A SEARCH ALGORITHM TO META-OPTIMIZE THE PARAMETERS FOR AN EXTENDED KALMAN FILTER TO IMPROVE CLASSIFICATION ON HYPER-TEMPORAL IMAGES

*†‡B.P. Salmon, †‡W. Kleynhans, ‡F. van den Bergh, §J.C. Olivier, *W.J. Marais, †♯T.L. Grobler, and ‡K.J. Wessels*

†Department of Electrical, Electronic and Computer Engineering, University of Pretoria, South Africa

‡Remote Sensing Research Unit, Meraka Institute, CSIR, Pretoria, South Africa
bsalmon@csir.co.za

♯Defense, Peace, Safety and Security Unit, CSIR, Pretoria, South Africa

§School of Engineering, University of Tasmania, Australia

*Space Science and Engineering Center, University of Wisconsin-Madison, Wisconsin, United States

ABSTRACT

In this paper the Bias Variance Search Algorithm is proposed as an algorithm to optimize a candidate set of initial parameters for an Extended Kalman filter (EKF). The search algorithm operates on a Bias Variance Equilibrium Point criterion to determine how to set the initial parameters. The candidate set is then used by the EKF to estimate state parameters to fit a triply modulated cosine function to time series of the first two spectral bands of the MODerate-resolution Imaging Spectroradiometer (MODIS) land product. The state parameters are then used for land cover classification. The results of the search algorithm was tested on classifying land cover in the Limpopo province, South Africa. An improvement in land cover classification was observed when the method was compared to a robust regression method.

***Index Terms*—** Hellinger distance, Kalman Filter, Time series analysis, Spatial information

1. INTRODUCTION

Reliable land cover and land cover change data are essential to environmental monitoring and regional development planning. The Limpopo province located in South Africa is experiencing rapid growth in informal settlement. Proper knowledge of land cover is critical to effectively manage the resources with such sparsely distributed settlements.

Kleynhans *et al.* [1] proposed the triply modulated cosine model which jointly estimates the mean and seasonal component of a NDVI time series to improve land cover separation. The EKF models the NDVI time series by updating a set of state parameters for each time index. Salmon *et al.* [2] expanded on the idea by modeling the spectral bands separately and introduced a meta-optimization method for the EKF that will be called the Bias Variance Equilibrium Point (BVEP) in this paper.

The objective of this paper is to introduce an unsupervised search algorithm called the Bias Variance Search Algorithm (BVSA), which will appropriately search for initial parameters using the BVEP criterion to improve the performance of the EKF.

The paper is organized as follows. Section 2 discusses the study area and data set. In section 3 the BVEP is discussed and the new BVSA method is presented. Section 4 presents the results by comparing several approaches to the new proposed search algorithm. Section 5 presents the conclusions.

2. STUDY AREA AND DATA DESCRIPTION

The Limpopo province is largely covered by natural vegetation which is used as grazing for cattle and wildlife. Development of settlements across the province potentially has detrimental effects on the environment, but the formation of these settlements is monitored only on an ad hoc basis. A study area of approximately 808 km^2 area was selected for validation, which is predominantly covered by vegetation and settlements. Land cover type (vegetation and settlement) in the validation area was obtained through visual inspection. The EKF that is set using the BVSA, was then applied to the entire Limpopo province, covering an area of 147553 km^2.

MODIS spectral bands 1 and 2 time series data were extracted from the 8-day composite MCD43A4 bidirectional reflectance distribution function (BRDF) adjusted MODIS land surface reflectance product, with a spatial resolution of 500 m for the time period January 2001 to January 2009.

3. METHODOLOGY

An EKF estimates a set of state parameters for an underlying model. These state parameters are used to separate a set of time series into different classes. The triply modulated cosine function is used to model the two spectral bands, and is given as

$$x_{i,k,b} = \mu_{i,k,b} + \alpha_{i,k,b} \cos(2\pi f_{\text{samp}} i + \theta_{i,k,b}) + v_{i,k,b}. \quad (1)$$

The variable $x_{i,k,b}$ denotes the observed value of the b^{th} spectral band's time series, $b \in \{1, 2\}$, of the k^{th} pixel, $k \in [1, K_{\max}]$, at time index i, $i \in [1, \mathcal{I}_{\max}]$. The noise sample of the k^{th} pixel at time i for each spectral band is denoted by $v_{i,k,b}$. The noise is assumed to be additive, with a normal distribution on all the spectral bands. The cosine function model is based on several parameters; the frequency f_{samp} which is computed on the annual vegetation growth cycle and the sampling rate of the MODIS sensor, f_{samp} is set to $\frac{8}{365}$. The non-zero mean of the b^{th} spectral band of the k^{th} pixel at time index i is denoted by $\mu_{i,k,b}$, the amplitude by $\alpha_{i,k,b}$ and the phase by $\theta_{i,k,b}$. The values of $\mu_{i,k,b}$, $\alpha_{i,k,b}$ and $\theta_{i,k,b}$ are dependent on time and must be estimated for each pixel k, $\forall k \in [1, K_{\max}]$, given the spectral band observation vectors $x_{i,k,b}$ for i, $\forall i \in [1, \mathcal{I}_{\max}]$, and b, $b \in \{1, 2\}$. The MODIS spectral bands were assumed to be uncorrelated and were treated independently. The index b is omitted for convenience with no loss of generality in the description of this method. As stated earlier, a state vector is estimated by the EKF at each time increment i, and is expressed as

$$\vec{W}_{i,k} = [W_{i,k,1} \;\; W_{i,k,2} \;\; W_{i,k,3}] = [\mu_{i,k} \;\; \alpha_{i,k} \;\; \theta_{i,k}]. \quad (2)$$

For the present case of land cover classification, it was assumed that the state vector $\vec{W}_{i,k}$ does not change significantly through time; hence, the process model is linear. The measurement model, however, contains the cosine function and, as such, is evaluated via the standard Jacobian formulation, through linear approximation of the non-linear measurement function around the current state vector.

The method of operation for an operator is to set the initial parameters using a training set, which include the initial state vector, process noise covariance matrix and observation noise covariance matrix. Section 3.1 summarizes the BVEP criterion presented in [2] and in section 3.2 a novel unsupervised search algorithm termed as the BVSA is introduced to automate the initialization of the EKF parameters.

3.1. BVEP criterion to setting the initial parameters

The general approach to estimating and initializing state vectors, as well as the observation noise covariance matrix $\mathcal{R}$ and process noise covariance matrix $\mathcal{Q}$ for the EKF, is for an analyst to determine these parameters offline.

The BVEP method uses temporal and spatial information to design a parameter space where desirable system behaviour is expected [2]. Let the uncorrelated observation covariance matrix's diagonals be placed into a vector called the observation candidate vector $\Upsilon_{\mathcal{R}}$, $\Upsilon_{\mathcal{R}} \in \upsilon_{\mathcal{R}}$, which is expressed as

$$\Upsilon_{\mathcal{R}} = 10^{\zeta/10} = E[(x_k - E[x_k])^2]. \quad (3)$$

Let the uncorrelated process covariance matrix's diagonals be placed into a vector called the process candidate vector $\Upsilon_{\mathcal{Q}}$, $\Upsilon_{\mathcal{Q}} \in \upsilon_{\mathcal{Q}}$, which is expressed as

$$\Upsilon_{\mathcal{Q}} = 10^{\vec{\zeta_s}/10} = E[(W_{k,\vec{s}} - E[W_{k,\vec{s}}])^2]. \quad (4)$$

The first desired behaviour is the tracking of the observation vector with minimal residual. The minimal achievable sum of absolute residuals $\sigma_{\mathcal{E}}$ is computed as

$$\sigma_{\mathcal{E}} = \min_{\Upsilon_{\mathcal{R},k} \in \upsilon_{\mathcal{R}}, \Upsilon_{\mathcal{Q},k} \in \upsilon_{\mathcal{Q}}} \left\{ \sum_{k=1}^{K_{\max}} \sum_{i=1}^{\mathcal{I}_{\max}} \left\| \hat{x}_{i,k} - x_{i,k} \right\| \right\}, \quad (5)$$

then

$$[\mathcal{R}_{\sigma_{\mathcal{E}}}, \mathcal{Q}_{\sigma_{\mathcal{E}}}] = \operatorname*{argmin}_{\Upsilon_{\mathcal{R},k} \in \upsilon_{\mathcal{R}}, \Upsilon_{\mathcal{Q},k} \in \upsilon_{\mathcal{Q}}} \left\{ \sum_{k=1}^{K_{\max}} \sum_{i=1}^{\mathcal{I}_{\max}} \left\| \hat{x}_{i,k} - x_{i,k} \right\| \right\}. \quad (6)$$

The set $[\mathcal{R}_{\sigma_{\mathcal{E}}}, \mathcal{Q}_{\sigma_{\mathcal{E}}}]$ represents the parameters required to achieve this value. The second criterion is to have internal stability of the state vector. The minimal achievable absolute deviation σ_s for each state variable is computed the same way as $\sigma_{\mathcal{E}}$ in (5). The set $[\mathcal{R}_{\sigma_s}, \mathcal{Q}_{\sigma_s}]$ represents the parameters required to achieve this condition.

The spatial information is included through the use of the set of $K_{\max}$ time series for a geographical area, which is denoted by $\{x_{i,k}\}$. Let $q_{i,\mathcal{E}}$ denote the probability density function derived at time index i from the residuals given over the set of observations $\{x_{i,k}\}_{k=1}^{k=K_{\max}}$ such that $P[a \leq \mathcal{E} \leq b] = \int_a^b f(e)\mathrm{d}e = \int_a^b f(e, \mathcal{R}, \mathcal{Q})\mathrm{d}e$, i.e.,

$$P[a \leq \mathcal{E} \leq b] = \int_a^b q(e, \mathcal{R}, \mathcal{Q})\mathrm{d}e = \int_a^b q_{i,\mathcal{E}}\mathrm{d}e. \quad (7)$$

A conditioned observation probability density function $q^*_{i,\mathcal{E}}$ is defined as the probability density function $q_{i,\mathcal{E}}$ in equation (7) using the set $[\mathcal{R}_{\sigma_{\mathcal{E}}}, \mathcal{Q}_{\sigma_{\mathcal{E}}}]$ to minimize $\sigma_{\mathcal{E}}$ as

$$P[a \leq \mathcal{E} \leq b] = \int_a^b q(e, \mathcal{R}_{\sigma_{\mathcal{E}}}, \mathcal{Q}_{\sigma_{\mathcal{E}}})\mathrm{d}e = \int_a^b q^*_{i,\mathcal{E}}\mathrm{d}e. \quad (8)$$

The conditioned state variable probability density function $q^*_{i,s}$ and corresponding $q_{i,s}$ is computed in the same way. The performance of the current estimate $\Upsilon_{\mathcal{R}}$ and $\Upsilon_{\mathcal{Q}}$ is defined by a performance metric that evaluates how well $\sigma_{\mathcal{E}}$ and σ_s, $\forall s$ are satisfied. The current estimates are recursively updated and are denoted by $\hat{\Upsilon}^{\iota}_{\mathcal{R}}$ and $\hat{\Upsilon}^{\iota}_{\mathcal{Q}}$, where ι denotes the iteration number. The current estimates $\hat{\Upsilon}^{\iota}_{\mathcal{R}}$ and $\hat{\Upsilon}^{\iota}_{\mathcal{Q}}$ are used to derive the set of probability density functions $\{\hat{q}^{\iota}_{i,\mathcal{E}}\}$, $\forall i$, and $\{\hat{q}^{\iota}_{i,s}\}$, $\forall i$. A f-divergent distance known as the Hellinger distance is used to measure the similarity between the probability density functions $\hat{q}^{\iota}_{i,\mathcal{E}}$ and $q^*_{i,\mathcal{E}}$. The modified Hellinger distance $\mathcal{H}_{i,\mathcal{E}}$, $\mathcal{H}_{i,\mathcal{E}} \in [0, 1]$, is computed as

$$\mathcal{H}_{i,\mathcal{E}} = 1 - \sqrt{1 - \sqrt{\int_{-\infty}^{\infty} \hat{q}^{\iota}_{i,\mathcal{E}} \, q^*_{i,\mathcal{E}}\mathrm{d}e}}, \quad (9)$$

where a value of $\mathcal{H}_{i,\mathcal{E}} \to 1$ means high similarity between $\hat{q}^{\iota}_{i,\mathcal{E}}$ and $q^*_{i,\mathcal{E}}$, while $\mathcal{H}_{i,\mathcal{E}} \to 0$ means low similarity. The hellinger distance $\mathcal{H}_{i,s}$ for each state variable is computed the same way.

Let $\mathcal{I}_T$ denote the number of time steps required to ensure the EKF converges to a stable state. The performance metric Γ defined in [2] is deemed accurate at $\mathcal{I}_T$, which is defined as

$$\Gamma_{\mathcal{I}_T} = \min\Big\{\{\mathcal{H}_{\mathcal{I}_T,s}\}\ \{\mathcal{H}_{\mathcal{I}_T,\mathcal{E}}\}\Big\}. \tag{10}$$

3.2. Bias Variance Search Algorithm

The BVSA method is proposed that will attempt to estimate $\hat{\Upsilon}^{\iota}_{\mathcal{R}}$ and $\hat{\Upsilon}^{\iota}_{\mathcal{Q}}$ to satisfy the performance metric $\Gamma_{\mathcal{I}_T}$. The search algorithm starts by creating ideal operating conditions for each parameter in the EKF; followed by using a hill-climbing algorithm to search for a set of $\hat{\Upsilon}^{\iota}_{\mathcal{R}}$ and $\hat{\Upsilon}^{\iota}_{\mathcal{Q}}$ that will attempt to satisfy the ideal operating conditions for all the parameters within the EKF. The first ideal condition is a system that employs perfect tracking of the observation vectors and is obtained by setting

$$q^*_{i,\mathcal{E}} = \big\{q_{i,\mathcal{E}} : \{\zeta\} \to -\infty; \{\varsigma_s\} \to \infty, \forall\ s\big\}. \tag{11}$$

The second ideal condition is a system that employs a stable state variable and is obtained by setting

$$q^*_{i,s} = \big\{q_{i,s} : \{\zeta\} \to \infty; \{\varsigma_{\{s\}\backslash s}\} \to \infty; \{\varsigma_s\} \to -\infty\big\}. \tag{12}$$

After the probability density functions $q^*_{i,\mathcal{E}}$ and $q^*_{i,s}$ are computed, a hill-climbing search algorithm is applied to find a set of initial parameters that will satisfy these conditions. The BVSA iteratively searches the parameter space and is described below.

Step 1: The search algorithm starts with an unbiased EKF where $\zeta^0 = 0\text{dB}$, and $\varsigma^0_s = 0\text{dB}$, $\forall s$.

Step 2: Compute the state vector $\vec{W}_{\mathcal{I}_T,k}$ using the same $\hat{\Upsilon}^{\iota}_{\mathcal{R}}$ and $\hat{\Upsilon}^{\iota}_{\mathcal{Q}}$ for every time series in the set.

Step 3: Obtain the probability density functions $q^{\iota}_{\mathcal{I}_T,\mathcal{E}}$ and $q^{\iota}_{\mathcal{I}_T,s}$ over the $K_{\max}$ time series.

Step 4: Compute the Hellinger distance $\mathcal{H}_{\mathcal{I}_T,\mathcal{E}}$ and $\mathcal{H}_{\mathcal{I}_T,s}$.

Step 5: Determine the best performing condition $\mathcal{H}_{\text{best}}$ as

$$\mathcal{H}_{\text{best}} = \max\big\{\{\mathcal{H}_{\mathcal{I}_T,\mathcal{E}}\}\ \{\mathcal{H}_{\mathcal{I}_T,s}\}\big\}. \tag{13}$$

Step 6: Determine the worst performing condition $\mathcal{H}_{\text{worst}}$ as

$$\mathcal{H}_{\text{worst}} = \min\big\{\{\mathcal{H}_{\mathcal{I}_T,\mathcal{E}}\}\ \{\mathcal{H}_{\mathcal{I}_T,s}\}\big\}. \tag{14}$$

Step 7: Adjust the new ζ^{ι} according to its relative position to the best and worst performing parameters using a threshold $\rho_{\mathcal{H}}$, $\rho_{\mathcal{H}} \in [0,1]$, $\rho_{\mathcal{H}} \in \mathbb{R}$. The adjustment is made as

$$\zeta^{\iota+1} = \begin{cases} \zeta^{\iota} + \gamma^{\iota} & \text{if } \left(\frac{\mathcal{H}_{\mathcal{I}_T,\mathcal{E}} - \mathcal{H}_{\text{worst}}}{\mathcal{H}_{\text{best}} - \mathcal{H}_{\text{worst}}} > \rho_{\mathcal{H}}\right) \\ \zeta^{\iota} - \gamma^{\iota} & \text{if } \left(\frac{\mathcal{H}_{\mathcal{I}_T,\mathcal{E}} - \mathcal{H}_{\text{worst}}}{\mathcal{H}_{\text{best}} - \mathcal{H}_{\text{worst}}} \le \rho_{\mathcal{H}}\right) \end{cases} \tag{15}$$

The variable γ^{ι} is a decreasing scalar measured in decibels and is a non-negative real number.

Step 8: Adjust the new ς^{ι} according to its relative position to the best and worst performing parameters using a threshold $\rho_{\mathcal{H}}$, $\rho_{\mathcal{H}} \in [0,1]$, $\rho_{\mathcal{H}} \in \mathbb{R}$. The adjustment is made as

$$\varsigma^{\iota+1}_s = \begin{cases} \varsigma^{\iota}_s + \gamma^{\iota} & \text{if } \left(\frac{\mathcal{H}_{\mathcal{I}_T,s} - \mathcal{H}_{\text{worst}}}{\mathcal{H}_{\text{best}} - \mathcal{H}_{\text{worst}}} > \rho_{\mathcal{H}}\right) \\ \varsigma^{\iota}_s - \gamma^{\iota} & \text{if } \left(\frac{\mathcal{H}_{\mathcal{I}_T,s} - \mathcal{H}_{\text{worst}}}{\mathcal{H}_{\text{best}} - \mathcal{H}_{\text{worst}}} \le \rho_{\mathcal{H}}\right) \end{cases} \tag{16}$$

The variable γ^{ι} is a decreasing scalar measured in decibels and is a non-negative real number.

Repeat steps 2–8 until one of the parameters ζ or ς_s stabilizes. After the search algorithm converges, the estimates $\hat{\Upsilon}^{\iota}_{\mathcal{R}}$ and $\hat{\Upsilon}^{\iota}_{\mathcal{Q}}$ are used to initialize the EKF.

4. RESULTS

4.1. Meta-optimization of EKF parameters

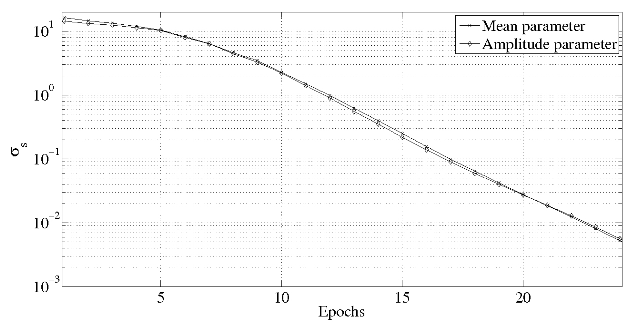

Fig. 1. Average standard deviation for both mean and amplitude parameters computed as a function of epoch.

In this section the standard deviation of the state parameters and the average residuals are reported as a function of epoch produced by the BVSA. In figure 1 the standard deviations of the mean and amplitude parameters are shown and it is observed that with each passing epoch the parameters become more stable (reduced standard deviation).

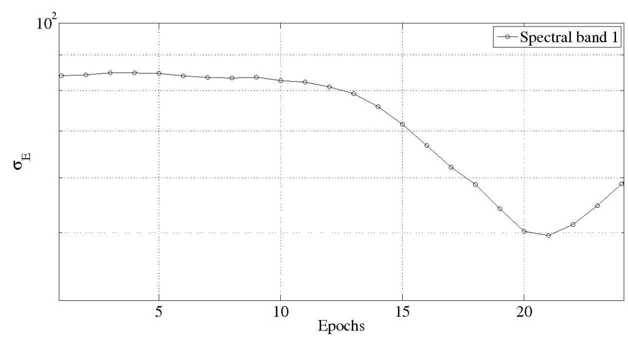

Fig. 2. The average residual computed as a function epoch.

In figure 2 the average measured residual between the estimated observations and actual observations is reported. It was

Table 1. Comparison between an M-estimate and BVEP.

Spectral Band		Mode M-estimate	BVEP
Band 1	$\sigma_{\mathcal{E}}$	118.7	87.1
	σ_{μ}	28.1	0.04
	σ_{α}	36.1	0.02
Band 2	$\sigma_{\mathcal{E}}$	144.7	95.7
	σ_{μ}	37.4	0.01
	σ_{α}	57.6	0.36

observed that with each passing epoch the residual become smaller until an overfit is observed after 21 epochs.

The observation noise covariance matrix and process noise covariance matrix used in the 21st epoch are then used to initialize the EKF. In table 1, a comparison is made between the initialized EKF and an M-estimated model fit. The M-estimate is obtained using the Nelder-Mead algorithm to fit the triply modulated cosine function. The comparison shows a significant reduction in standard deviation of the state parameters and residuals when comparing the initialized EKF to the M-estimate.

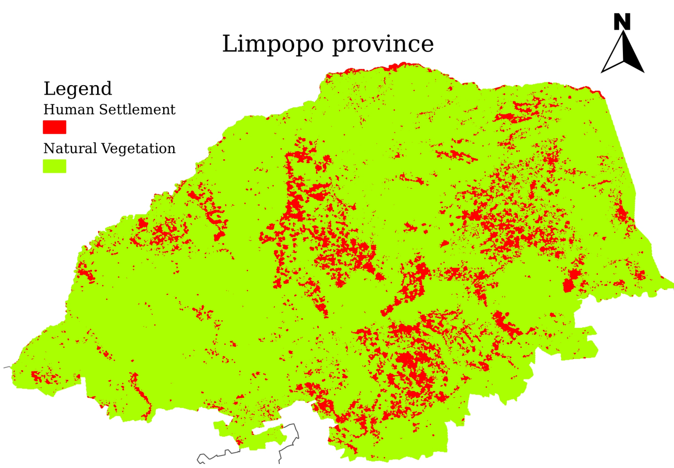

Fig. 3. Classification of the entire Limpopo province, South Africa in January 2009. The settlement are coded in red and vegetation are coded in green.

4.2. Limpopo province – case study

The BVSA uses the BVEP criterion to set the initial parameters of the EKF. The optimized EKF is then applied to time series and produces a stream of state vectors. These state vectors are classified after clustering with the K-means algorithm. A silhouette graph was used to determine the optimal number of clusters for partitional clustering [3]. The clusters were evaluated and grouped into settlement and vegetation.

Table 2. Classification accuracy of K-means on the labelled data set. Each entry gives the average classification accuracy and corresponding standard deviation.

	Method M-estimator	BVEP
Vegetation	$81.5\% \pm 3.6\%$	$84.4\% \pm 0.2\%$
Settlement	$80.6\% \pm 3.0\%$	$82.3\% \pm 0.2\%$

The classification accuracy on the labelled data set is reported in Table 2, as class separability is not solely based on a decrease in the parameters' standard deviation [2]. An improvement of more than 2% is observed in the classification accuracy when the BVEP criterion is used on the labelled data set.

The optimized EKF was then applied to the entire Limpopo province, which covers an area of 147553 km^2 (Figure 3). The clusters allocated to the settlements had a total land coverage of 10.3% (15198 km^2).

5. CONCLUSIONS

This paper demonstrated that improved features can be obtained by using the BVEP criterion [2]. The proposed BVSA is not dependent on acquiring a labelled training data set. It was shown that with proper selection of the initial state parameters, observation noise and process noise covariance matrix, the class separability of features extracted with the EKF is improved.

6. REFERENCES

[1] W. Kleynhans et al., "Improving land cover class separation using an Extended Kalman Filter on MODIS NDVI Time-Series Data," *IEEE Geoscience and Remote Sensing Letters*, vol. 7, no. 2, pp. 381–385, April 2010.

[2] B.P. Salmon et al., "Meta-optimization of the Extended Kalman filter's parameters for improved feature extraction on hyper-temporal images," in *IEEE IGARSS*, 2011, pp. 2543–2546.

[3] L. Kaufman and P.J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley-Interscience, New Jersey, ninth edition, 1990.