiently implement such a system using Sentinel-2 imagery.

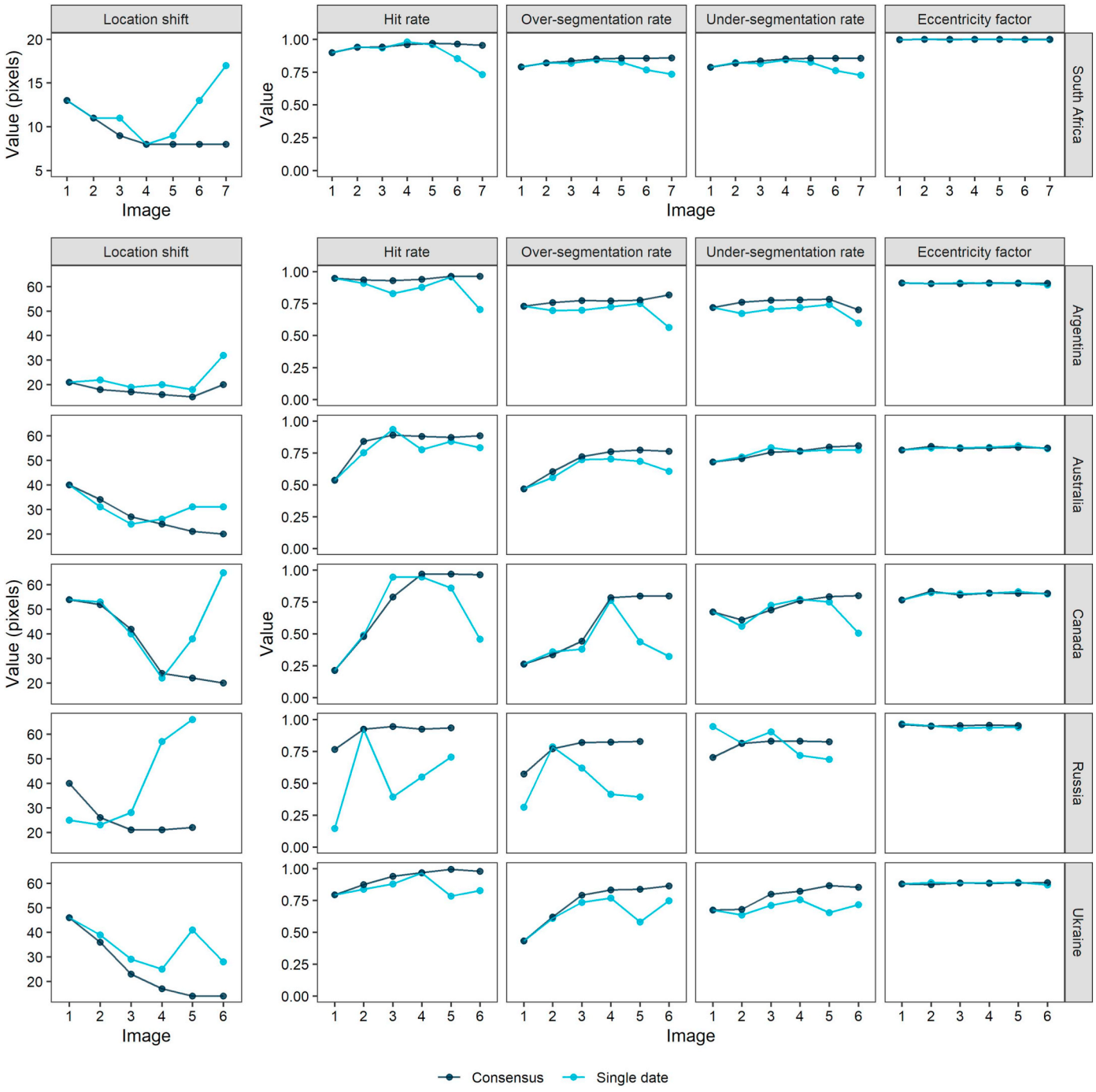


Fig. 9. Model accuracy for space-time generalisation. The x-axis represents the image position in the time series (single-date processing) or the number of images that were averaged to build consensus (so that consensus for image 3 is a time-average of image dates 1, 2 and 3). The model was able to generalise across space and time, but considerable temporal variability was observed. Building consensus is a simple and effective approach to mitigate this variability in accuracy: it was generally at least as good as single-date predictions. Most benefit of the consensus approach was realised with four images.

6.2.1. Train on composite images using blocks of continuous reference data

Deep learning relies on large amounts of training data, which are usually costly and time-consuming to collect. For field boundary detection, it appeared that the model was able to generalise across a range of locations from a local data set. This illustrates that, if one wishes to collect new training data sets, sample sizes are reasonable, especially as it can be artificially inflated using data augmentation techniques. New training data should be collected in blocks so that the model can learn to identify field boundaries in their context. While one must strive to collect accurate training data, ours were not completely free of errors but this did not seem to significantly impact the model's performance. Leveraging existing data sets, such as those from the European Land Parcel Identification System, is another option to cut down data collection costs. We also recommend training the model on monthly composites from across the season. Monthly composites are particularly appealing for two reasons: 1) in most cropping systems, they provide consistent, cloud-free observations across large areas, and 2) they minimise the amount of training data wasted because of contamination from clouds and cloud shadows. With Sentinel-2's five-day revisit frequency, several monthly composites per year should be obtainable in most cropping systems.

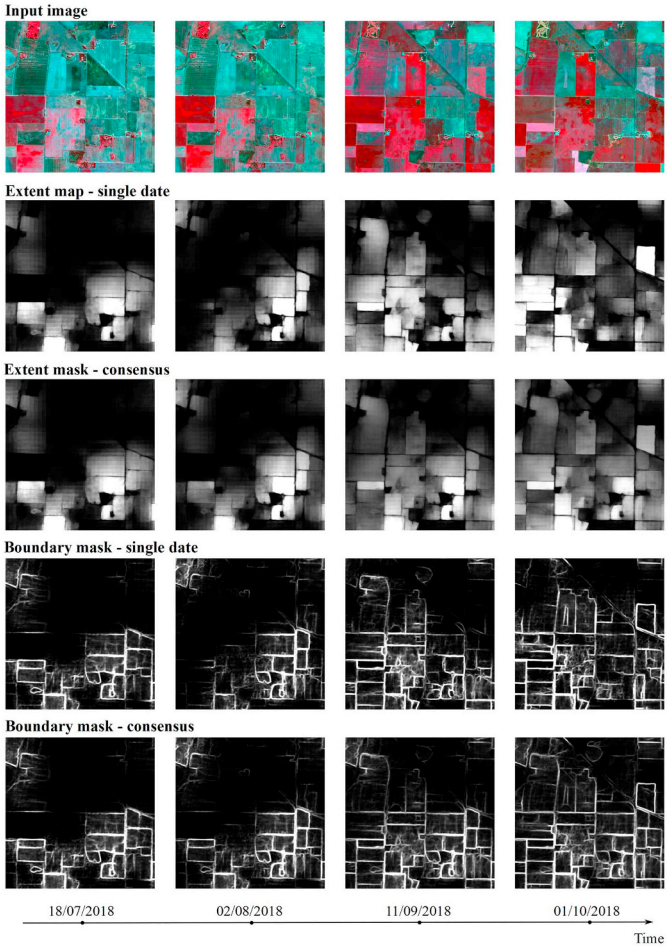


Fig. 10. Building consensus for an image subset in Australia. Single-date masks fail to capture all fields and all boundaries but missed patterns can be detected at later in the season. By averaging multiple predictions, the consensus approach is a computationally-cheap option to safeguard against loss of accuracy.

6.2.2. Predict on single-date images and build consensus predictions

This paper demonstrates that a model trained on a composite image can generalise to cloud-free single-date images. This is convenient because most space agencies are now providing surface reflectance data for download. Nonetheless, we highly recommend building consensus by averaging predictions from multiple dates to increase the robustness and confidence of the field extraction process. Our results indicate that averaging the averaging predictions from four images well-spread across the season safeguards against temporal variations of accuracy. Cloud-contaminated images can also be used for inference provided input images with clouds and shadows are properly discarded.

6.2.3. Optimise thresholds locally

Fine-tuning thresholds during post-processing had a significant impact when applying the model to previously-unseen locations. Unlike training data, which are to be collected in continuous blocks, reference data to adjust thresholds should be distributed across the region of interest and cover a range of field types. Optimal threshold values can then be determined locally, *e.g.*, with moving windows.

These evidence-based principles provide practical guidance for organising field boundary extraction across vast areas with an automated, data-driven approach and minimum image preprocessing requirements.

6.3. Perspectives

In advancing the ability to retrieve individual fields from satellite imagery, our work established that semantic segmentation combined with multitasking is a state-of-the-art approach. It also indicates future direction of research. Foremost among these is testing other image segmentation approaches, such as instance segmentation, where individual labels are assigned to objects of the same class, *e.g.*, with Faster Regional Convolutional Neural Network (R-CNN) or Mask R-CNN (Ren et al., 2015; He et al., 2017). These models have the potential to produce closed boundaries in a single pass, obviating any post-processing of the semantic segmentation outputs. However, empirical evidence showed that the accuracy of Mask R-CNN is similar to that of a UNet, and that combining their predictive performance produce more accurate results than either alone (Vuola et al., 2019).

Our work suggests further topics of investigation to better characterise the model's inner working and data requirements. These topics include 1) to identify the minimum viable training set size, which has direct implications cost of collecting reference data, 2) to evaluate the effect of errors in the training set (*e.g.*, missing fields or approximative field outlines) on accuracy, and 3) and to untangle the relationship between accuracy and crop growth processes and crop calendars in the areas being mapped. Finally, more work is needed to evaluate our approach in a wider range of cropping systems such as smallholder farming systems. In that case, very high spatial resolution is likely to be required to resolve small fields.

7. Conclusion

The ability to automatically extract field boundaries from satellite imagery is increasingly needed by many providers of digital agriculture services. In this paper, we formulated this problem as a semantic segmentation task and trained a deep convolutional neural network to solve it. Our model relied on multi-tasking and conditioned inference to predict, for each pixel, the probability of belonging to a field, to a field boundary, as well as to predict the distance to the closest boundary. These predictions were then post-processed to achieve instance segmentation and extract individual fields. By exploiting spectral and contextual information, our neural network demonstrated state-of-the-art performance for field boundary detection and good abilities to generalise across space, time, resolutions, and sensors. We believe that learning from multiple, correlated tasks is pivotal to minimise overfitting and achieve high generalisation.

Our work provides evidence-based principles for field boundary extraction at scale using deep learning: 1) to train models on monthly cloud-free composites to maximise usage of training data; 2) to predict field boundaries using single-date images because these can replace composites used for prediction with marginal loss of accuracy; 3) to build average predictions from at least four images to cope with the temporal variability of accuracy; 4) to use data-driven procedures to combine model outputs. These principles minimise image preprocessing and replace arbitrary parameter selection with data-driven processes. We think that following these principles will facilitate the production of accurate field boundary information required by many digital agriculture applications.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.rse.2020.111741.

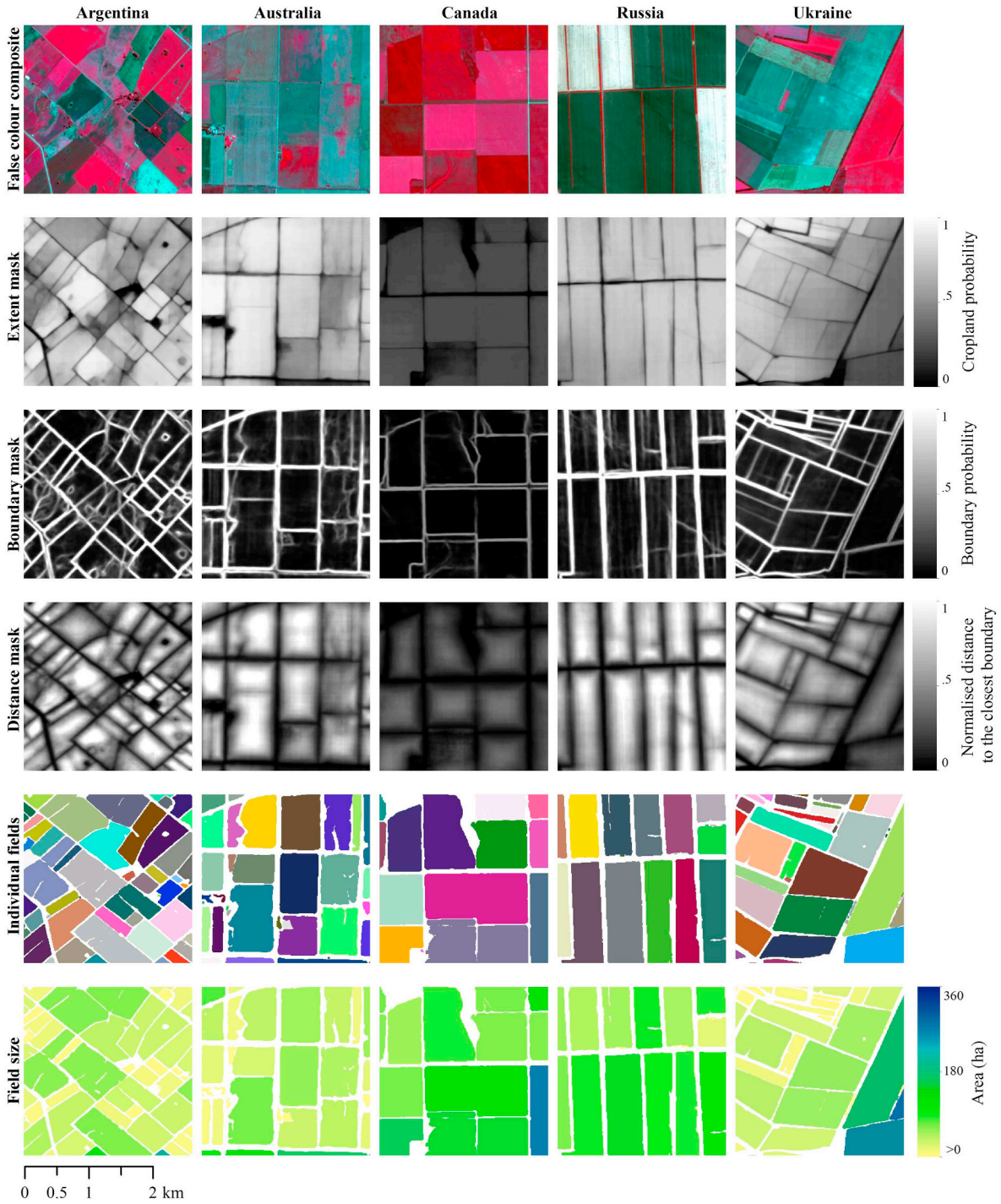


Fig. 11. Field extraction in the secondary sites with the consensus approach: Argentina (34°32′S, 59°06′W), Australia (36°12′S, 143°33′E), Canada (50°54′N, 97°45′W), Russia (45°98′N, 42°99′E), and Ukraine (50°51′N, 29°96′E). Each inset represents a 2.5-km × 2.5-km subset of the full 100-km × 100-km secondary sites.

CRediT authorship contribution statement

François Waldner: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft, Writing - review & editing. **Foivos I. Diakogiannis:** Methodology, Software, Supervision, Validation, Visualization, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We received funding from “Grains”, a project of the Digiscape Future Science Platform supported by the CSIRO. We are grateful to Dr. Sophie Bontemps and Annelize Collet for sharing their data. Finally, we would like to thank Dr. Robert Brown, Dr. Dave Henry, Dr. Joel Dabrowski and the three anonymous reviewers for their insightful comments on earlier versions of the paper.

References

Agravat, R.R., Raval, M.S., 2018. Deep learning for automated brain tumor segmentation in mri images. In: Soft Computing Based Medical Image Analysis. Elsevier, pp. 183–201.

Belgiu, M., Csillik, O., 2018. Sentinel-2 cropland mapping using pixel-based and object-based time-weighted dynamic time warping analysis. Remote Sens. Environ. 204, 509–523.

Bergstra, J., Bengio, Y., 2012. Random search for hyper-parameter optimization. J. Mach. Learn. Res. 13, 281–305.

Blaes, X., Vanhalle, L., Defourny, P., 2005. Efficiency of crop identification based on optical and SAR image time series. Remote Sens. Environ. 96, 352–365.

Borgefors, G., 1986. Distance transformations in digital images. Comput. Vision Graph. Image Process. 34, 344–371. https://doi.org/10.1016/S0734-189X(86)80047-0.

Boughorbel, S., Jarray, F., El-Anbari, M., 2017. Optimal classifier for imbalanced data using Matthews correlation coefficient metric. PLoS One 12, e0177678.

Brodrick, P.G., Davies, A.B., Asner, G.P., 2019. Uncovering ecological patterns with convolutional neural networks. Trends Ecol. Evol. 34 (8), 734–745.

Chai, D., Newsam, S., Zhang, H.K., Qiu, Y., Huang, J., 2019. Cloud and cloud shadow detection in Landsat imagery based on deep convolutional neural networks. Remote Sens. Environ. 225, 307–316.

Chen, B., Qiu, F., Wu, B., Du, H., 2015a. Image segmentation based on constrained spectral variance difference and edge penalty. Remote Sens. 7, 5980–6004.

Chen, J., Chen, J., Liao, A., Cao, X., Chen, L., Chen, X., He, C., Han, G., Peng, S., Lu, M., et al., 2015b. Global land cover mapping at 30 m resolution: a POK-based operational approach. ISPRS J. Photogramm. Remote Sens. 103, 7–27.

Chen, L.C., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking Atrous Convolution for Semantic Image Segmentation. arXiv preprint arXiv:1706.05587.

Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2018a. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. 40, 834–848.

Chen, Y., Fan, R., Yang, X., Wang, J., Latif, A., 2018b. Extraction of urban water bodies from high-resolution remote-sensing imagery using deep learning. Water 10, 585.

Cheng, G., Wang, Y., Xu, S., Wang, H., Xiang, S., Pan, C., 2017. Automatic road detection and centerline extraction via cascaded end-to-end convolutional neural network. IEEE Trans. Geosci. Remote Sens. 55, 3322–3337.

Coello, C.A., 2000. An updated survey of ga-based multiobjective optimization techniques. ACM Computing Surveys (CSUR) 32, 109–143.

Crop Estimates Consortium, 2017. Field Crop Boundary Data Layer.

De Wit, A., Clevers, J., 2004. Efficiency and accuracy of per-field classification for operational crop mapping. Int. J. Remote Sens. 25, 4091–4112.

Defourny, P., Bontemps, S., Bellemans, N., Cara, C., Dedieu, G., Guzzonato, E., Hagolle, O., Inglada, J., Nicola, L., Rabaute, T., et al., 2019. Near real-time agriculture monitoring at national scale at parcel resolution: performance assessment of the sen2-agri automated system in various cropping systems around the world. Remote Sens. Environ. 221, 551–568.

Diakogiannis, F.I., Waldner, F., Caccetta, P., Wu, C., 2020. Resunet-a: a deep learning framework for semantic segmentation of remotely sensed data. ISPRS J. Photogramm. Remote Sens. 162, 94–114.

Egorov, A.V., Roy, D.P., Zhang, H.K., Li, Z., Yan, L., Huang, H., 2019. Landsat 4, 5 and 7 (1982 to 2017) analysis ready data (ARD) observation coverage over the conterminous United States and implications for terrestrial monitoring. Remote Sens. 11, 447.

Evans, C., Jones, R., Svalbe, I., Berman, M., 2002. Segmenting multispectral Landsat tm images into field units. IEEE Trans. Geosci. Remote Sens. 40, 1054–1064.

Garcia-Pedrero, A., Gonzalo-Martn, C., Lillo-Saavedra, M., 2017. A machine learning approach for agricultural parcel delineation through agglomerative segmentation. Int. J. Remote Sens. 38, 1809–1819.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT press.

Graesser, J., Ramankutty, N., 2017. Detection of cropland field parcels from Landsat imagery. Remote Sens. Environ. 201, 165–180.

Hagolle, O., Huc, M., Pascual, D.V., Dedieu, G., 2010. A multi-temporal method for cloud detection, applied to FORMOSAT-2, VENμS, LANDSAT and SENTINEL-2 images. Remote Sens. Environ. 114 (8), 1747–1755.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Identity mappings in deep residual networks. CoRR abs/1603.05027. arXiv:1603.05027.

He, K., Zhang, X., Ren, S., Sun, J., 2016b. Identity mappings in deep residual networks. In: European Conference on Computer Vision. Springer, pp. 630–645.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969.

Inglada, J., Vincent, A., Arias, M., Tardy, B., Morin, D., Rodes, I., 2017. Operational high resolution land cover map production at the country scale using satellite image time series. Remote Sens. 9 (1), 95.

Ioffe, S., Szegedy, C., 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.

Isikdogan, F., Bovik, A., Passalacqua, P., 2018. Learning a river network extractor using an adaptive loss function. IEEE Geosci. Remote Sens. Lett. 15, 813–817.

Ji, S., Zhang, C., Xu, A., Shi, Y., Duan, Y., 2018. 3d convolutional neural networks for crop classification with multi-temporal remote sensing images. Remote Sens. 10, 75.

Kingma, D.P., Ba, J., 2014. Adam: a method for stochastic optimization. CoRR 1–15. http://arxiv.org/abs/1412.6980 arXiv:1412.6980.

Lesiv, M., Laso Bayas, J.C., See, L., Duerauer, M., Dahlia, D., Durando, N., Hazarika, R., Kumar Sahariah, P., Vakolyuk, M., Blyshchyk, V., et al., 2019. Estimating the global distribution of field size using crowdsourcing. Glob. Chang. Biol. 25, 174–186.

Massey, R., Sankey, T., Yadav, K., Congalton, R., Tilton, J., Thenkabail, P., 2017. Nasa Making Earth System Data Records for Use in Research Environments (Measures) Global Food Security-support Analysis Data (GFSAD) Cropland Extent 2010 North America 30 m v001 [Data Set]. https://doi.org/10.5067/MEaSUREs/GFSAD/GFSAD30NACE.001.

Matthews, B.W, 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta Protein Struct. 405 (2), 442–451.

Matton, N., Canto, G.S., Waldner, F., Valero, S., Morin, D., Inglada, J., Arias, M., Bontemps, S., Koetz, B., Defourny, P., 2015. An automated method for annual cropland mapping along the season for various globally-distributed agrosystems using high spatial and temporal resolution time series. Remote Sens. 7, 13208–13232.

Meyer, F., Beucher, S., 1990. Morphological segmentation. J. Vis. Commun. Image Represent. 1, 21–46.

Milletari, F., Navab, N., Ahmadi, S., 2016. V-net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. CoRR abs/1606.04797. arXiv:1606.04797.

Mueller, M., Segl, K., Kaufmann, H., 2004. Edge-and region-based segmentation technique for the extraction of large, man-made objects in high-resolution satellite imagery. Pattern Recogn. 37, 1619–1628.

Novikov, A.A., Major, D., Lenis, D., Hladuvka, J., Wimmer, M., Bühler, K., 2017. Fully Convolutional Architectures for Multi-class Segmentation in Chest Radiographs. CoRR abs/1701.08816. arXiv:1701.08816.

Oliphant, A.J., Thenkabail, P.S., Teluguntla, P., Xiong, J., Gumma, M.K., Congalton, R.G., Yadav, K., 2019. Mapping cropland extent of Southeast and Northeast Asia using multi-year time-series Landsat 30-m data using a random forest classifier on the Google Earth Engine Cloud. Int. J. Appl. Earth Obs. Geoinf. 81, 110–124.

Pan, S.J., Yang, Q., 2009. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22 (10), 1345–1359.

Penatti, O.A., Nogueira, K., Dos Santos, J.A., 2015. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 44–51.

Persello, C., Bruzzone, L., 2010. A novel protocol for accuracy assessment in classification of very high resolution images. IEEE Trans. Geosci. Remote Sens. 48, 1232–1244.

Persello, C., Tolpekin, V.A., Bergado, J.R., de By, R.A., 2019. Delineation of agricultural fields in smallholder farms from satellite images using fully convolutional networks and combinatorial grouping. Remote Sens. Environ. 231, 111253.

Phalke, A., Ozdogan, M., Thenkabail, S.P.G., Congalton, R., Yadav, K., Massey, R., Teluguntla, P.P.J., Smith, C., 2017. Nasa Making Earth System Data Records for Use in Research Environments (Measures) Global Food Security-support Analysis Data (GFSAD) Cropland Extent 2015 Europe, Central Asia, Russia, Middle East 30 m v001 [Data Set]. https://doi.org/10.5067/MEaSUREs/GFSAD/GFSAD30EUCEARUMECE.001. (doi:10.5067/MEaSUREs/GFSAD/GFSAD30EUCEARUMECE.001).

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention. Springer, pp. 234–241.

Ruder, S., 2017. An Overview of Multi-task Learning in Deep Neural Networks. arXiv preprint arXiv:1706.05098.

Rydberg, A., Borgefors, G., 2001. Integrated method for boundary delineation of agricultural fields in multispectral satellite images. IEEE Trans. Geosci. Remote Sens. 39, 2514–2520.

Salman, N., 2006. Image segmentation based on watershed and edge detection techniques. Int. Arab J. Inf. Technol. 3, 104–110.

Shrivakshan, G., Chandrasekar, C., 2012. A comparison of various edge detection techniques used in image processing. International Journal of Computer Science Issues (IJCSI) 9, 269.

Soille, P.J., Ansoult, M.M., 1990. Automated basin delineation from digital elevation models using mathematical morphology. Signal Process. 20, 171–182.

Turker, M., Kok, E.H., 2013. Field-based sub-boundary extraction from remote sensing imagery using perceptual grouping. ISPRS J. Photogramm. Remote Sens. 79, 106–121.

Vuola, A.O., Akram, S.U., Kannala, J., 2019. Mask-rcnn and U-Net Ensembled for Nuclei Segmentation. arXiv preprint arXiv:1901.10170.

Waldner, F., Hansen, M.C., Potapov, P.V., Löw, F., Newby, T., Ferreira, S., Defourny, P., 2017. National-scale cropland mapping based on spectral-temporal features and outdated land cover information. PloS one 12 (8).

Waldner, F., Duveiller, G., Defourny, P., 2018. Local adjustments of image spatial resolution to optimize large-area mapping in the era of big data. Int. J. Appl. Earth Obs. Geoinf. 73, 374–385.

Watkins, B., van Niekerk, A., 2019. A comparison of object-based image analysis approaches for field boundary delineation using multi-temporal sentinel-2 imagery. Comput. Electron. Agric. 158, 294–302.

Wilcoxon, F., 1945. Individual comparisons by ranking methods. Biom. Bull. 1, 80–83.

Yan, L., Roy, D., 2014. Automated crop field extraction from multi-temporal Web Enabled Landsat Data. Remote Sens. Environ. 144, 42–64.

Zhan, Q., Molenaar, M., Tempfli, K., Shi, W., 2005. Quality assessment for geo-spatial objects derived from remotely sensed data. Int. J. Remote Sens. 26, 2953–2974.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890.

Zhong, Y., Giri, C., Thenkabail, P., Teluguntla, P., Congalton, G.R., Yadav, K., Oliphant, J.A., Xiong, J., Poehnelt, J., Smith, C., 2017. Nasa Making Earth System Data Records for Use in Research Environments (Measures) Global Food Security-support Analysis Data (GFSAD) Cropland Extent 2015 South America 30 m v001 [Data Set]. https://doi.org/10.5067/MEaSUREs/GFSAD/GFSAD30SACE.001. (doi:10.5067/MEaSUREs/GFSAD/GFSAD30SACE.001).