CHAPTER FIVE
FEATURE EXTRACTION

5.1 OVERVIEW

In this chapter, four different feature extraction methods that could be used on time series are investigated. The chapter starts with a discussion on how a series of images are used to create a time series of reflectance values for a particular geographical area. From there the feature extraction methods are discussed, which are:

- EKF,
- least squares model fitting,
- M-estimator model fitting, and
- Fourier transform.

The EKF is a regression approach which uses a process model and an internal state space. The least squares and M-estimator methods are regression approaches that aim to minimise the fitting error (residuals) of a predefined model on a time series. The Fourier transform is a frequency analysis approach, which decomposes time series into several harmonic frequencies.

5.2 TIME SERIES REPRESENTATION

A time series is a sequence of data points measured at successive (often uniformly spaced) time intervals. A time series $\mathbf{x}$ of length $\mathcal{I}$, is defined as

$$\mathbf{x} = [\vec{x}_1\ \vec{x}_2\ \ldots\ \vec{x}_{\mathcal{I}}], \tag{5.1}$$

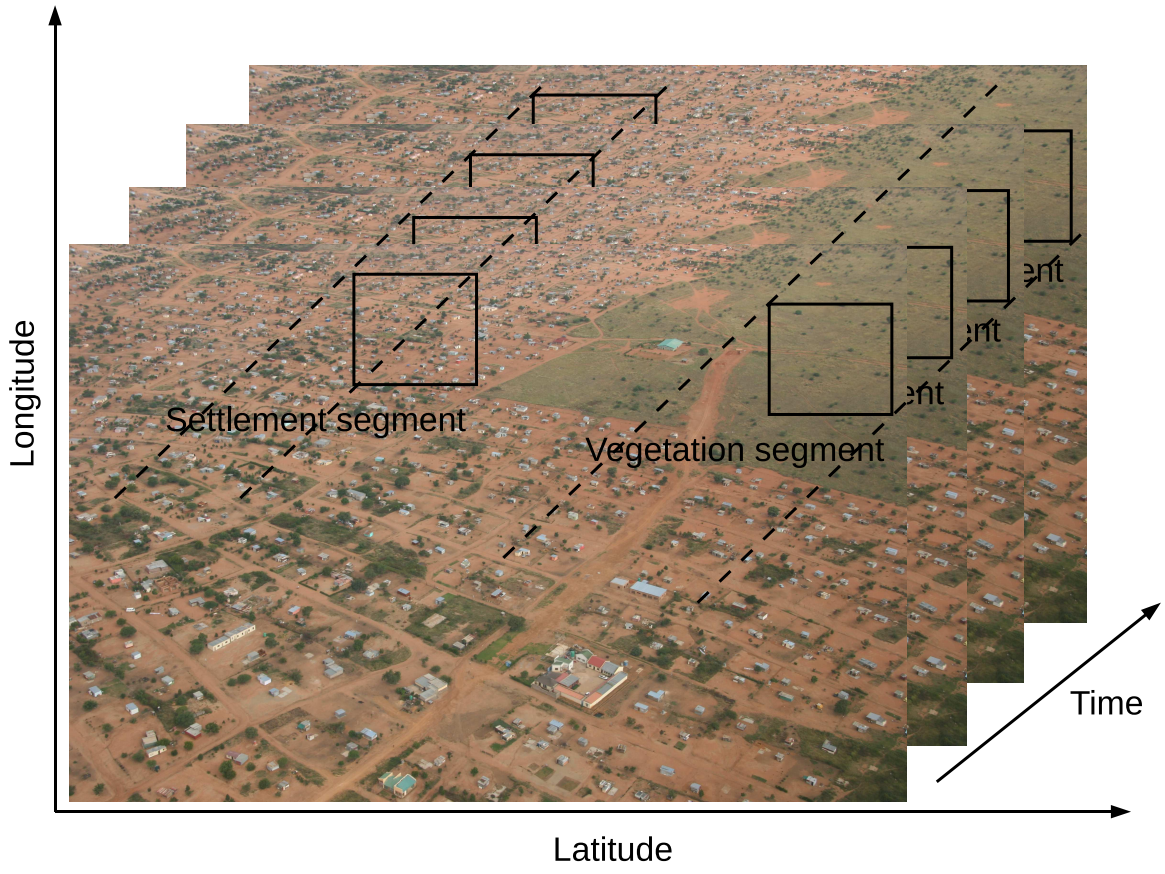

FIGURE 5.1: Multiple aerial photos are acquired in the Limpopo province at different time intervals of the same geographical area. Natural vegetation and human settlement segments are mapped out to form a set of time series.

with

$$\vec{x}_i = [x_{i,1}\ x_{i,2}\ \ldots\ x_{i,T}]. \tag{5.2}$$

The variable T denotes the number of elements in vector $\vec{x}_i$.

The analysis of time series comprises methods that attempt to understand the underlying structure of the data gathered. Analysing the structure allows the identification of patterns and trends, detection of change, clustering, modelling and forecasting [40]. A time series which is extracted from multiple images is used in this chapter to illustrate various concepts.

Land cover example: In figure 5.1, multiple aerial photos are acquired of the same geographical area with segments mapped out over a duration of time. These segments illustrate an example of two different land cover types which do not change over time. The two land cover types are: natural vegetation and human settlement. These hyper-temporal segments are processed to provide a single reflectance value for a given geographical segment at each time interval. A

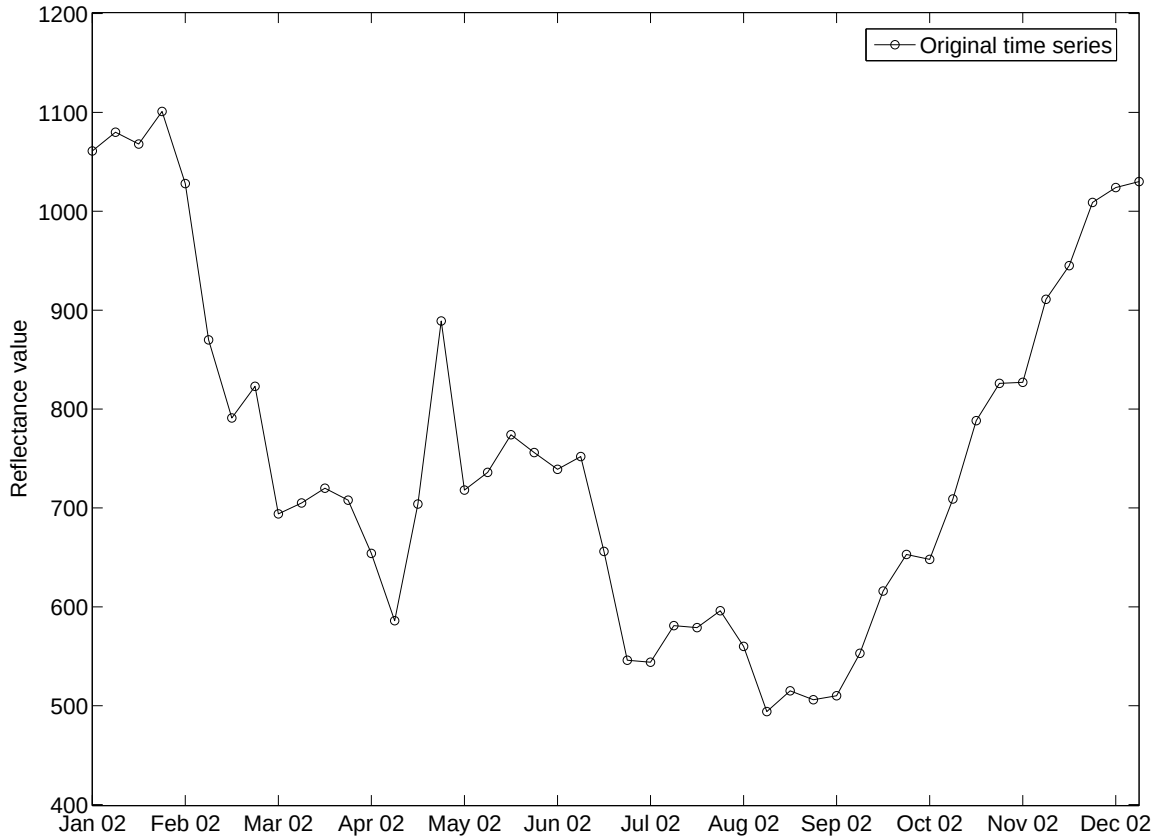

FIGURE 5.2: Time series consisting of reflectance values reported through time for a single image segment shown in figure 5.1.

single reflectance value is obtained from a linear mixture of all the intensities within a segment. The reflectance values for a segment creates a time series shown in figure 5.2. It is observed that the reflectance values in the time series undergo seasonal changes through the course of the year. □

5.3 STATE-SPACE REPRESENTATION

Numerous real world systems are approximated with an underlying process description. This process determines the output of a system which is driven by an internal state. The behaviour at time i of such a system can be predicted based on the information observed from the system at time $(i-1)$. This description of a system's internal operation is known as a state-space model. It was originally developed by control engineers [184, Ch. 3 p. 41]. A state-space model is a mathematical representation frequently used to model a system with a set of state-space variables. The state-space model uses a set of state-space variables to predict the next output of the system.

The state-space variables in most applications are a function of time; as such the use of a time

domain representation is a convenient method for analysing the state-space model of a system [184, Ch. 3 p. 41]. The current state is thus represented by a first order differential function in the time domain. The assumption thus far has been that the process function used within the state-space model and the set of state-space variables are known and that all the system's internal operations have been incorporated. This is usually not the case, as both should be estimated. This results in an erroneous prediction of the output, which leads to assessing the accuracy of the system.

Assessing the accuracy of the state-space model requires the comparison of the actual system's output to the predicted output. The output is usually observed with the addition of noise [185, Ch. 1]. The noise is contributed by several factors, which include:

1. the limited description of the process function,
2. the state-space variables that are not estimated perfectly, and
3. any unknown internal or external source of noise.

This leads to two models required to express the dynamic model: the process model and observation model. The process model is used to describe the adaptation of the state-space variables from time $(i-1)$ to time i. The state-space variables are encapsulated at time i in a state-space vector $\vec{W}_i$ as

$$\vec{W}_i = [W_{i,1} \;\; W_{i,2} \;\; \dots \;\; W_{i,S}], \tag{5.3}$$

where S denotes the number of elements in the state-space vector. The adaptation of the state-space vector is known as the prediction step. The state-space vector $\vec{W}_i$ for time i is predicted using the transition equation, which is given as

$$\vec{W}_i = \mathbf{f}(\vec{W}_{i-1}) + \vec{z}_{i-1}. \tag{5.4}$$

The relation between $\vec{W}_i$ and $\vec{W}_{i-1}$ is described by a known transition function $\mathbf{f}$. A process noise vector $\vec{z}_{i-1}$ is added owing to the incomplete description ability inherent in function $\mathbf{f}$ and/or any previous incorrect estimates of the state-space vector $\vec{W}_{i-1}$. The noise vector $\vec{z}_{i-1}$ is assumed to be a stochastic vector with a zero-mean and covariance matrix $\mathcal{Q}_{i-1}$.

The observation model is used to describe the relation between the state-space vector $\vec{W}_i$ and the actual output of the system at time i. The actual output at time i is termed the observation vector $\vec{x}_i$ and is used in the updating step. The updating step uses a measurement equation which is given as

$$\vec{x}_i = \mathbf{h}(\vec{W}_i) + \vec{v}_i. \tag{5.5}$$

The state-space vector $\vec{W}_i$ is related to the observation vector $\vec{x}_i$ by means of the known measurement function $\mathbf{h}$. The measurement function $\mathbf{h}$ and state-space vector $\vec{W}_i$ might not be perfectly estimated. This is compensated for by including an observation noise vector $\vec{v}_i$, where the noise vector $\vec{v}_i$ is a stochastic vector with zero mean and covariance matrix $\mathcal{R}_i$. Equations (5.4) and (5.5) are known as the state-space form of a linear dynamic model. The time domain approach to state-space model representation provides an iterative model that recursively processes each observation vector sequentially.

It is assumed that both the noise vectors $\vec{z}_{i-1}, \vec{z}_{i-1} \sim \mathcal{N}_u(0, \mathcal{Q}_{i-1})$, and $\vec{v}_i, \vec{v}_i \sim \mathcal{N}_u(0, \mathcal{R}_i)$, are uncorrelated and distributed by a known distribution $\mathcal{N}_u$ for all time increments. This property is expressed as

$$\begin{pmatrix} \vec{z}_{i-1} \\ \vec{v}_i \end{pmatrix} = \mathcal{N}_u \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathcal{Q}_{i-1} & 0 \\ 0 & \mathcal{R}_i \end{pmatrix} \right), \quad \forall i. \tag{5.6}$$

It is also assumed that the noise vectors are uncorrelated with the initial state-space vector $\vec{W}_0$, which is expressed as

$$E[\vec{W}_0 \vec{z}_{i-1}] = E[\vec{W}_0 \vec{v}_i] = 0, \quad \forall i. \tag{5.7}$$

The recursive nature of a linear dynamic model requires that a state-space vector must be adapted at each time increment i using the newest observation vector $\vec{x}_i$. This requires the derivation of a posterior probability density function of the state-space vector, given that all previous observation vectors are available [185, Ch. 1]. This is accomplished by obtaining the initial state-space vector $P(\vec{W}_i)$, after which the posterior probability density function $p(\vec{W}_i|\vec{x}_i, \vec{x}_{i-1}, \dots \vec{x}_0)$ is recursively estimated using the predict (equation (5.4)) and update (equation (5.5)) steps. The posterior probability $p(\vec{W}_i|\vec{x}_{i-1}, \vec{x}_{i-2}, \dots \vec{x}_0)$ is obtained using the Chapman-Kolmogoroff equation given as

$$p(\vec{W}_i|\vec{x}_{i-1}, \vec{x}_{i-2}, \dots \vec{x}_0) = \int p(\vec{W}_i|\vec{W}_{i-1}) p(\vec{W}_{i-1}|\vec{x}_{i-1}, \vec{x}_{i-2}, \dots \vec{x}_0) d\vec{W}_{i-1}. \tag{5.8}$$

The conditional probability density function $p(\vec{W}_i|\vec{W}_{i-1})$ is estimated using the transition equation shown in equation (5.4) and known covariance matrix $\mathcal{Q}_{i-1}$. In this prediction step the transition equation expands the current state-space probability density function. The measurement equation then uses the newest observation vector $\vec{x}_i$ to tighten the state-space probability density function [185, Ch. 1]. The state-space probability density function is updated using the observation vector $\vec{x}_i$ via Bayes' rule as

$$p(\vec{W}_i|\vec{x}_i, \vec{x}_{i-1}, \dots \vec{x}_0) = \frac{p(\vec{x}_i|\vec{W}_i)p(\vec{W}_i|\vec{x}_{i-1}, \vec{x}_{i-2}, \dots \vec{x}_0)}{p(\vec{x}_i|\vec{x}_{i-1}, \vec{x}_{i-2}, \dots \vec{x}_0)}, \tag{5.9}$$

which is expanded to

$$p(\vec{W}_i|\vec{x}_i, \vec{x}_{i-1}, \dots \vec{x}_0) = \frac{p(\vec{x}_i|\vec{W}_i)p(\vec{W}_i|\vec{x}_{i-1}, \vec{x}_{i-2}, \dots \vec{x}_0)}{\int p(\vec{x}_i|\vec{W}_i)p(\vec{W}_i|\vec{x}_{i-1}, \vec{x}_{i-2}, \dots \vec{x}_0)d\vec{W}_i}. \tag{5.10}$$

The conditional probability density function $p(\vec{x}_i|\vec{W}_i)$ is calculated using equation (5.5) and known covariance matrix $\mathcal{R}_i$. The accuracy of the state-space vector can be measured if knowledge of the posterior probability density function $p(\vec{W}_i|\vec{x}_i, \vec{x}_{i-1}, \dots \vec{x}_0)$ is available [185, Ch. 1].

5.4 KALMAN FILTER

The Kalman filter was originally developed by Rudolf Kalman in 1960 and was published in two journals [186, 187]. The Kalman filter was designed to recursively solve the state-space form of the linear dynamic model given in equations (5.4) and (5.5). The Kalman filter assumes that the transition function $\mathbf{f}$ is a known linear matrix $\mathbf{F}$ and the process noise vector $\vec{z}_{i-1}, \vec{z}_{i-1} \sim \mathcal{N}(0, \mathcal{Q}_{i-1})$, is normally distributed. This simplifies the transition equation given in equation (5.4) to

$$\vec{W}_i = \mathbf{F}\vec{W}_{i-1} + \vec{z}_{i-1}. \tag{5.11}$$

The Kalman filter also assumes that the measurement function $\mathbf{h}$ is a known linear matrix $\mathbf{H}$ and the observation noise vector $\vec{v}_i, \vec{v}_i \sim \mathcal{N}(0, \mathcal{R}_i)$, is normally distributed. This simplifies the measurement equation given in equation (5.5) to

$$\vec{x}_i = \mathbf{H}\vec{W}_i + \vec{v}_i. \tag{5.12}$$

The distributions $p(\vec{W}_i|\vec{x}_{i-1}, \dots, \vec{x}_0)$, $p(\vec{W}_{i-1}|\vec{x}_{i-1}, \dots, \vec{x}_0)$ and $p(\vec{W}_i|\vec{x}_i, \dots, \vec{x}_0)$ in equation (5.8) and equation (5.10) are assumed to be normally distributed. The posterior probability $p(\vec{W}_i|\vec{x}_{i-1}, \dots \vec{x}_0)$ is thus expressed as

$$p(\vec{W}_i|\vec{x}_{i-1}, \dots \vec{x}_0) = \sqrt{|2\pi\mathfrak{P}_{(i|i-1)}|}\exp(A_1), \tag{5.13}$$

with

$$A_1 = -\frac{1}{2}(\vec{W}_i - \hat{\vec{W}}_{(i|i-1)})^{\mathrm{T}}\mathfrak{P}^{-1}_{(i|i-1)}(\vec{W}_i - \hat{\vec{W}}_{(i|i-1)}). \tag{5.14}$$

The matrix $\mathfrak{P}_{(i|i-1)}$ denotes the covariance matrix at time i, given all the previous covariance matrices

up to and including time $(i-1)$. The vector $\vec{\hat{W}}_{(i|i-1)}$ denotes the estimate of the state-space vector $\vec{W}$ at time i, given all estimates of state-space vectors up to and including time $(i-1)$. The other posterior probability given in equation (5.8) is expressed as

$$p(\vec{W}_{i-1}|\vec{x}_{i-1},\dots\vec{x}_0) = \sqrt{|2\pi\mathfrak{P}_{(i-1|i-1)}|}\exp(A_2), \tag{5.15}$$

with

$$A_2 = -\frac{1}{2}(\vec{W}_{i-1} - \vec{\hat{W}}_{(i-1|i-1)})^{\mathrm{T}}\mathfrak{P}^{-1}_{(i-1|i-1)}(\vec{W}_{i-1} - \vec{\hat{W}}_{(i-1|i-1)}). \tag{5.16}$$

The matrix $\mathfrak{P}_{(i-1|i-1)}$ denotes the covariance matrix at time $(i-1)$, given all the previous covariance matrices up to and including time $(i-1)$. The vector $\vec{\hat{W}}_{(i-1|i-1)}$ denotes the estimate of the state-space vector $\vec{W}$ time $(i-1)$, given all the previous estimates of state-space vectors up to and including time $(i-1)$. The posterior probability given in equation (5.10) is expressed as

$$p(\vec{W}_i|\vec{x}_i,\dots\vec{x}_0) = \sqrt{|2\pi\mathfrak{P}_{(i|i)}|}\exp\left(-\frac{1}{2}(\vec{W}_i - \vec{\hat{W}}_{(i|i)})^{\mathrm{T}}\mathfrak{P}^{-1}_{(i|i)}(\vec{W}_i - \vec{\hat{W}}_{(i|i)})\right), \tag{5.17}$$

where $\mathfrak{P}_{(i|i)}$ denotes the covariance matrix at time i, given all the previous covariance matrices up to and including time i. The vector $\vec{\hat{W}}_{(i|i)}$ denotes the estimate of the state-space vector $\vec{W}$ at time i, given all estimates of state-space vectors up to and including time i.

The Kalman filter recursively estimates the probability density functions given in equations (5.13)–(5.17). The prediction parameters used in the prediction step (equation (5.4)) include the predicted state-space vector $\vec{\hat{W}}_{(i|i-1)}$ and predicted covariance matrix $\mathfrak{P}_{(i|i-1)}$. The predicted state-space vector's estimate $\vec{\hat{W}}_{(i|i-1)}$ is computed as

$$\vec{\hat{W}}_{(i|i-1)} = \mathbf{F}\vec{\hat{W}}_{(i-1|i-1)}, \tag{5.18}$$

and the predicted estimate of the covariance matrix is computed with

$$\mathfrak{P}_{(i|i-1)} = \mathcal{Q}_{i-1} + \mathbf{F}\mathfrak{P}_{(i-1|i-1)}\mathbf{F}^{\mathrm{T}}. \tag{5.19}$$

The parameters used in the updating step (equation (5.5)) include the posterior estimate of the state-space vector $\vec{\hat{W}}_{(i|i)}$ and posterior estimate of the covariance matrix $\mathfrak{P}_{(i|i)}$. These parameters require the computation of the innovation term and optimal Kalman gain. The innovation term $\mathcal{S}_i$ is computed as

$$\mathcal{S}_i = \mathbf{H}\mathfrak{P}_{(i|i-1)}\mathbf{H}^{\mathrm{T}} + \mathcal{R}_i. \tag{5.20}$$

The optimal Kalman gain $\mathfrak{K}_i$ is computed as

$$\mathfrak{K}_i = \mathfrak{P}_{(i|i-1)}\mathbf{H}^{\mathrm{T}}\mathcal{S}_i^{-1}. \tag{5.21}$$

The posterior estimate of the state-space vector $\hat{\vec{W}}_{(i|i)}$ is computed as

$$\hat{\vec{W}}_{(i|i)} = \hat{\vec{W}}_{(i|i-1)} + \mathfrak{K}_i(\vec{x}_i - \mathbf{H}\hat{\vec{W}}_{(i|i-1)}), \tag{5.22}$$

and the posterior estimate of the covariance matrix $\mathfrak{P}_{(i|i)}$ is computed as

$$\mathfrak{P}_{(i|i)} = \mathfrak{P}_{(i|i-1)} - \mathfrak{K}_i\mathcal{S}_i\mathfrak{K}_i^{\mathrm{T}}. \tag{5.23}$$

If the process function is precise and the initial estimates of $\hat{\vec{W}}_{(0|0)}$ and $\mathfrak{P}_{(0|0)}$ are accurate, then the following five properties will hold. The first two properties, which are relevant to the state-space vector's estimate, are

$$E[\vec{W}_i - \hat{\vec{W}}_{(i|i)}] = E[\vec{W}_i - \hat{\vec{W}}_{(i|i-1)}] = 0, \tag{5.24}$$

$$E[\vec{x}_i - \mathbf{H}\hat{\vec{W}}_{(i|i-1)}] = 0. \tag{5.25}$$

The last three properties hold a relation to the covariance matrices, which accurately reflect the estimated covariance as

$$\mathfrak{P}_{(i|i)} = \mathrm{cov}(\vec{W}_i - \hat{\vec{W}}_{(i|i)}), \tag{5.26}$$

$$\mathfrak{P}_{(i|i-1)} = \mathrm{cov}(\vec{W}_i - \hat{\vec{W}}_{(i|i-1)}), \tag{5.27}$$

$$\mathcal{S}_i = \mathrm{cov}(\vec{x}_i - \mathbf{H}\hat{\vec{W}}_{(i|i-1)}). \tag{5.28}$$

The performance of the Kalman filter is usually inhibited by the poor estimation of the observation noise's covariance matrix $\mathcal{R}_i$ and the process noise's covariance matrix $\mathcal{Q}_{i-1}$. The Kalman filter is unable to compute the mean and covariance of the Gaussian posterior probability $p(\vec{W}_i|\vec{x}_i, \vec{x}_{i-1}, \ldots \vec{x}_0)$ accurately if poor initial estimates are made of the observation and process noise's covariance matrices.

5.5 EXTENDED KALMAN FILTER

The EKF is the non-linear extension of the standard Kalman filter in estimation theory. The EKF has been considered to be the de facto standard in the theory of non-linear state estimate, navigation systems and global positioning system (GPS) [188].

The EKF is similar to the standard Kalman filter as a state-space vector $\vec{W}_i$ is estimated at each time increment i. The state-space vector $\vec{W}_i$ is estimated at time i recursively by using the set of observation vectors $\{\vec{x}_i, \vec{x}_{i-1}, \ldots, \vec{x}_0\}$. The state-space model's equations are reformulated for the EKF in this section. The transition equation in equation (5.11) is rewritten as

$$\vec{W}_i = \mathbf{f}(\vec{W}_{i-1}) + \vec{z}_{i-1}. \tag{5.29}$$

The transition function $\mathbf{f}$ is a non-linear function, and the process noise vector $\vec{z}_{i-1}, \vec{z}_{i-1} \sim \mathcal{N}(0, \mathcal{Q}_{i-1})$, is assumed to be normally distributed. The measurement equation in equation (5.12) is rewritten as

$$\vec{x}_i = \mathbf{h}(\vec{W}_i) + \vec{v}_i. \tag{5.30}$$

The measurement function $\mathbf{h}$ is a non-linear function and the observation noise vector $\vec{v}_i, \vec{v}_i \sim \mathcal{N}(0, \mathcal{R}_i)$ is assumed to be normally distributed. The idea behind the EKF is that the non-linear transition function $\mathbf{f}$ and non-linear measurement function $\mathbf{h}$ can be sufficiently described using local linearisation of the two functions.

The posterior probability density function $p(\vec{W}_i|\vec{x}_i, \ldots, \vec{x}_0)$ is approximated by means of a Gaussian distribution, which implies that equations (5.13)–(5.17) described in the Kalman filter section (section 5.4) still hold. Prediction parameters and updating parameters are reformulated to take into account the non-linear transition and measurement functions. The predicted state-space vector's estimate $\hat{\vec{W}}_{(i|i-1)}$ is expressed as

$$\hat{\vec{W}}_{(i|i-1)} = \mathbf{f}(\hat{\vec{W}}_{(i-1|i-1)}), \tag{5.31}$$

where $\mathbf{f}$ denotes the non-linear transition function. The predicted estimate of the covariance matrix $\mathfrak{P}_{(i|i-1)}$ is expressed as

$$\mathfrak{P}_{(i|i-1)} = \mathcal{Q}_{i-1} + \mathbf{F}_{\text{est}} \mathfrak{P}_{(i-1|i-1)} \mathbf{F}_{\text{est}}^{\text{T}}. \tag{5.32}$$

The matrix $\mathbf{F}_{\text{est}}$ is the local linearisation of the non-linear transition function $\mathbf{f}$. The matrix $\mathbf{F}_{\text{est}}$ is defined as the Jacobian evaluated at $\hat{\vec{W}}_{(i-1|i-1)}$ as [185, Ch. 2]

$$\mathbf{F}_{\text{est}} = \left\| \left[\frac{\partial}{\partial W_{i,1}} \; \cdots \; \frac{\partial}{\partial W_{i,S}} \right] \mathbf{f}^{\text{T}}(\vec{W}_i) \right\|_{\vec{W}_i = \vec{\hat{W}}_{(i-1|i-1)}}. \tag{5.33}$$

In the case of the updating parameters, the posterior estimate of the state-space vector $\vec{\hat{W}}_{(i|i)}$ is expressed as

$$\vec{\hat{W}}_{(i|i)} = \vec{\hat{W}}_{(i|i-1)} + \mathfrak{K}_i(\vec{x}_i - \mathbf{h}(\vec{\hat{W}}_{(i|i-1)})). \tag{5.34}$$

The function $\mathbf{h}$ denotes the non-linear measurement function and $\mathfrak{K}_i$ denotes the EKF's optimal Kalman gain given as

$$\mathfrak{K}_i = \mathfrak{P}_{(i|i-1)}\mathbf{H}_{\text{est}}^{\text{T}}\mathcal{S}_i^{-1}. \tag{5.35}$$

The matrix $\mathbf{H}_{\text{est}}$ is the local linearisation of the non-linear measurement function $\mathbf{h}$. The matrix $\mathbf{H}_{\text{est}}$ is defined as the Jacobian evaluated at $\vec{\hat{W}}_{(i|i-1)}$ as [185, Ch. 2]

$$\mathbf{H}_{\text{est}} = \left\| \left[\frac{\partial}{\partial W_{i,1}} \; \cdots \; \frac{\partial}{\partial W_{i,S}} \right] \mathbf{h}^{\text{T}}(\vec{W}_i) \right\|_{\vec{W}_i = \vec{\hat{W}}_{(i|i-1)}}. \tag{5.36}$$

The innovation term for the EKF is defined as

$$\mathcal{S}_i = \mathbf{H}_{\text{est}}\mathfrak{P}_{(i|i-1)}\mathbf{H}_{\text{est}}^{\text{T}} + \mathcal{R}_i. \tag{5.37}$$

The posterior estimate of the covariance matrix $\mathfrak{P}_{(i|i)}$ is expressed as

$$\mathfrak{P}_{(i|i)} = \mathfrak{P}_{(i|i-1)} - \mathfrak{K}_i\mathcal{S}_i\mathfrak{K}_i^{\text{T}}. \tag{5.38}$$

Land cover example: The time series example given in figure 5.1 produces a time series which is shown in figure 5.2. Kleynhans *et al.* proposed a triply modulated cosine function for the process function [30]. The triply modulated cosine function is expressed as

$$\vec{x}_i = \mu_i + \alpha_i \cos(2\pi f_{\text{samp}} i + \theta_i). \tag{5.39}$$

The variable i denotes the time index and f_{samp} denotes the temporal sampling rate of the image acquisitions. The cosine function is characterised by three variables: μ_i, α_i and θ_i. These three variables form the state-space vector, which is defined as

$$\vec{W}_i = [W_{i,1}\; W_{i,2}\; W_{i,3}] = [W_{i,\mu}\; W_{i,\alpha}\; W_{i,\theta}]. \tag{5.40}$$

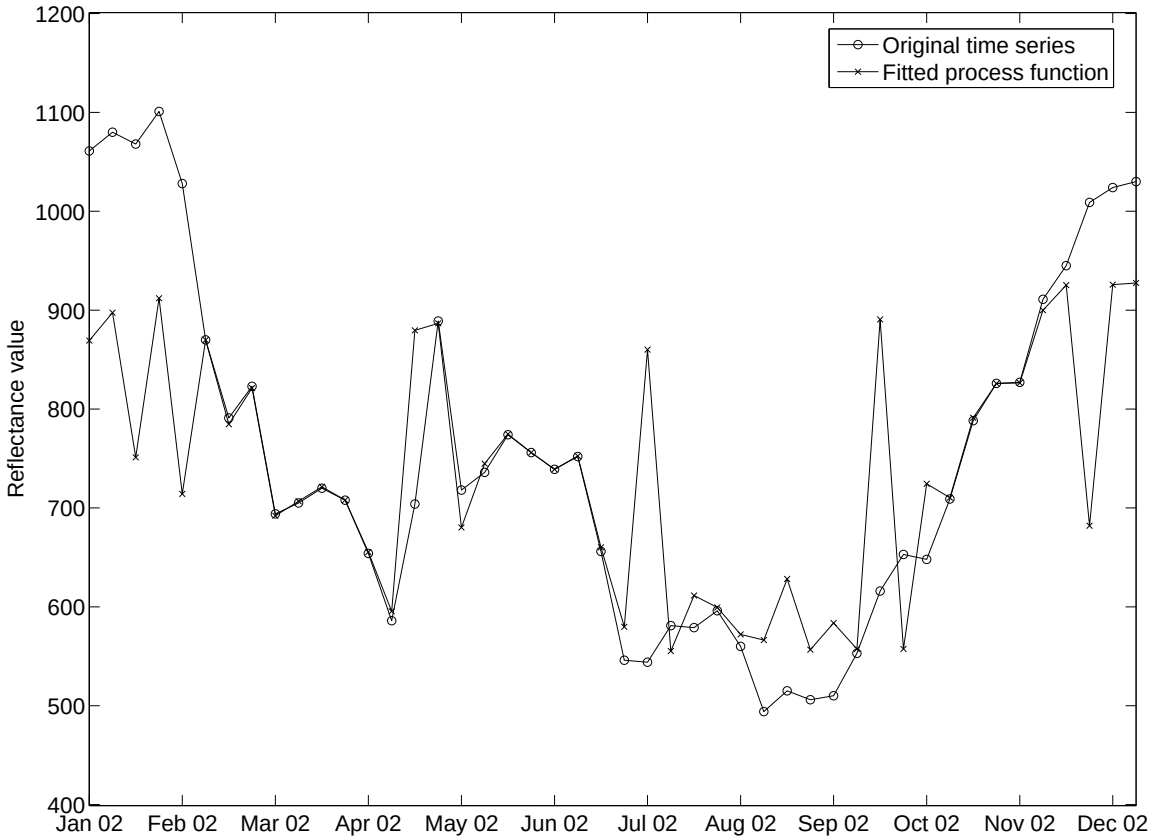

FIGURE 5.3: The Extended Kalman filter estimates the parameters of the state-space vector $\vec{W}_i$ to fit the triply modulated cosine function onto the time series shown in figure 5.2. The estimated state-space vector is used to create a fitted process function to measure the accuracy of the fit.

The triply modulated cosine function is a non-linear function and the EKF was proposed to solve the state-space model. It is assumed that the state-space vector remains constant from one time increment to the next. This reduces the transition equation given in equation (5.29) to

$$\vec{W}_i = \vec{W}_{i-1} + \vec{z}_{i-1}. \tag{5.41}$$

The measurement equation shown in equation (5.30) is defined for this example as

$$\vec{x}_i = \mathbf{h}(\vec{W}_i) + \vec{v}_i, \tag{5.42}$$

where the measurement function $\mathbf{h}$ is the triply modulated cosine function given in equation (5.39) as

$$\mathbf{h}(\vec{W}_i) = W_{i,\mu} + W_{i,\alpha}\cos(2\pi f_{\text{samp}} i + W_{i,\theta}). \tag{5.43}$$

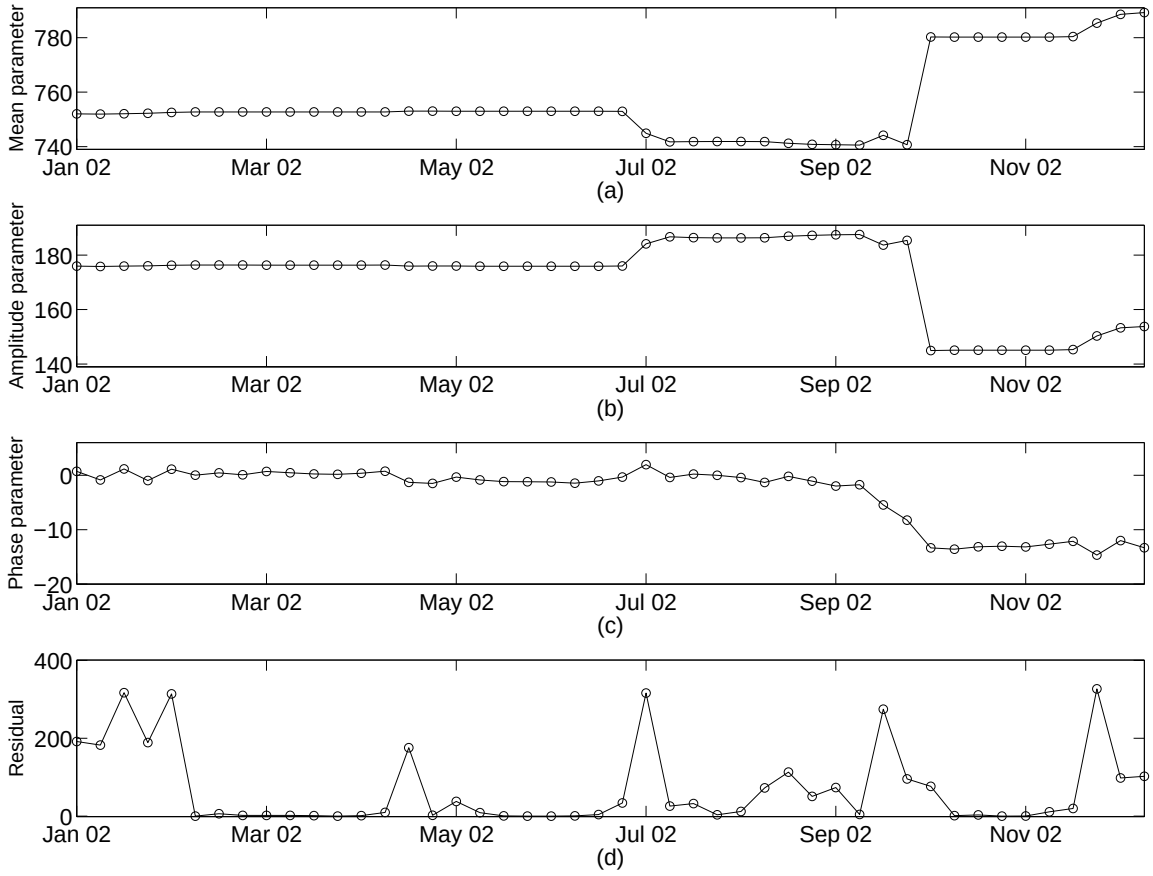

FIGURE 5.4: The Extended Kalman filter estimates the parameters in the state-space vector $\vec{W}_i$. Figure (a) shows the mean parameter μ_i estimates. Figure (b) shows the amplitude parameters α_i estimates. Figure (c) shows the phase parameter θ_i estimates. Figure (d) shows the absolute error in tracking the output of the system.

It should be noted that the measurement function produces a vector with a single dimension. Thus for this example, equation (5.42) is further reduced to a single output as

$$x_i = \mathbf{h}(\vec{W}_i) + v_i. \tag{5.44}$$

The predicted state-space vector's estimate $\hat{\vec{W}}_{(i|i-1)}$ shown in equation (5.31) is rewritten by substituting the transition function with the identity matrix for the example as

$$\hat{\vec{W}}_{(i|i-1)} = \mathbf{f}(\hat{\vec{W}}_{(i-1|i-1)}) = \hat{\vec{W}}_{(i-1|i-1)}. \tag{5.45}$$

The matrix $\mathbf{F}_{\text{est}}$ is an identity matrix, which simplifies the predicted estimate for the covariance matrix $\mathfrak{P}_{(i|i-1)}$ shown in equation (5.32) to

$$\mathfrak{P}_{(i|i-1)} = \mathcal{Q}_{i-1} + \mathbf{F}_{\text{est}}\mathfrak{P}_{(i-1|i-1)}\mathbf{F}_{\text{est}}^{\text{T}} = \mathcal{Q}_{i-1} + \mathfrak{P}_{(i-1|i-1)}. \tag{5.46}$$

The posterior estimate of the state-space vector $\hat{\vec{W}}_{(i|i)}$ shown in equation (5.34) is expressed for this example as

$$\begin{aligned}
\hat{\vec{W}}_{(i|i)} &= \hat{\vec{W}}_{(i|i-1)} + \mathfrak{K}_i(\vec{x}_i - \mathbf{h}(\hat{\vec{W}}_{(i|i-1)})) \\
&= \hat{\vec{W}}_{(i|i-1)} + \mathfrak{K}_i(\vec{x}_i - \mathbf{H}_{\text{est}}(\hat{\vec{W}}_{(i|i-1)})) \\
&= \hat{\vec{W}}_{(i|i-1)} + \mathfrak{K}_i\left(\vec{x}_i - \left\|\left[\frac{\partial \mathbf{h}^{\text{T}}(\vec{W}_i)}{\partial W_{i,\mu}} \; \frac{\partial \mathbf{h}^{\text{T}}(\vec{W}_i)}{\partial W_{i,\alpha}} \; \frac{\partial \mathbf{h}^{\text{T}}(\vec{W}_i)}{\partial W_{i,\theta}}\right]\right\|_{\vec{W}_i = \hat{\vec{W}}_{(i|i-1)}}\right),
\end{aligned} \tag{5.47}$$

with

$$\frac{\partial \mathbf{h}(\vec{W}_i)}{\partial W_{i,\mu}} = 1 \tag{5.48}$$

$$\frac{\partial \mathbf{h}(\vec{W}_i)}{\partial W_{i,\alpha}} = \cos(2\pi f_{\text{samp}} i + \hat{\vec{W}}_{(i|i-1),\theta}) \tag{5.49}$$

$$\begin{aligned}
\frac{\partial \mathbf{h}(\vec{W}_i)}{\partial W_{i,\theta}} &= -\hat{\vec{W}}_{(i|i-1),\alpha}\Big[\sin(2\pi f_{\text{samp}} i)\cos(\hat{\vec{W}}_{(i|i-1),\theta}) + \\
&\qquad \cos(2\pi f_{\text{samp}} i)\sin(\hat{\vec{W}}_{(i|i-1),\theta})\Big].
\end{aligned} \tag{5.50}$$

The time series shown in figure 5.2 is fitted with the triply modulated cosine function by estimating a state-space vector $\vec{W}_i$ for each time increment. The estimated output of the EKF, using the newest available observation vector at time i, is plotted with the actual observation vector $\vec{x}_i$ in figure 5.3. It is observed that the EKF requires an initial number of observations before the state-space vector starts to stabilise. The stabilised state-space vector corresponds to a more accurate tracking of the actual observations.

The progressive estimation of the state-space vectors is shown in figure 5.4. Figure 5.4(a) illustrates the estimation of the mean parameter μ_i (the first element in the state-space vector denoted by $W_{i,\mu}$). Figure 5.4(b) illustrates the estimation of the amplitude parameter α_i (the second element in the state-space vector denoted by $W_{i,\alpha}$). Figure 5.4(c) illustrates the estimation of the phase parameter θ_i (the third element in the state-space vector denoted by $W_{i,\theta}$). The absolute error in the tracking of the output is illustrated in figure 5.4(d).

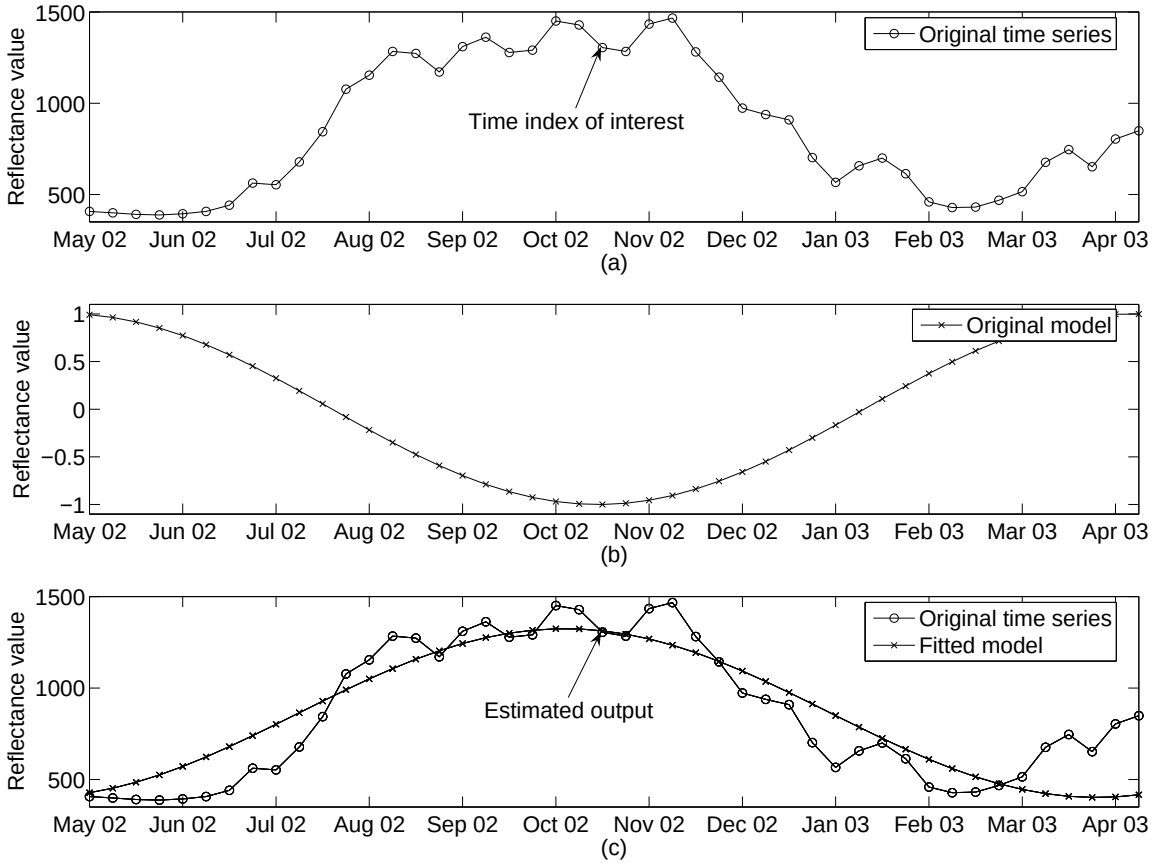

FIGURE 5.5: Least squares estimates the parameter vector $\vec{W}_i$ to fit the model onto the time series.

5.6 LEAST SQUARES MODEL FITTING

The least squares method was first discovered by Carl Friedrich Gauss in 1795 and was later published by the French mathematician Legendre in 1805. The least squares is a method used to fit the triply modulated cosine model with a parameter vector $\vec{W}_i$. It estimates the parameter vector by evaluating the fit of the model to the actual observation vector. The parameter vector in this context can be viewed as the state-space vector defined in the state-space model and the model can be viewed as the process function (section 5.3).

The least squares is a linear regression method, which uses a model $\mathbf{h}$ to predict a set of dependent parameter vectors $\{\vec{W}_i\}$ from a set of independent observation vectors $\{\vec{x}_i\}$. The least squares' goal is to find a parameter vector $\vec{W}_i$ that will minimise the sum of squares between the observation vectors $\vec{x}_i$ and the model's estimated output vector $\hat{\vec{x}}_i$. The sum of squares is computed as a summation of the error residuals to measure the performance and is expressed as

$$\mathcal{E}_{\text{LS}} = \sum_{i=1}^{\mathcal{I}} (\vec{x}_i - \hat{\vec{x}}_i)^2 = \sum_{i=1}^{\mathcal{I}} (\vec{x}_i - \mathbf{h}(\vec{x}_i, \vec{W}_i))^2. \tag{5.51}$$

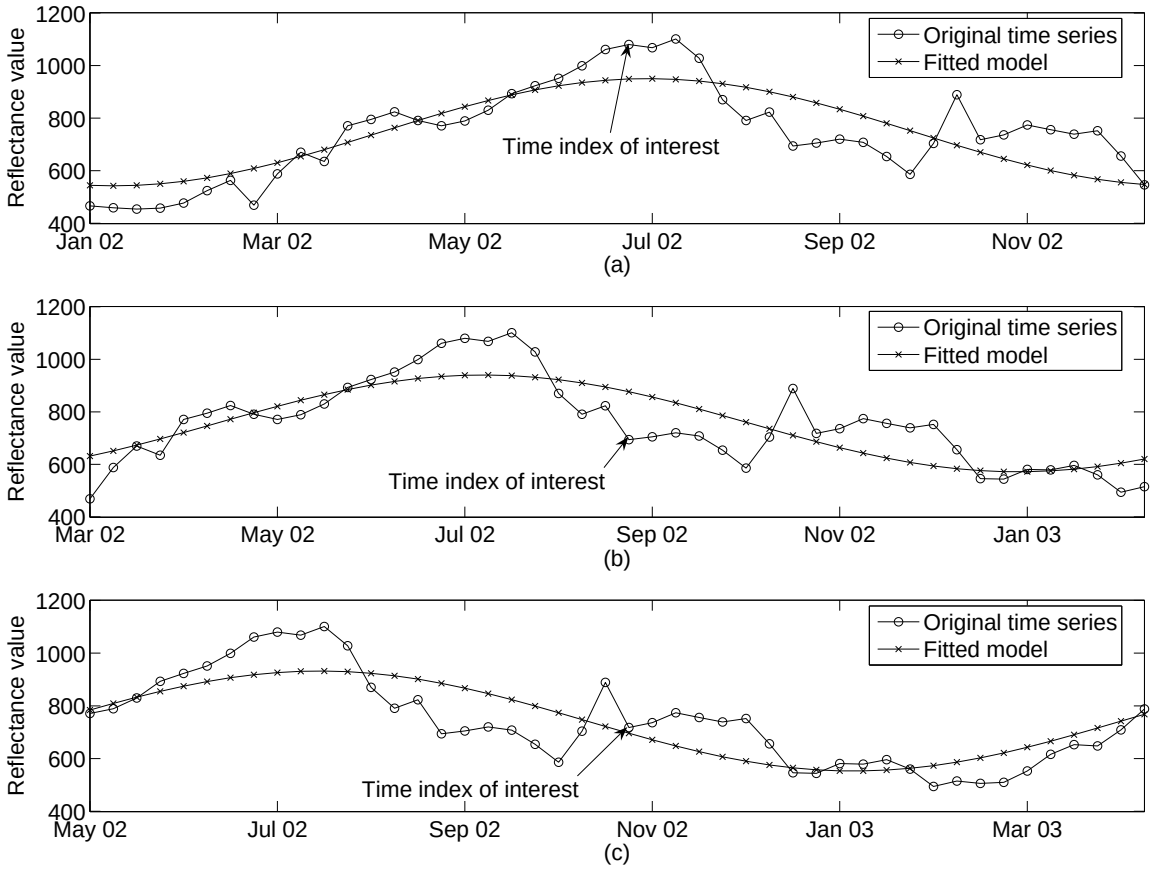

FIGURE 5.6: Least squares estimates the parameter vector $\vec{W}_i$ by shifting the model over the time series.

The variable $\mathcal{E}_{\mathrm{LS}}$ denotes the sum of squares and $\mathbf{h}$ denotes the model. The sum of squares can be minimised using standard approaches, which evaluate the partial derivatives. The partial derivative of the sum of squares is solved as

$$\frac{d\mathcal{E}_{\mathrm{LS}}}{d\vec{W}_i} = 2\sum_{j=1}^{\mathcal{I}}(\vec{x}_j - \hat{\vec{x}}_j)\frac{d(\vec{x}_j - \hat{\vec{x}}_j)}{d\vec{W}_i} = 0, \qquad \forall i. \tag{5.52}$$

Several variations of the least squares exist; the most popular method is the ordinary least squares (OLS) algorithm. The OLS assumes the observation noise vector $\vec{v}_i$ is normally distributed and the model $\mathbf{h}$ is linear.

The least squares is considered optimal when a set of criteria is met in the estimates of the parameter vector. These criteria are:

1. The observation vectors are randomly sampled from a well defined data set.

2. The underlying structure within the data set is linear.

3. The difference between the observation vector $\vec{x}_i$ and the fitted model has an expected zero mean.

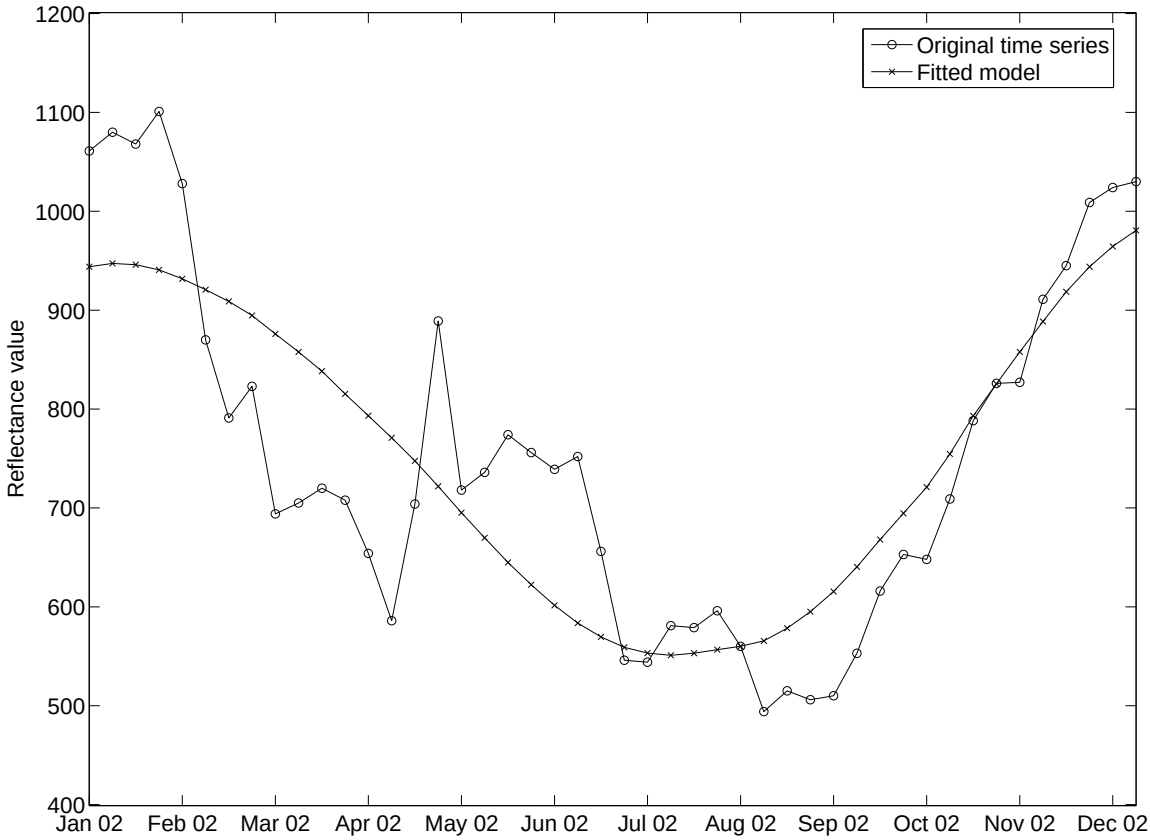

FIGURE 5.7: Least squares estimates the parameter vector $\vec{W}_i$ to fit triply modulated cosine model onto a time series.

4. The parameter vector's variables are linearly independent from each other.

5. The difference between the observation vector $\vec{x}_i$ and the fitted model is normally distributed and uncorrelated to the parameter vector.

In addition to the five criteria stated, if the Gauss-Markov condition also holds; then the OLS estimates are considered to be equivalent to the maximum likelihood estimates of the parameter vectors. More sophisticated adaptations have been made to the OLS and the most frequently used of these are: the weighted least squares, alternating least squares and partial least squares.

The OLS can be extended to include the field of non-linear models. The drawback is that the standard approach of evaluating the derivative of a non-linear model in equation (5.52) is not always possible. This is because the derivatives of $d(\vec{x}_j - \hat{\vec{x}}_j)/d\vec{W}_i$ are functions which are dependent on both the observation vectors $\{\vec{x}_i\}$ and the parameter vectors $\{\vec{W}_i\}$.

This changes the least squares from a closed-form solution for the linear case to a non closed-form solution for the non-linear case. This requires that the estimation of the set of parameter vectors $\{\vec{W}_i\}$ is derived using an analytical iterative algorithm. The algorithm iterates through the parameter vector's

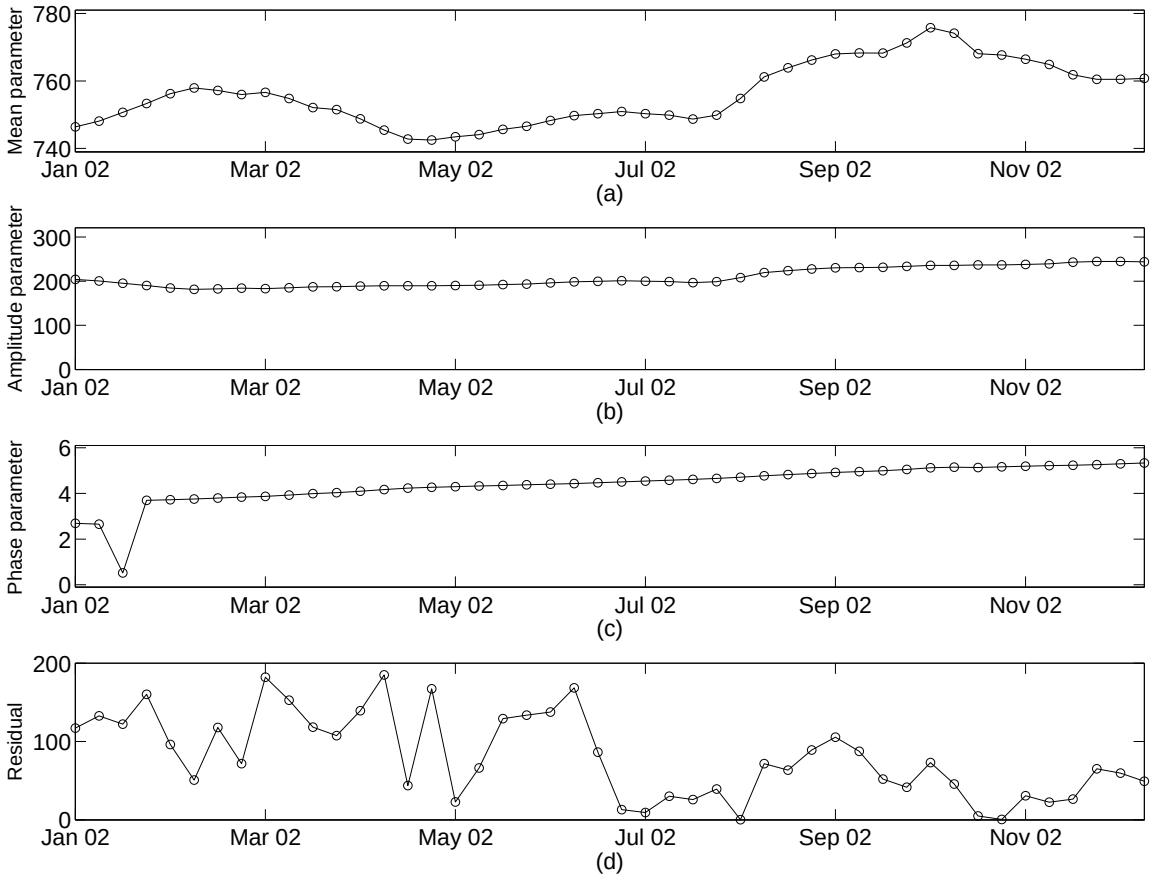

FIGURE 5.8: Least squares estimates the parameter vector $\vec{W}_i$ to fit triply modulated cosine model onto a time series.

space using the derivative of the sum of squares $\mathcal{E}_{\text{LS}}$ at each epoch. The gradient descent algorithm is a popular iterative method used in this case.

Land cover example: In this example the least squares predicts the set of parameter vectors for the time series shown in figure 5.2. The problem lies in the fact that the least squares requires a set of observation vectors $\{\vec{x}_i\}$ to estimate a single parameter vector $\vec{W}_i$. The lowest number of observation vectors required to estimate the parameter vector is $(|\vec{W}_i| + 1)$.

This concept is illustrated in figure 5.5 by using a set of observation vectors the length of a single year. In figure 5.5(a) the time series in figure 5.2 is shown with a time index of interest. The parameter vector $\vec{W}_i$ for observation vector $\vec{x}_i$ is estimated using the set $\{\vec{x}_{i-N}, \vec{x}_{i-N+1}, \ldots, \vec{x}_{i+N-1}, \vec{x}_{i+N}\}$ of observation vectors. The variable N is chosen to encapsulate the entire period of the model shown in figure 5.5(b). The parameter vector $\vec{W}_i$ is then determined using the least squares to minimise the sum of squares to produce the fitted model shown in figure 5.5(c).

The next step is to estimate a parameter vector $\vec{W}_i, \forall i$. This is accomplished by moving the

model across the time index. The parameter vector $\vec{W}_{i+c}$ for observation vector $\vec{x}_{i+c}$ is estimated using the set $\{\vec{x}_{i-N+c}, \vec{x}_{i-N+c+1}, \ldots, \vec{x}_{i+N+c-1}, \vec{x}_{i+N+c}\}$. This iterative approach to moving the model is shown in three different figures in figure 5.6.

After shifting through the entire time series, the predicted output of the least squares is plotted, along with the actual observation vectors in figure 5.7.

The progressive estimation of the parameter vectors is shown in figure 5.8. Figure 5.8(a) illustrates the estimation of the model's mean parameter μ_i. Figure 5.8(b) illustrates the estimation of the model's amplitude parameter α_i. Figure 5.8(c) illustrates the estimation of the model's phase parameter θ_i. The absolute error in tracking of the output is illustrated in figure 5.8(d). □

5.7 M-ESTIMATE MODEL FITTING

Various attempts have been made to create robust statistical estimators, which are used to fit models. M-estimates rely on the maximum likelihood approach to estimate the parameters of a particular statistical model. An M-estimator is generally defined as a zero of the estimating function, while the estimating function is usually the derivative of a statistical function of interest. The advantage of a M-estimator is that it does not assume that the residuals are normally distributed. M-estimators attempt to minimise the mean absolute deviation in the residuals for a given distribution using a maximum likelihood approach.

The assessment of different distributions in the M-estimator allow for different weighting functions to be associated with outliers. Normally distributed residuals usually associate greater weights to outliers when compared to a Lorentzian distribution of residuals [189, Ch. 15]. This deviant behaviour in relative weighting points in a model makes it difficult to apply standard gradient descent. The Nelder-Mead method is thus the chosen optimisation method, as it only requires function evaluations and not the derivatives [189, Ch. 15].

The Nelder-Mead algorithm was first proposed by John Nelder and Roger Mead in 1965 [190]. The Nelder-Mead algorithm is a non-linear method which estimates the parameter vector $\vec{W}_i$ for a particular model. The Nelder-Mead algorithm is a well-defined numerical method that operates on a twice differentiable, unimodal, multi-dimensional function. The method makes use of a direct search by evaluating a function at the vertices of a simplex. A N-simplex is a N-dimensional polytope which is the convex hull of (N+1) vertices. The algorithm then iteratively moves and scales the simplex's vertices through the set of dimensions in search of the minimum. It continually attempts to improve the evaluated function until a predefined bound is reached.

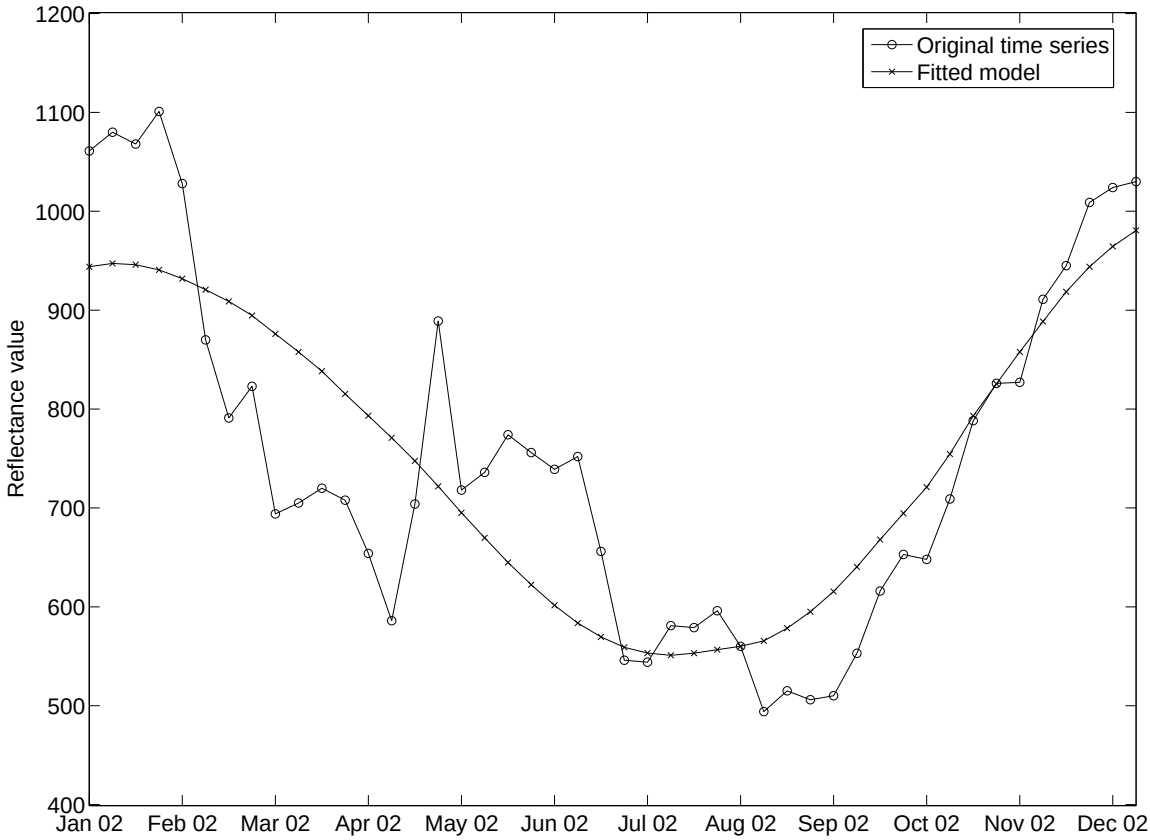

FIGURE 5.9: M-estimator estimates the parameter vector $\vec{W}_i$ to fit the triply modulated cosine model onto a time series.

Each epoch requires the execution of six steps to compute the new position of the simplex. The algorithm in summary starts with initialising the vertices of the simplex. It then iteratively rejects and replaces the worst performing vertex point with a new vertex point. This process of setting new vertex points creates a sequence of new N-simplexes. The initialisation with a small initial N-simplex converges rapidly to a local minimum, while a large N-simplex becomes trapped in non-stationary points in the vector space.

Land cover example: In this example the M-estimator predicts a set of parameter vectors for the time series shown in figure 5.2. The same problem exists for the M-estimator, as for the least squares, when estimating the sequence of parameter vectors. The parameter vector $\vec{W}_i$ for observation vector $\vec{x}_i$ is estimated using the set $\{\vec{x}_{i-N}, \vec{x}_{i-N+1}, \ldots, \vec{x}_{i+N-1}, \vec{x}_{i+N}\}$ of observation vectors. This is rectified by shifting the model through all the time indices. The initial estimate of the M-estimator is contained in a certain parameter space by using the mean and standard deviation of the time series as the initial parameter vector for the model. The previous parameter vector $\vec{W}_{i-1}$ is then used to initialise the M-estimator when determining the current parameter vector $\vec{W}_i$.

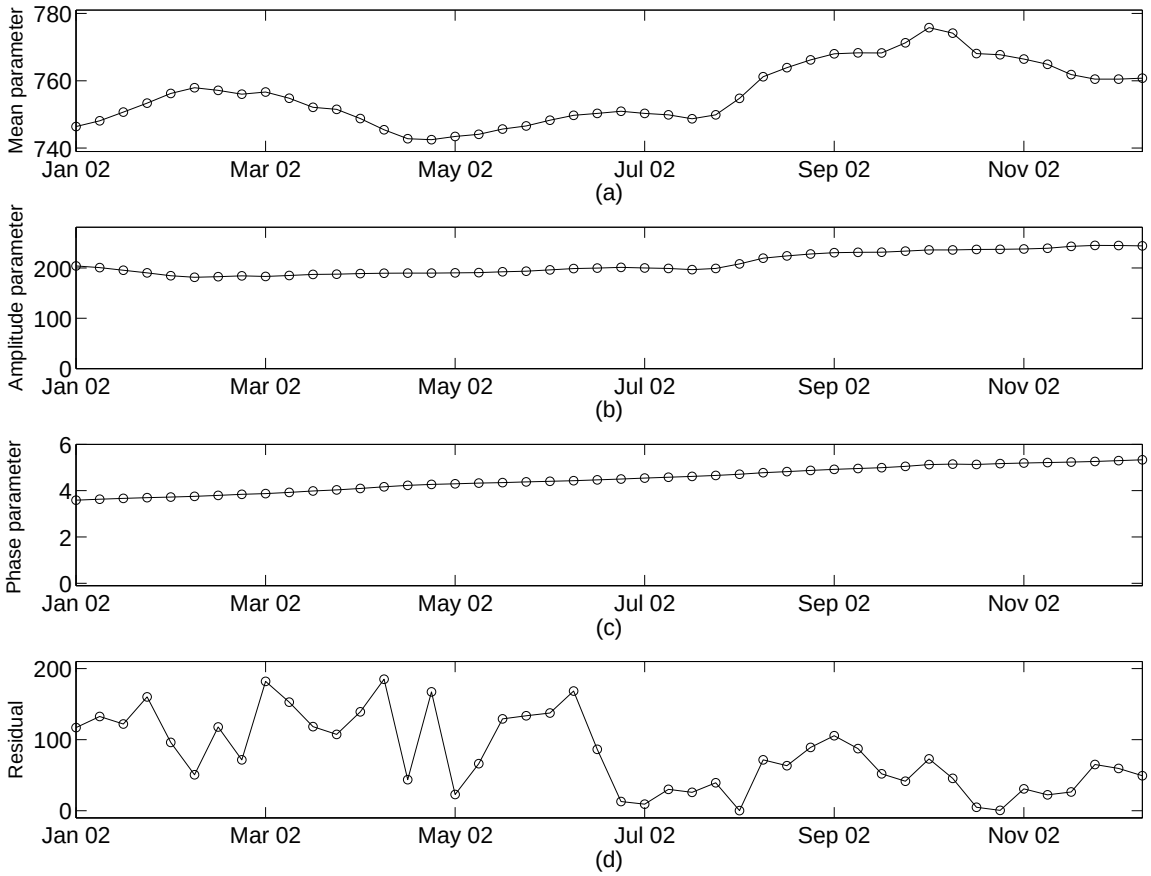

FIGURE 5.10: M-estimator estimates the parameter vector $\vec{W}_i$ to fit the triply modulated cosine model onto a time series.

The predicted output of the M-estimator is plotted with the actual observation vectors $\vec{x}_i$ in figure 5.9.

The progressive estimation of the parameter vectors are shown in figure 5.10. Figure 5.10(a) illustrates the estimation of the model's mean parameter μ_i. Figure 5.10(b) illustrates the estimation of the model's amplitude parameter α_i. Figure 5.10(c) illustrates the estimation of the model's phase parameter θ_i. The absolute error in the tracking of the output is illustrated in figure 5.10(d). □

5.8 FOURIER TRANSFORM

The Fourier transform of a discrete time series is a representation of the sequence in terms of the complex exponential sequence $\{e^{j2\pi fi}\}$, where f is the frequency variable. The Fourier transform representation of a time series, if it exists, is unique and the original time series can be recovered by applying an inverse Fourier transform [115, Ch. 3].

Let $\mathbf{x}$, $\mathbf{x} = [x_1\ x_2\ \ldots\ x_{\mathcal{I}}]$, denote the time series and let $\mathcal{I} \to \infty$, then the Fourier transform $\mathcal{X}(e^{j2\pi f})$ is defined as

$$\mathcal{X}(e^{j2\pi f}) = \sum_{i=-\infty}^{\infty} x_{(\mathcal{I}/2)} e^{j2\pi f i}. \tag{5.53}$$

The Fourier transform $\mathcal{X}(e^{j2\pi f})$ is a complex function and is written in rectangular form as

$$\mathcal{X}(e^{j2\pi f}) = \mathcal{X}_{\text{real}}(e^{j2\pi f}) + j\mathcal{X}_{\text{imag}}(e^{j2\pi f}), \tag{5.54}$$

where $\mathcal{X}_{\text{real}}(e^{j2\pi f})$ denotes the real part and $\mathcal{X}_{\text{imag}}(e^{j2\pi f})$ denotes the imaginary part of $\mathcal{X}(e^{j2\pi f})$. The components of the rectangular form are expressed as

$$\mathcal{X}_{\text{real}}(e^{j2\pi f}) = |\mathcal{X}(e^{j2\pi f})| \cos\theta_{\mathcal{X}}, \tag{5.55}$$

$$\mathcal{X}_{\text{imag}}(e^{j2\pi f}) = |\mathcal{X}(e^{j2\pi f})| \sin\theta_{\mathcal{X}}. \tag{5.56}$$

The quantity $|\mathcal{X}(e^{j2\pi f})|$ denotes the magnitude function of the Fourier transform. The quantity $\theta_{\mathcal{X}}$ denotes the phase function, which is given as

$$\theta_{\mathcal{X}} = \arctan\left(\frac{\mathcal{X}_{\text{imag}}(e^{j2\pi f})}{\mathcal{X}_{\text{real}}(e^{j2\pi f})}\right). \tag{5.57}$$

In the case of a finite length time series $\mathbf{x}$, $\mathbf{x} = [x_1\ x_2\ \ldots\ x_{\mathcal{I}}]$, $\mathcal{I} \in \mathbb{N}$, $\mathcal{I} < \infty$, there is a simpler relation between the time series and its corresponding Fourier transform $\mathcal{X}(e^{j2\pi f})$ [115, Ch. 3]. For a time series $\mathbf{x}$ of length $\mathcal{I}$, only $\mathcal{I}$ values of $\mathcal{X}(e^{j2\pi f})$ at $\mathcal{I}$ distinct harmonic functions at frequency points, $0 \leq f \leq \mathcal{I}$, are sufficient to construct the unique time series $\mathbf{x}$. This leads to the concept of a second transform domain representation that operates on a finite length time series [115, Ch. 3].

This second transform is known as the discrete Fourier transform (DFT). The relation between a finite length time series $\mathbf{x}$, $\mathbf{x} = [x_1\ x_2\ \ldots\ x_{\mathcal{I}}]$, and its corresponding Fourier transform $\mathcal{X}(e^{j2\pi f})$ is obtained by uniformly sampling $\mathcal{X}(e^{j2\pi f})$ on the frequency domain between $0 \leq f \leq 1$ at increments of $f = i/\mathcal{I}$, $0 \leq i \leq (\mathcal{I}-1)$. The DFT is computed by sampling equation (5.53) uniformly as

$$\mathcal{X}_i = \mathcal{X}(e^{j2\pi f})\Big|_{f=i/\mathcal{I}} = \sum_{n=0}^{\mathcal{I}-1} x_n e^{j2\pi i n/\mathcal{I}}, \quad 0 \leq i \leq (\mathcal{I}-1). \tag{5.58}$$

The inverse discrete Fourier transform (IDFT) is given by

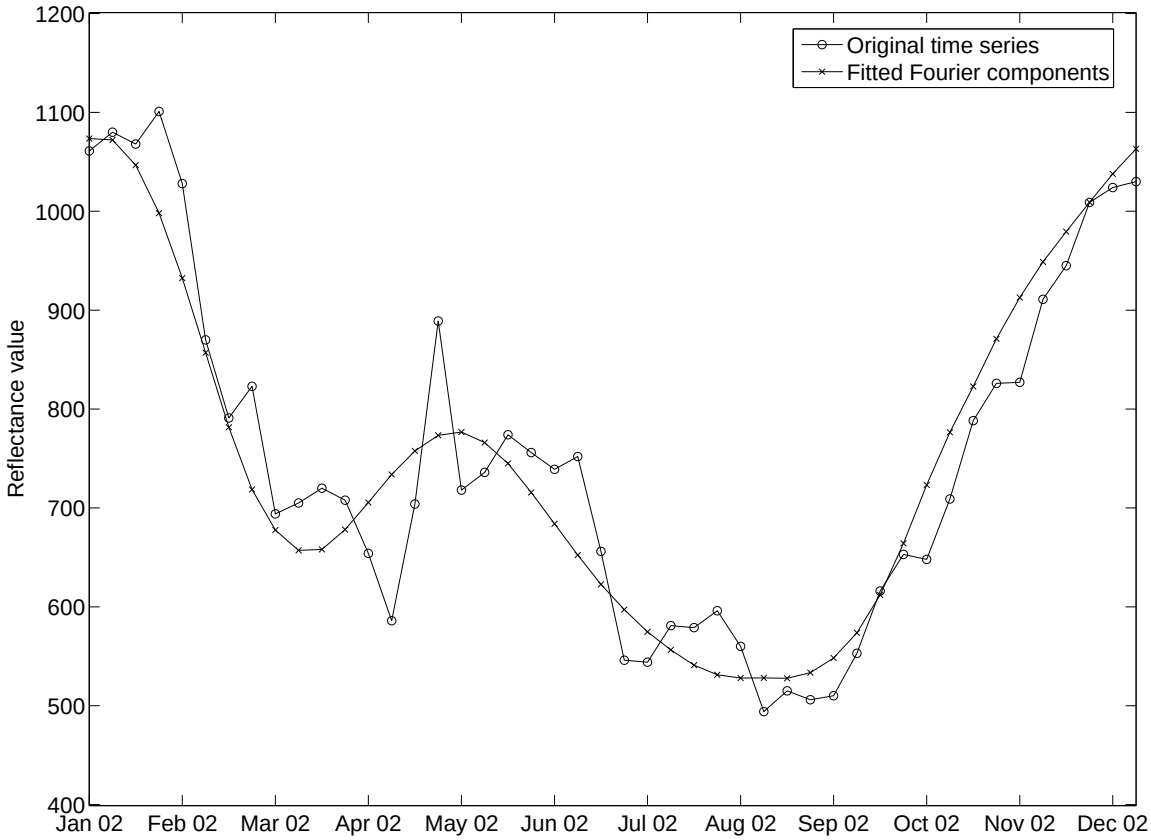

FIGURE 5.11: Fast Fourier transform (FFT) estimates the parameters of the vector $\vec{W}_i$ to fit multiple harmonics onto time series $\mathbf{x}$.

$$x_n = \sum_{i=0}^{\mathcal{I}-1} \mathcal{X}_i e^{-j2\pi in/\mathcal{I}}, \quad 0 \leq n \leq (\mathcal{I}-1). \tag{5.59}$$

The computation of the DFT and IDFT requires $\mathcal{O}(\mathcal{I}^2)$ complex multiplications and $\mathcal{O}(\mathcal{I}^2 - \mathcal{I})$ complex additions. A fast Fourier transform (FFT) refers to an algorithm that has been developed to reduce the computational complexity of computing the DFT to about $\mathcal{O}(\mathcal{I}(\log_2 \mathcal{I}))$ operations. As there is no loss in precision in using these fast computing algorithms, they will be used throughout this thesis when referring to the DFT of a time series. Similarly, an inverse fast Fourier transform (IFFT) algorithm has been developed to compute the IDFT efficiently.

The FFT function is denoted by $\mathfrak{F}$ and is mathematically computed as

$$\mathcal{X} = \mathfrak{F}(\mathbf{x}). \tag{5.60}$$

The sequence $\mathcal{X}$ is the DFT of the time series $\mathbf{x}$. The time series $\mathbf{x}$ is a process in the time domain and the value of $\mathbf{x}$ is dependent on the corresponding time index i. The DFT $\mathcal{X}$, on the other hand, is a process in the frequency domain by which the process is defined by the amplitude $|\mathbf{x}_f|$ and phase $\angle \mathbf{x}_f$

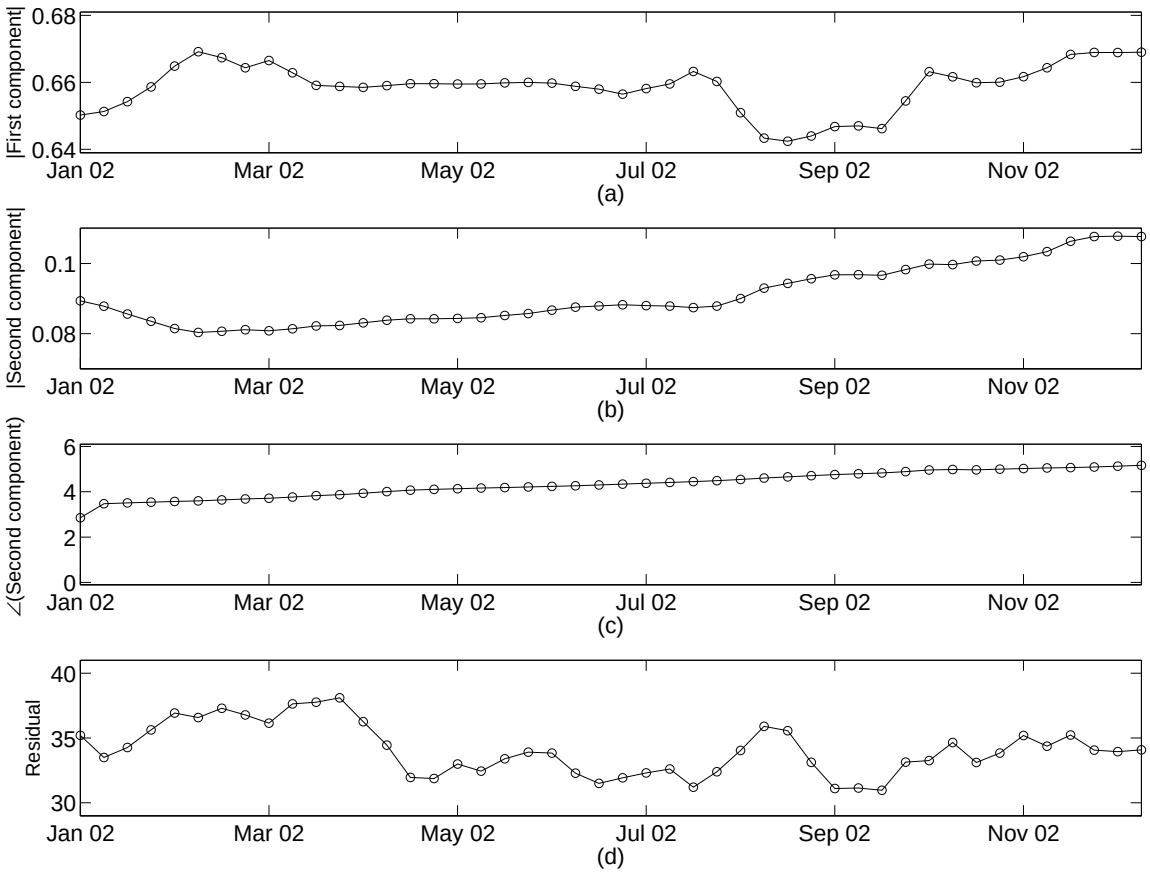

FIGURE 5.12: Fast Fourier transform (FFT) estimates the parameters of the vector $\vec{W}_i$ to fit multiple harmonics onto time series $\mathbf{x}$.

of harmonic frequency samples f, $f \in \{-\infty, \infty\}$.

The inverse Fourier transform is denoted by $\mathfrak{F}^{-1}$ and is mathematically computed as

$$\mathbf{x} = \mathfrak{F}^{-1}(\mathcal{X}). \tag{5.61}$$

The conversion to the frequency domain allows the analysis of periodic (such as seasonal) effects and trends within the time series $\mathbf{x}$.

Land cover example: In this example the fast Fourier transform is used to predict a set of Fourier components for the time series shown in figure 5.2.

The Fourier components are stored in a vector $\vec{W}_i$ for observation vector $\vec{x}_i$ and are estimated using the set $\{\vec{x}_{i-N}, \vec{x}_{i-N+1}, \ldots, \vec{x}_{i+N-1}, \vec{x}_{i+N}\}$ of observation vectors. The variable N is chosen to capture enough energy in each harmonic function of interest. This happens to be the entire process function of a complete phenological cycle of one year.

A set of harmonic functions is stored in the state-space model as

$$\vec{W}_i = [W_{i,1}\ W_{i,2}\ W_{i,3}] = [W_{i,\mu}\ W_{i,\alpha}\ W_{i,\theta}] = [|\mathcal{X}_1|\ \ 2|\mathcal{X}_2|\ \ \angle(\mathcal{X}_2)]. \tag{5.62}$$

The next step is to estimate a vector $\vec{W}_i, \forall i$. This is accomplished by moving a window across the time index. The vector $\vec{W}_{i+c}$ for observation vector $\vec{x}_{i+c}$ is estimated using the set $\{\vec{x}_{i-N+c}, \vec{x}_{i-N+c+1}, \ldots, \vec{x}_{i+N+c-1}, \vec{x}_{i+N+c}\}$. This iterative approach moves the window of the DFT similar to the least squares and M-estimator. The predicted output of the Fourier components is plotted along with the actual observation vectors in figure 5.11.

The progressive estimation of the vectors is shown in figure 5.12. Figure 5.12(a) illustrates the estimation of the magnitude of the first frequency component in $\mathcal{X}$. Figure 5.12(b) illustrates the estimation of the magnitude of the second frequency component in $\mathcal{X}$. Figure 5.12(c) illustrates the phase of the second frequency component $\mathcal{X}$. The absolute error in tracking of the output is illustrated in figure 5.12(d). □

5.9 SUMMARY

In this chapter, four different feature extraction methods were investigated. The feature extraction methods are all based on the same principle of fitting a cosine model to the time series. The first three methods; EKF, least squares model fitting and M-estimator model fitting, are regression approaches, which attempt to estimate the mean, amplitude, and phase component of the cosine function. All three features are comparable among the three regression methods. The Fourier transform method is similar to the other three methods, except for the fact that a complex vector is estimated, which contains the combined power of both a cosine and sine function. The feature vectors extracted using these methods will be used by machine learning methods to determine the corresponding class labels.