Fast automatic deforestation detectors and their extensions for other spatial objects.

Jesper Muren\(^2\), Vilhelm Niklasson\(^2\), Dmitry Otryakhin\(^1,2\), Maxim Romashin

\(^1\) d.otryakhin.acad@protonmail.ch
\(^2\) Department of Mathematics, Stockholm University, SE-10691 Stockholm, Sweden

December 3, 2021

Abstract

This paper is devoted to the problem of detection of forest and non-forest areas on Earth images. We propose two statistical methods to tackle this problem: one based on multiple hypothesis testing with parametric distribution families, another one- on non-parametric tests. The parametric approach is novel in the literature and relevant to a larger class of problems—detection of natural objects, as well as anomaly detection. We develop mathematical background for each of the two methods, build self-sufficient detection algorithms using them and discuss numerical aspects of their implementation. We also compare our algorithms with those from standard machine learning using satellite data.

1 Introduction

Within the last decade, there have been some attempts to create remote deforestation detectors. Contemporary literature on the topic seems to be geared towards machine learning approaches. Just to name a few examples we refer to [10] that employed deep learning for deforestation detection in the Brazilian Amazon and [15] that presented k-nearest neighbour classification to track deforestation in Peru. Bayesian approach has also seen use, e.g. in [12] on time series data from Fiji. The present work aims to create methods that not only show excellent performance, but are also, unlike machine learning methods, fast, simple and interpretable. To achieve these characteristics we employ sample means and covariances and empirical characteristic functions (ECFs) of colour intensities of image samples. The idea behind it is that forests seen from above, like many other natural objects, show rather simple homogeneous structure. Moreover, in many cases natural objects are distinguishable due to their colour features rather than their spatial patterns. Thus, operating on distributions and their ECFs and completely erasing all the spatial dependencies seems to be reasonable.

The general statistical setting of the problem is as follows. In visible spectrum images of geographical objects are represented as 3 matrices. Each of them consists of intensities of one of the red, green or blue colours. We employ supervised learning where tags (forest / non-forest) are assigned to entries of the training data and learn features of the distributions of the matrices in those entries. During classification, test images are split into much smaller square sub images for the purpose of obtaining better resolution, and these sub-images are tested against reference images that are known to contain forest.
We intend to investigate two different approaches to the problem: non-parametric multiple hypothesis testing and parametric one. In the case of parametric multiple hypothesis testing, we assume parametric distributions for the colour intensities of pixels and use training data to estimate the parameters of said distributions. New data is then tested against this distribution. In the case of the non-parametric setting, we do not assume any known distribution for the colour intensities but instead test if the multivariate distributions of the colour intensities of the training data and new data have equal means and covariance matrices.

We build both the parametric and the non-parametric algorithms in such a way that they work on images in visible spectrum, in particular, RGB composites. Thus, the algorithms do not require any certain pre-processing technique for the data and can be applied to a broad range of sources of images of Earth. There are at least two ways of obtaining large datasets of Earth images: satellite data obtained through open access Earth observation missions, e.g. Landsat [13] and Sentinel-2 [4] and aerial photo imaging. In this work, Sentinel-2 images are used because they provide good spatial and temporal resolution. Sentinel-2 mission provides data from visible, near infrared and short-wave infrared sensors in 13 corresponding spectral bands. Although we use RGB composites obtained via a post-processing procedure of Sentinel-2 original tiles, our methods could work on all of the data in those 13 bands.

1.1 A practical example

In order to give more intuition of what task we tackle, we present a real-life example of deforestation detection. Figure 1a shows a Sentinel-2 image of a piece of land near the city of Arkhangelsk in Russia suffering from logging. Figure 1b is a black-and-white image representing the detection result. White represents forest and black is used for all types of terrains not covered by it.
One can see that the detector highlighted the areas with logging traces (black rectangles) as well as natural objects: lakes, bare ground and pieces of rivers.

The rest of the paper is organized as follows. Section 2 briefly mentions the key steps in data preparation and discusses what features forest data possess. Sections 3 and 4 describe correspondingly novel non-parametric and parametric methods. Section 5 discusses theoretical properties of the two methods, while the last section compares the two with well-known methods from machine learning using Satellite data.

Our main contribution contains the application of the known theory of Hotelling $T^2$ statistic to the problem of deforestation detection, which occupies Section 3. There, we prove Theorem 1 and using it, build classification and training algorithms in 3.2 and 3.3. Another major contribution is Section 4 (excluding 4.1), where we develop a new setting of hypothesis testing for stable models based on empirical characteristic functions in 4.2 and use this setting to build algorithms for detection of forest in 4.3 and 4.4.

## 2 Data and exploratory analysis

To obtain data in visible spectrum we use a simple preprocessing procedure, collecting red, green and blue bands from Sentinel-2 Level-1C products and combining them into one GeoTIFF file. It is then handled by GDAL [5], which creates three numeric matrices (one per each of the red, green and blue channels) with values between 0 and 1. The procedure is described in appendix A in more detail.

Initial inspection of the empirical density of the forest data gave us some additional ideas for potentially suitable distributions. We considered distributions including, but not limited to, stable, Normal, Log-Normal, Gamma, Cauchy, Weibull and Levy. We estimated parameters for each of the colour intensity distributions and calculated the root-mean-square-error (RMSE)

\[
\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{f}(x_i) - f^*(x_i))^2}{n}}
\]

where $\hat{f}(\cdot)$ is the probability density function using the estimated parameters, $f^*(\cdot)$ is the empirical density and $x_i$ is colour intensity of data point $i = 1, \ldots, n$. To choose the best fitting distributions we then plotted the densities of the three distributions with the lowest RMSE’s together with the empirical densities and compared. These distributions appeared to be Normal, Gamma and Stable consistently throughout the set of images. An example of a forest image with such comparison can be seen in Figure 2, where we can observe the three best fitting distributions for a forest image with all parameters estimated. Further from the plots we see that the Normal distributions do not seem to fit very well while the Gamma and Stable distributions are quite close, with Stable edging out Gamma especially for the intensity of the colour red.

Since the focus in this work is to be able to detect if an area has forest or not, we describe detection of presence or lack of this type of natural objects. However, our exploratory data analysis shows us that the framework developed could be used to detect other natural objects from satellite images as well. In Figures 11 and 12 in appendix B, we can see promising results of using the stable distribution to describe the colour intensities of images of mountains and sea. When it comes to man-made objects such as, for example a city, our approach does not work, and we have to at least assume a multimodal mixture of some kind to be appropriate since satellite images contain distinctly different parts, i.e., an orange roof, green grass and grey pavement.
Figure 2: Plots of empirical and estimated probability density functions for distribution of colour intensities in an image of a forest. The means of the red, green and blue intensities are respectively 0.16, 0.24 and 0.20.
3 Detection with a non-parametric method

In this section, the non-parametric method for forest detection is developed. Later, in Section 4, the ideas presented here are used to create the theory and implementation for the parametric one.

3.1 Comparing images with Hotelling T-square statistic

The Mahalanobis distance $D$ was originally introduced as a distance between a given multivariate point $p$ and a distribution. For a point $p$ and a distribution with mean vector $\mu$ and covariance matrix $\Sigma$, the distance is given by

$$D = \sqrt{(p - \mu)^T \Sigma^{-1} (p - \mu)}.$$ 

Where, in practice, the mean and covariance matrix of the distribution would be estimated by the sample mean and covariance matrix. Further extensions can be made to calculate the distance between two multivariate populations with

$$D = \sqrt{(\bar{X}_1 - \bar{X}_2)^T \hat{\Sigma}^{-1} (\bar{X}_1 - \bar{X}_2)}.$$  \hspace{1cm} (1)

Where $\bar{X}_1$ and $\bar{X}_2$ are the sample mean vectors of the two populations. The variable $\hat{\Sigma}^{-1}$ is the inverse of the pooled sample covariance matrix, given by $\hat{\Sigma} = (n_1 \hat{\Sigma}_1 + n_2 \hat{\Sigma}_2)/(n - 2)$ for samples with respective sample sizes $n_1, n_2$, total sample size $n = n_1 + n_2$ and sample covariance matrices $\hat{\Sigma}_1, \hat{\Sigma}_2$. An interested reader can find more on the topic in [3].

The Mahalanobis distance is often used in cases where there is correlation present in the considered multivariate distribution. As the reader might have noticed, the Mahalanobis distance expression is closely related to the statistical concept of standardization, i.e., subtracting the mean and dividing by the standard deviation. The idea is to decorrelate and scale each dimensions variance to be 1 and then calculate the Euclidean distance.

**Theorem 1.** Let $X_1$ and $X_2$ be independent samples of sizes $n_1 \times p$ and $n_2 \times p$ composed from i.i.d $p$-dimensional random vectors from distributions with mean $\mu$ and covariance matrix $\Sigma$. Then, for $D$ from (1) we have that $\frac{n_1 n_2}{n_1 + n_2} D^2 \rightarrow \chi^2(p)$.

**Proof.** By the multidimensional central limit theorem

$$(\bar{X}_1 - \bar{X}_2) \rightarrow N(0, \gamma \Sigma), \ \text{where} \ \gamma = n_1^{-1} + n_2^{-1}. \hspace{1cm} (2)$$

After introducing $A := (\gamma \Sigma)^{1/2}$ and $\hat{A} := (\gamma \hat{\Sigma})^{1/2}$ we can write

$$(\bar{X}_1 - \bar{X}_2) \rightarrow A Z, \ \text{where} \ Z \sim N(0, I_p). \hspace{1cm} (3)$$

Taking into account (3) and that $\hat{A}^{-1} \xrightarrow{P} A^{-1}$, where the later is a matrix of constants, by a property of weak convergence

$$[\hat{A}^{-1}, (\bar{X}_1 - \bar{X}_2)] \rightarrow [A^{-1}, AZ]. \hspace{1cm} (4)$$

Then, we introduce a new quantity

$$Q := [\hat{A}^{-1} (\bar{X}_1 - \bar{X}_2)]^T [\hat{A}^{-1} (\bar{X}_1 - \bar{X}_2)]. \hspace{1cm} (5)$$
The multivariate continuous mapping theorem applied to (4) implies \( Q \overset{d}{\rightarrow} Z^T Z \overset{d}{=} \chi^2(p) \). On the other hand,

\[
Q = (\bar{X}_1 - \bar{X}_2)^T (\hat{A}^{-1})^T \hat{A}^{-1} (\bar{X}_1 - \bar{X}_2),
\]
and using the following chain of transformations

\[
(\hat{A}^{-1})^T \hat{A}^{-1} = (\hat{A}^T)^{-1} \hat{A}^{-1} = (\hat{A}^T \hat{A})^{-1} = (\gamma \Sigma)^{-1}
\]

we obtain

\[
Q = (\bar{X}_1 - \bar{X}_2)^T (\gamma \Sigma)^{-1} (\bar{X}_1 - \bar{X}_2).
\]

Remark 1. Note, that \( X_1 \) and \( X_2 \) do not have to be drawn from the same distribution. If they are drawn from different ones, but possessing the same mean and covariance matrix, the result of the theorem still holds.

3.2 Hypothesis testing based on Mahalanobis distances between images

Suppose we have an RGB image of forest (Figure 3). It consists of three matrices corresponding to each of the red, green and blue channels. The matrices contain values of colour intensities of pixels in the range \([0, 1]\). Then, each matrix is split into column-vectors that are subsequently stack together to form one column-vector of the length \( p^2 \). The three obtained large vectors are bound to form a \((p^2) \times 3\) matrix \( X \). Since the same operation is applied to every matrix, the triplets of pixel intensities are kept unchanged, so it is possible to analyse dependencies between colours. On the contrary, matrix \( X \) does not contain any information about spatial locations of the pixels. From \( X \), a three-dimensional mean vector and a \( 3 \times 3 \) covariance matrix are calculated.

![Image of forest](image)

Figure 3: Data processing for the non-parametric testing
The assumptions of the theorem hold for data of type of $X$: since all entries are bounded between 0 and 1, the means and covariances will always be finite. This naturally leads us to a hypothesis testing framework. Having two images— one picture that is known to depict forest and one test image, we can use the theorem to accept or reject the null hypothesis that the test image is forest. This is done by first converting RGB composites into matrices $X_1$ and $X_2$, obtaining their sizes, means and covariance matrices, and second, computing the Mahalanobis distance between $X_1$ and $X_2$, and checking whether or not $T := \frac{n_1 n_2}{n_1 + n_2} D^2$ is lower than a certain threshold.

![Figure 4: Classification in the non-parametric model](image)

A scheme for multiple hypothesis testing framework is depicted on Figure 4. When it comes to classification of a certain image, rescaled Mahalanobis distances $T_1, T_2, \ldots$ between the image and every forest picture in the training set are computed. Out of all $T$’s, the minimal one is chosen, and is then compared to a threshold obtained during model training. If the minimal $T$ is smaller than the threshold, the test image is tagged forest, if greater, then— non-forest.

Remark 2. This setting is equivalent to multiple hypothesis testing with the thresholds set to be equal. This restriction can be dropped, in which case the precision of predictions will generally increase. But the downside of such a generalization would be that the additional computational complexity during both training and classification would increase linearly with the number of forest images.

Remark 3. This algorithm describes dependences between colours by means of the pooled sample covariance matrix. On the contrary, all the information about spatial features are erased. Also, distributions of colour intensities in each channel are described only by their location parameters.

### 3.3 Training and cross-validation of the non-parametric model

Training of the non-parametric model is schematically depicted on Figure 5. One can observe two sets of images there: reference forest images in the left column (set 1) and images of both types in the upper row (set 2). For every pair i-th image from set 1 and j-th image from set 2, statistic $T_{ij} = \frac{n_1 n_2}{n_1 + n_2} D^2_{ij}$ is computed, where $D_{ij}$ is the Mahalanobis distance between the two pictures.
Then, for every image from set 2, the minimal statistic $\min_i \hat{T}_{ij}$ is found. Lastly, the threshold $T$ is chosen to minimize misclassification error when classifying images from set 2 according to a result of evaluation of Boolean expression $\min_i \hat{T}_{ij} < T$. As previously, True corresponds to forest, while False— to non-forest.

Aside from simple training on a set of images, we adopt the setting of k-fold cross-validation for this model. Initially, we shuffle pictures and group them into k folds. In this procedure, in order to preserve internal features, we don’t split or merge pictures. We follow a similar procedure to what was described above. Once a fold is chosen, set 1 consists of all forest pictures of a training set, while set 2 is a holdout set. Then statistics $\min_i \hat{T}_{ij}$ are computed for each of the $k$ holdout sets and $T$ is chosen to minimize the misclassification error on all the combinations. Cross-validation is summarized in Algorithm 1.

![Figure 5: Training in the non-parametric model](image-url)
Algorithm 1 Non parametric cross validation procedure

1: **Inputs:** Data set of \(N\) images labelled forest or non-forest, number of folds \(k\) & max threshold range \(T_{\text{max}}\).
2: Compute means and covariances for every image.
3: Randomly shuffle images and group them into \(k\) folds.
4: **for** \(i = 1\) to \(k\) **do**
5: \quad **for** each image in fold \(i\) **do**
6: \quad \quad Compute \(\hat{T} = \frac{n_1 \sigma_2}{\sigma_1^2 + \sigma_2^2} D^2\) from Theorem 1, between the image and all images labelled forest in the remaining \(k - 1\) folds.
7: \quad \quad Keep only the smallest \(\hat{T}\) for the image.
8: \quad **end for**
9: **end for**
10: **for** \(t\) in \(0\) to \(T_{\text{max}}\) **do**
11: \quad For all images and all \(\hat{T}\)'s, tag images with \(\hat{T} < t\) as forest and \(\hat{T} > t\) as non forest.
12: \quad Compute the accuracy for the given threshold \(t\), \(\text{Acc}(t)\).
13: **end for**
14: **return** \(\arg\max_{t} \text{Acc}(t)\) and \(n\)'s, means and covariances of forest images.

4 Detection with a parametric method

The non-parametric approach described in Section 3 is simple and universal in terms of underlying data, it doesn’t make any assumptions on the data coming from RGB composites. On the other hand, it possesses a theoretical drawback - it doesn’t take advantage of the distribution of the data, only its mean and covariance matrix. The parametric approach to the detection was designed basing on the opposite ideas - to capture the actual distributions of the data and parametrize them in order to make use of parametric theoretical machinery.

4.1 Stable distributions for signal processing

The second classification method we develop in this study is based on univariate stable distributions as they proved to fit well to the empirical data (Section 2). In this short section we give some context for use of stable distributions in the field of signal processing, describe how they are fit to satellite data in the present paper and how we estimate sampled data.

The stable distribution family is a class generalising the Gaussian one. In signal processing it appears due to two factors. First, stable distribution is a result of the Generalised Central Limit Theorem which says that if there exists any limit of a mean of i.i.d. variables with possibly unbounded variations, then this limit must have a stable distribution. Thus, this class is a natural choice for modelling signals with potentially large deviations. Second, stable distributions might be asymmetric, and it is controlled by one of the parameters. This family of random variables has already seen use in remote sensing areas such as ship detection [16] and video background subtraction [2]. While it is unclear to us what role the Generalised Central Limit Theorem plays in shaping the distributions of forest colour intensities, it is certain that parametric control of their shape provides accurate fitting (Figure 2) when the general stable distribution is assumed.

**Definition 1.** A random variable \(X\) is said to have a stable distribution if there are parameters \(0 < \alpha \leq 2, \sigma \geq 0, -1 \leq \beta \leq 1\) and \(\delta \in \mathbb{R}\) such that its characteristic function has the following
\[ \phi(t) = E[e^{itX}] = \begin{cases} \exp\left\{ -\sigma |t|^\alpha (1 - i\beta (\text{sign} t) \tan \frac{\pi \alpha}{2}) + i\delta t \right\} & \text{if } \alpha \neq 1 \\ \exp\left\{ -\sigma |t|(1 + i\beta \frac{2}{\pi} (\text{sign} t) \ln |t|) + i\delta t \right\} & \text{if } \alpha = 1 \end{cases} \]  

(9)

Where \( \text{sign} t = 0 \) if \( t = 0 \). The parameter \( \alpha \) is known as the index of stability and controls the decay of the tails. The parameters \( \beta \) is not the classical statistical skewness but it indicates how the distribution is skewed, indicating left skewness if \( \beta < 0 \), right skewness if \( \beta > 0 \) and symmetry if \( \beta = 0 \). The parameter \( \sigma \) is a scale parameter controlling variability and \( \delta \) a location parameter which shifts the distribution. The location parameter only corresponds to the mean when \( \alpha > 1 \), otherwise the mean is undefined. In fact, for \( \alpha < 2 \) the \( p \)-th absolute moment \( E[|X|^p] < \infty \) iff \( p < \alpha \).

4.1.1 Parameter estimation

This section will cover estimation of the four parameters of the univariate stable distribution used in this article. There is a number of techniques available for estimation of these parameters, including quantile-based estimators, Maximum-likelihood-based estimators and those based on the method of moments. For this work, the Koutrouvelis regression-type estimator [6] was chosen, as it shows a great balance between precision and computational performance. Estimators which performed well but were very slow were also studied, including ones based on the generalized method of moments. Similarly, estimators which were faster than the Koutrouvelis one were assessed, but these did not perform well enough for our purposes: the McCulloch estimator [8] based on quantiles and the one developed by Press [11]. A more in-depth study on the performance of different parameter estimators may be of interest to individuals interested in improving the algorithm proposed in this thesis.

The Koutrouvelis regression-type estimator is quite fast, but one big downside is that the estimator of the location parameter \( \delta \) is often unreliable. For our purposes this turns out not to be a big issue though because experimentally we have seen that the parameter \( \alpha > 1 \) in all our forest data and when this is the case the location parameter \( \delta \) equals the mean of the distribution [14, Property 1.2.16]. Which implies that the sample mean is a consistent estimator for the location parameter \( \delta \) for our data.

4.2 A test for empirical characteristic functions of stable distributions

In this part, we develop a statistical test that rejects a null hypothesis that a sample comes from a stable distribution defined by (9) with certain parameters.

First, we provide a limit theorem for vectors composed of parts of ECFs of stable random variables. For a sequence \( \{X_k\}_{k=1}^n \) of iid stable random variables the ECF is given by

\[ \phi_n(t) := \frac{1}{n} \sum_{k=1}^n [\cos(tX_k) + i\sin(tX_k)], \]  

(10)

where we denote

\[ u_n(t) := \cos(tX_k) \quad v_n(t) := \sin(tX_k), \]  

(11)
and define new random vectors

\[ Z_n(t) := \frac{1}{n} \sum_{k=1}^{n} \begin{bmatrix} \cos(tX_k) \\ \sin(tX_k) \end{bmatrix} \quad Z_0(t) := \begin{bmatrix} \text{Re}[\phi(t)] \\ \text{Im}[\phi(t)] \end{bmatrix} \quad (12) \]

**Proposition 1.** Fix \( t \in R \), then \( \sqrt{n}[Z_n(t) - Z_0(t)] \overset{d}{\to} N(0, \Sigma_z) \), where

\[ \Sigma_z = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \quad (13) \]

and the elements of the matrix are given by

\[ \sigma_{11} = \frac{1}{4} [\phi(2t) + 2 + \phi(-2t) - (\phi(t))^2 - 2\phi(t)\phi(-t) - (\phi(-t))^2] \]
\[ \sigma_{22} = \frac{1}{4} (4 - 2\phi(t)\phi(-t) + [\phi(t)]^2) - \frac{1}{4} [\phi(2t) + \phi(-2t) - 2] \mid_{t=1} \]
\[ \sigma_{12} = \sigma_{21} = \frac{1}{4i} [\phi(2t) - \phi(-2t) + \phi(-t)] . \]

**Remark 4.** Note that when \( X_j, j = 1, \ldots, n \) are also symmetric and centered around the origin \( \Sigma_z \) reduces to the form

\[ \Sigma_z = \begin{bmatrix} 1/2 [1 + \phi(2t) - 2(\phi(t))] & 0 \\ 0 & 1/2 [1 - \phi(2t)] \end{bmatrix} \quad (15) \]

This case was described in Section 3 of [11].

**Proof of proposition 1.** First, rewrite \( u_j \) and \( v_j \):

\[ u_j = \frac{1}{2}(e^{itX_j} + e^{-itX_j}) \]
\[ v_j = \frac{1}{2i}(e^{itX_j} - e^{-itX_j}) \quad (16) \]

Since both of them are bounded, then \( Z_n(t) \) is a mean of random vectors with components having the second moment. Thus, the multidimensional central limit theorem immediately applies. Therefore, it is only left to find the entries of the matrix \( \Sigma_z \) (13).

\[
\text{Cov}(u_j, v_j) = \text{Cov}\left(\frac{1}{2}(e^{itX_j} + e^{-itX_j}), \frac{1}{2i}(e^{itX_j} - e^{-itX_j})\right) = \frac{1}{4i} \left[ \text{Cov}(e^{itX_j}, e^{itX_j}) - \text{Cov}(e^{itX_j}, e^{-itX_j}) + \text{Cov}(e^{-itX_j}, e^{-itX_j}) - \text{Cov}(e^{-itX_j}, e^{itX_j}) \right] = \]
\[ \frac{1}{4i} \left[ \text{Cov}(e^{itX_j}, e^{itX_j}) - \text{Cov}(e^{-itX_j}, e^{-itX_j}) \right] \]

\[ \text{Var}(e^{itX_j}) = \text{E}\left[ e^{2itX_j} \right] - (\text{E}\left[ e^{itX_j} \right])^2 = \phi(2t) - [\phi(t)]^2 \]
\[ \text{Var}(e^{-itX_j}) = \phi(-2t) - [\phi(-t)]^2 \quad (18) \]

Having substituted covariances in (17) using (18) one has

\[ \text{Cov}(u_j, v_j) = \frac{1}{4i} \left[ \phi(2t) - [\phi(t)]^2 - \phi(-2t) + [\phi(-t)]^2 \right] \quad (19) \]
Other elements of the covariance matrix Var($u_j$) and Var($v_j$) are computed in the same manner:

\[
\begin{align*}
\text{Var}(u_j) &= E[u_j^2] - (E[u_j])^2 = \frac{1}{4} E[e^{2t Z X_j} + 2e^0 + e^{-2t Z X_j}] - \frac{1}{4} [E(e^{t Z X_j} + e^{-t Z X_j})]^2 \\
&= \frac{1}{4} (\phi(2t) + 2 + \phi(-2t) - (\phi(t))^2 - 2\phi(t)\phi(-t) - (\phi(-t))^2) \\
\text{Var}(v_j) &= E[v_j^2] - (E[v_j])^2 = -\frac{1}{4} E[e^{2t Z X_j} - 2e^0 + e^{-2t Z X_j}] - \left(\frac{1}{2t} E[e^{t Z X_j} - e^{-t Z X_j}]\right)^2 \\
&= \frac{1}{4} (\phi(t)^2 - 2\phi(t)\phi(-t) + (\phi(-t))^2) - \frac{1}{4} (\phi(t) - \phi(-t))^2 = \frac{1}{4} (\phi(t))^2 - 2\phi(t)\phi(-t) + (\phi(-t))^2 - \frac{1}{4} (\phi(t) - \phi(-t))^2
\end{align*}
\]

(20)

Proposition 1 provides a limiting distribution for a two-dimensional vector $Z_n - Z_0$, making the result inconvenient for statistical inference. To address that, we define an analogy of Mahalanobis distance— a statistic characterising how far $Z_n$ is from $Z_0$. The following theorem provides a limiting distribution for this statistic and plays the key role in the parametric inference.

**Theorem 2.** Fix $t$. Assume that $\hat{\Sigma}_z \overset{d}{\rightarrow} \Sigma_z$ and that $Z_n^{(1)}(t)$, $Z_n^{(2)}(t)$ come from iid $X_1$ and $X_2$ with distribution determined by $Z_0(t)$. Then

\[
a) \tilde{T}_{1,0} := n \left[ Z_n^{(i)}(t) - Z_0(t) \right]' \hat{\Sigma}_z^{-1} \left[ Z_n^{(i)}(t) - Z_0(t) \right] \overset{d}{\rightarrow} \chi^2(2), \quad i = 1, 2
\]

(22)

\[
b) \tilde{T}_{1,2} := n \left[ Z_n^{(1)}(t) - Z_n^{(2)}(t) \right]' \hat{\Sigma}_z^{-1} \left[ Z_n^{(1)}(t) - Z_n^{(2)}(t) \right] \overset{d}{\rightarrow} \chi^2(2)
\]

Proof. a) According to Proposition 1 and properties of weak convergence

\[
\left( \frac{1}{n} \Sigma_z \right)^{-1/2} (Z_n - Z_0) \overset{d}{\rightarrow} N(0, I).
\]

(23)

Hence, by Continuous mapping theorem,

\[
\left[ \left( \frac{1}{n} \Sigma_z \right)^{-1/2} (Z_n - Z_0) \right]' \left[ \left( \frac{1}{n} \Sigma_z \right)^{-1/2} (Z_n - Z_0) \right] = \left[ n(Z_n - Z_0)' (\Sigma_z^{-1/2})' \Sigma_z^{-1/2} \Sigma_z^{-1/2} (Z_n - Z_0) \right] \overset{d}{\rightarrow} \chi^2(2)
\]

(24)

b) follows from independence of $Z_n^{(1)}$ and $Z_n^{(2)}$. \qed

### 4.3 Hypothesis testing based on ECFs of colour intensities

Now, let us return to the data extraction (Figure 6). Assume we have an image of forest and a test image. Having extracted matrix $X$, one can estimate parameters of the stable distributions of the forest picture channels and compute vectors $Z_0^r(t)$, $Z_0^g(t)$ and $Z_0^b(t)$. Similarly, after obtaining $X$ from the test image, empirical characteristic functions are computed and subsequently $Z_n^r(t)$, $Z_n^g(t)$ and $Z_n^b(t)$ are obtained. Note that we use CFs and ECFs of each channel separately, so dependencies between colours are ignored so as spatial dependencies.
Hypothesis testing in the parametric setting is depicted on Figure 7. A test image is compared with each of the forest images in the training set. When comparing a test image and a forest one, \([Z_n^r(t), Z_0^r(t), Z_0^b(t)]\) and \([Z_n^g(t), Z_n^g(t), Z_n^b(t)]\) are used to test separately that the colour intensities of the test image are those of forest:

\[
T_r := n(Z_n^r - Z_0^r)\hat{\Sigma}_{z;r}^{-1}(Z_n^r - Z_0^r) \xrightarrow{d} \chi^2(2)
\]

\[
T_g := n(Z_n^g - Z_0^g)\hat{\Sigma}_{z;g}^{-1}(Z_n^g - Z_0^g) \xrightarrow{d} \chi^2(2)
\]

\[
T_b := n(Z_n^b - Z_0^b)\hat{\Sigma}_{z;b}^{-1}(Z_n^b - Z_0^b) \xrightarrow{d} \chi^2(2)
\]

(25)

It implies that in order to be tagged as forest, a test image must have all three colour samples passing their corresponding test. Since the limiting distributions in (25) are the same, thresholds for each channel are set to be equal.

For each pair the test image and a forest image, \(T_{tf} := \min(T_r, T_g, T_b)\) is computed and then \(T_{min} := \min_f(T_{tf})\) is obtained and compared with the threshold.
4.4 Training and cross-validation of the parametric model

The training procedure starts with finding $\hat{\Sigma}_z$, $\hat{\Sigma}_r$’s, $Z_0$’s and $Z_n$’s for forest images and $Z_n$’s for non-forest ones in formulas (25). After that, training proceeds as in 3.3: $T$ statistics for pairs of images are computed and the threshold is set to minimise the misclassification rate. Throughout classification, $\hat{\Sigma}_z$, $\hat{\Sigma}_r$’s and $Z_0$’s will be available in advance, so that only $Z_n$’s will be left to compute to perform hypothesis testing based on (25). For that reason we do not build the parametric framework on Theorem 2 b) with $\hat{\Sigma}_z$, $\hat{\Sigma}_r$ obtained from both the test image and a training forest image: estimation of parameters of stable distribution is quite computationally demanding. Although, this approach would provide better accuracy.

Cross-validation for the parametric model is summarised in Algorithm 2.

5 Comparison of the methods

There are features typical for both of our methods. First, if by accident there is a non-forest image in the training forest set, test images of its type start to be tagged as forest because they pass the corresponding test. As a consequence, additional false positive results will occur. Specially designed cross-validation may be exploited to eliminate this problem. Second, the setting is multiple hypothesis testing, so it could naturally be re-set to classify forests images into multiple types such as rain forest, coniferous forest, etc. Another thing is that clouds and sunbeam scattering affect the algorithms because they introduce noise to the distributions of objects. It could be cured by atmospheric correction [7]. Also, both models increase accuracy with growth of the numbers of points in the training data and test data. Improvement of prediction accuracy could be achieved in three ways. Increasing resolution, e.g. by switching from satellite data to aerial imagery, would add more pixels to objects under consideration while adding new tagged images to the training set would provide better training of thresholds and more tagged classes. In practice, merely increasing the size of an image with fixed resolution (adding more area) works well only in the case of training pictures: enlarging test images would imply observing pictures with objects belonging to different
Algorithm 2 Parametric cross validation procedure

1: **Inputs:** Data set of images labelled forest or non-forest, number of folds $k$, max threshold $T_{\text{max}}$, parameter for CF’s ($t$ in (9)).
2: For every forest image compute $Z_0$, parameters and the covariance matrix.
3: For every non-forest image compute $Z_n$.
4: Randomly shuffle images and group them into $k$ folds.
5: for $i = 1$ to $k$ do
6: for each image in fold $i$ do
7: Compute $3\hat{T}$’s from (25), between the image and all images labelled forest in the remaining $k-1$ folds.
8: Keep only the smallest $\hat{T}$ for the image.
9: end for
10: end for
11: for $t$ in $0$ to $T_{\text{max}}$ do
12: For all images and all $\hat{T}$’s, tag images with $\hat{T} < t$ as forest and $\hat{T} > t$ as non forest.
13: Compute the accuracy for the given threshold $t$, $\text{Acc}(t)$.
14: end for
15: return $\arg\max_t \text{Acc}(t)$ and $Z_0$’s, parameters and the covariance matrices of forest images.

There is also a number of differences between the two models. An advantage of the parametric model is that it can work on incomplete data: e.g. when only 1 channel is available or different channels are available at different times. The non-parametric method needs at least two channels simultaneously. On the other hand, the disadvantage of the parametric model is that the algorithm generally requires more pixels per image in training as it involves parameter estimation and because we chose to use theorem 2 a). The estimation procedure also makes the parametric method slower than the other one. A con of the non-parametric method is that the model does not take into account the actual distributions of colour intensities, which can contribute to additional false positives.

To summarise the comparison, we point out that at the current stage of research it is not possible to claim superiority of one method over the other as they both have pros and cons.

6 Empirical application

Sentinel-2 satellite images are used to illustrate how the new Mahalanobis distance classifier (MDC) and stable distribution classifier (SDC) perform in comparison to some other commonly used classification methods. The methods used for comparison are support vector machine (SVM), logistic regression (LR) and deep neural network (DNN). Three experiments with real data are shown here: two comparing accuracy of the methods and one giving visual representation of performance. As the source of terrain images, the region Occitanie in France is used, more precisely, vicinity of the city of Albi.

In each of the experiments SVM is implemented with the radial basis function kernel and the DNN has four hidden layers of sizes 350, 200, 100 and 50 together with ReLu as activation function. The feature space for the comparison methods consists of all pixels and their RGB colour values.
In practise, one would typically use some dimensionality reduction technique, e.g. PCA, before implementing the classification methods, but since no such technique is used for our new methods we want to be consistent with using all of the available data. Considering that the images are rather small this should not cause any issues.

In the first two experiments we look at how the methods perform using a training and test split of the available dataset. In the first one the training set consists of 510 images of sizes 20 × 20 pixels with an equal number forest and non-forest images. The test set is also equally balanced between forest and non-forest images but it consists of 200 images of sizes 20 × 20 pixels. The second experiment is similar, except that each 20 × 20 image is split into four equal pieces, making 2040 images of sizes 10 × 10 pixels in the training set and 1000 10 × 10 images in the test set. It should be noted that all of the images are collected from the same region and under similar conditions. This means that the models are not trained for different seasons or different types of forests. However, it would be straightforward to make the classification models more general by including more diverse data.

The performance measure used to evaluate the methods is the error rate, i.e., the number of incorrectly classified images in the test set divided by the total number of test images. Figure 8 illustrates how the methods perform when the training set varies between 10 % and 100 % of the total training set using images of sizes 10 × 10 pixels and 20 × 20 pixels, respectively. Even though the training set size varies, an equal number of forest and non-forest images are always used for training. The test set is kept unchanged.

It is clear in Figure 8 that the new methods outperform all of the comparison methods. The error rate drops to zero almost immediately in the 10 × 10 pixel case when using SDC whereas the error rate corresponding to MDC gradually decreases from 2.4% and reaches zero when the training set size is about 1800. All of the standard classification methods perform worse and their error rates do not fall below 2%. The best performing methods among the comparison ones are DNN and SVM. They show similar results apart from the fact that SVM is more robust to changes in performance when adding new data to the training set. Analogous results are presented in the 20 × 20 pixel case. Here, MDC generally has a lower error rate, and it drops to zero after the size of the training set reaches 300. On the contrary, the error rate of SDC is higher than in the previous case. Moreover, when the size of the training set is lower than 62, the error rate is higher than those of DNN and
SVM. When the size of the training set reaches 62, the error sharply drops and becomes lower or comparable to those two.

In the third experiment we visualise how the new methods and DNN (one of the best performing comparison methods) behave when every sub-image of size $10 \times 10$ pixels is classified of an image of size $2000 \times 2000$ pixels. Sub-images that are classified as forest are marked as white and non-forest sub-images are marked as black. This makes it possible to get a better feeling for how the methods differ and identify their strengths and weaknesses. The same (complete) training set as in the first part is used to calibrate the methods. This training set does not contain images that are part of the $2000 \times 2000$ pixels image but they were collected from a nearby region under similar conditions. The result of the classification is illustrated in Figure 9.
Generally, all three models performed well, though there are differences in their outcomes. One can see that unlike MDC and SDC, DNN is not so sensitive to the intensity of the green colour. It is best illustrated by performances of the models on forest covered by a shadow of the cloud in the right part of Figure 9. MDC and SDC incorrectly treat the shadowed vegetation as non-forest,
while DNN tags it as forest. Shades from clouds generally could be accounted for by cloud detection, which is a solved problem in Earth image processing (see e.g. [7]). In some other cases DNN tags non-forest vegetation as forest. In particular, in Figure 10 there is forest rounded by non-forest vegetation. Unlike MDC and SDC, DNN fails to detect some non-forest parts of the circle.

Figure 10: Central parts of images in Figure 9.
References

[1] PEPS - Plateforme d’Exploitation des Produits Sentinel. URL https://peps.cnes.fr.

[2] H. Bhaskar, L. Mihaylova, and A. Achim. Video foreground detection based on symmetric alpha-stable mixture models. *IEEE Transactions on Circuits and Systems for Video Technology*, 20(8):1133–1138, 2010.

[3] J. Bibby, K. Mardia, and J. Kent. *Multivariate analysis*. Academic Press, 1979.

[4] M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, et al. Sentinel-2: ESA’s optical high-resolution mission for GMES operational services. *Remote sensing of Environment*, 120:25–36, 2012.

[5] GDAL/OGR contributors. *GDAL/OGR Geospatial Data Abstraction software Library*. Open Source Geospatial Foundation, 2021. URL https://gdal.org.

[6] I. A. Koutrouvelis. Regression-type estimation of the parameters of stable laws. *Journal of the American statistical association*, 75(372):918–928, 1980.

[7] M. Main-Knorn, B. Pfiff, J. Louis, V. Debacker, U. Müller-Wilm, and F. Gascon. Sen2Cor for Sentinel-2. In *Image and Signal Processing for Remote Sensing XXIII*, volume 10427, page 1042704. International Society for Optics and Photonics, 2017.

[8] J. H. McCulloch. Simple consistent estimators of stable distribution parameters. *Communications in Statistics-Simulation and Computation*, 15(4):1109–1136, 1986.

[9] J. P. Nolan. Numerical calculation of stable densities and distribution functions. *Communications in statistics. Stochastic models*, 13(4):759–774, 1997.

[10] M. Ortega Adarme, R. Queiroz Feitosa, P. Nigri Happ, C. Aparecido De Almeida, and A. Rodrigues Gomes. Evaluation of deep learning techniques for deforestation detection in the Brazilian Amazon and cerrado biomes from remote sensing imagery. *Remote Sensing*, 12(6):910, 2020.

[11] S. J. Press. Estimation in univariate and multivariate stable distributions. *Journal of the American Statistical Association*, 67(340):842–846, 1972.

[12] J. Reiche, S. De Bruin, D. Hoekman, J. Verbesselt, and M. Herold. A Bayesian approach to combine Landsat and ALOS PALSAR time series for near real-time deforestation detection. *Remote sensing*, 7(5):4973–4996, 2015.

[13] D. Roy, M. Wulder, T. Loveland, W. C.E., R. Allen, M. Anderson, D. Helder, J. Irons, D. Johnson, R. Kennedy, T. Scambos, C. Schaaf, J. Schott, Y. Sheng, E. Vermote, A. Belward, R. Bindschadler, W. Cohen, F. Gao, J. Hipple, P. Hostert, J. Huntington, C. Justice, A. Kilic, V. Kovalskyy, Z. Lee, L. Lymburner, J. Masek, J. McCorkel, Y. Shuai, R. Trenza, J. Vogelmann, R. Wynne, and Z. Zhu. Landsat-8: Science and product vision for terrestrial global change research. *Remote Sensing of Environment*, 145:154–172, 2014. ISSN 0034-4257. doi: https://doi.org/10.1016/j.rse.2014.02.001. URL https://www.sciencedirect.com/science/article/pii/S003442571400042X.

[14] G. Samorodnitsky and M. S. Taqqu. *Stable Non-Gaussian Random Processes. Stochastic Models with Infinite Variance. Stochastic Modeling*. Routledge, 2017.
[15] J. Shermeyer and B. Haack. Remote sensing change detection methods to track deforestation and growth in threatened rainforests in Madre de Dios, Peru. *Journal of Applied Remote Sensing, 9*(1):096040, 2015.

[16] C. Wang, M. Liao, and X. Li. Ship detection in SAR image based on the alpha-stable distribution. *Sensors, 8*(8):4948–4960, 2008.
A Data pre-processing technology

We have built a preprocessing procedure of creating RGB composites from Sentinel satellite images using Python programming language.

First, we download Sentinel 2 Level-1C products from PEPS[1] with the means of EODAG. The PEPS (Plateforme d’Exploitation des Produits Sentinel) platform, redistributes the products of Sentinel satellites, S1A, S1B, S2A and S2B, S3A and S3B from COPERNICUS, the European system for Earth monitoring. EODAG (Earth Observation Data Access Gateway) is a command line tool and a Python package for searching and downloading remotely sensed images while offering a unified API for data access regardless of the data provider.

Second, for each product we do a preprocessing procedure of creating RGB composites from Sentinel 2 Level-1C products. Using rasterio package1 we extract red, green and blue bands from channels B04, B03 and B02 respectively and combine them in a single file using GeoTIFF file format2.

Third, we compress RGB-composites with lossless compression algorithm to decrease the amount of storage space needed for data using gdal_translate utility [5].

Lastly, in order to obtain images with sufficient brightness, we normalize values in the channels to range from 0.0 to 1.0 by dividing them by the constant 2000. We truncate values with normalised intensity more than 1.0 by setting them to 1.0, which primarily affects bright white clouds and conforms with typical ”crop and scale” image processing procedure. Deforestation detection algorithms use this normalised values.

B Distributions of colour intensities of other natural objects

In this section we explore two RGB images: one- of a mountain and another one- of sea. Each picture is accompanied by empirical density functions plotted for all channels. For comparison, best fitting stable, gamma and normal distributions are shown.

1https://rasterio.readthedocs.io/en/latest/
2https://gdal.org/drivers/raster/gtiff.html#raster-gtiff
(a) Mountain image analysed

(b) Comparison of densities for red color intensity

(c) Comparison of densities for green color intensity

(d) Comparison of densities for blue color intensity

Figure 11: Plots of empirical and estimated probability density functions for distribution of colour intensities in an image of mountain.
(a) Sea image analysed

(b) Comparison of densities for red color intensity

(c) Comparison of densities for green color intensity

(d) Comparison of densities for blue color intensity

Figure 12: Plots of empirical and estimated probability density functions for distribution of colour intensities in an image of sea.