Chapter 2
Methodology

2.1 Extreme Value Theory

The main idea of this thesis is the application of Extreme Value Econometrics to analyse the palm oil prices. Extreme Value Theory (EVT) is the concept of modeling and measuring extreme events which occurs with a very small probability (Brodin and Kluppelberg, 2008). It provides the methods for quantifying such events and their consequences statistically. Generally, there are two principal approaches to identify the extremes in real data. The Block Maxima (BM) and Peaks-Over-Threshold (POT) are central for the statistical analysis of maxima or minima and of exceedances over a higher or lower threshold (Lai and Wu, 2007). This thesis uses both univariate and bivariate extreme value, as each type of estimation adopt BM and POT. Finally, the third objective of this paper employs the extreme value copulas model.

The first objective applies both univariate BM and POT to predict extreme events of palm oil prices in the future.

2.2 Univariate Block Maxima

The BM model studies the statistical behaviour of the largest or the smallest value in a sequence of independent random variables (Lei and Qiao, 2010; Lei et al., 2011). One approach to working with extreme value data is to group the data into blocks of equal length and to fit the data to the maximums of each block whilst assuming that n (number of blocks) is correctly identified.

Let Z_i (i=1,…,n) denote the maximum observations in each block (Coles, 2001). Z_n is normalized to obtain a non-degenerated limiting distribution. The BM approach is closely associated with the use of Generalized Extreme Value (GEV) distribution with cumulative density function (c.d.f) (Lei and Qiao, 2010):

$$G(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \qquad (1)$$

Where μ, σ > 0 and ξ are location, scale and shape parameter, respectively. The GEV includes three extreme value distributions as special cases: the Frechet distribution is ξ > 0, the Fisher-Tippet or Weibull distribution is ξ < 0, and the Gumbel or double-exponential distribution is ξ = 0. Depending on the parameter ξ, a distribution function is classified as fat tailed (ξ > 0), thin tailed (ξ = 0) and short tailed (ξ < 0) (Odening and Hinrichs, 2003). Under the assumption that $Z_1, \ldots, Z_n$ are independent variables having the GEV distribution, the log-likelihood for the GEV parameters when ξ ≠ 0 is given by (Coles, 2001):

$$\ell(\xi, \mu, \sigma) = -n\log \sigma - (1+1/\xi) \sum_{i=1}^{n} \log\left[1+\xi\left(\frac{Z_i-\mu}{\sigma}\right)\right] - \sum_{i=1}^{n}\left[1+\xi\left(\frac{Z_i-\mu}{\sigma}\right)\right]^{-1/\xi} \quad (2)$$

provided that $1+\xi\left(\frac{Z_i-\mu}{\sigma}\right) > 0$, for i=1,….,n

The case ξ = 0 requires separate treatment using the Gumbel limit of the GEV distribution. The log-likelihood in that case is:

$$\ell(\mu, \sigma) = -n\log \sigma - \sum_{i=1}^{n}\left(\frac{Z_i-\mu}{\sigma}\right) - \sum_{i=1}^{n}\exp\left\{-\left(\frac{Z_i-\mu}{\sigma}\right)\right\} \quad (3)$$

The maximization of this equation with respect to the parameter vector (μ, σ, ξ) leads to the maximum likelihood estimate with respect to the entire GEV family (Coles, 2001).

Estimates and confidence intervals with the maximum likelihood estimate of z_p for $0 < p < 1$, the $1/p$ return level, is obtained as

$$\hat{z}_p = \mu - \frac{\hat{\sigma}}{\hat{\xi}}\left[1 - y_p^{-\hat{\xi}}\right], \textit{ for } \hat{\xi} \neq 0,$$

$$\hat{z}_p = \mu - \sigma \hat{\log} y_p, \textit{ for } \hat{\xi} = 0, \quad (4)$$

where $y_p = -\log(1-p)$. Moreover, by the delta method,

$$Var(\hat{z}_p) \approx \nabla z_p^{\mathrm{T}} V \nabla z_p, \quad (5)$$

where V is the variance-covariance matrix of $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ and

$$\nabla z_p^{T} = \left[\frac{\partial z_p}{\partial \mu}, \frac{\partial z_p}{\partial \sigma}, \frac{\partial z_p}{\partial \xi}\right] = [1, -\xi^{-1}(1-y_p^{-\xi}), \sigma\xi^{-2}(1-y_p^{-\xi}) - \sigma\xi^{-1}y_p^{-\xi}\log y_p]$$

evaluated at $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ (Coles, 2001).

2.3 Univariate Peak-Over-Threshold

The POT approach is based on the Generalized Pareto Distribution (GPD) introduced by Pickands (1975) (cited in Lei and Qiao, 2010). The GPD estimation involves two steps, the choice of threshold u and the parameter estimations for ξ and σ which can be done using Maximum Likelihood Estimation (Bensalah, 2000). These are models for all large observations that exceed a high threshold. The POT approach deals with the distribution of excess over a given threshold wherein the modelling is to understand the behaviour of the excess loss once a high threshold (loss) is reached (McNeil, 1999). Previous studies have shown that if the block maxima have an approximate distribution of GEV, then the excesses from the threshold have a corresponding Generalized Pareto Distribution (GPD) with c.d.f. (Lai and Wu, 2007, Lei and Qiao, 2010):

$$H(y) = 1 - \left(1 + \frac{\xi y}{\sigma}\right)^{-1/\xi} \qquad (6)$$

defined on {y: y > 0 and $\left(1 + \frac{\xi y}{\sigma}\right)$ > 0}, where y (growth rate price exceeds) is random variable, σ (σ > 0) and ξ (-∞ < ξ < +∞) are scale and shape parameters, respectively. The family of distributions defined by this equation is called the GPD family. Having determined a threshold, the parameters of GPD can be estimated by log-likelihood.

Suppose that the values $Y_1, \ldots, Y_n$ are the n excesses of a threshold u. For ξ ≠ 0, the log-likelihood is (Coles, 2001)

$$\ell(\sigma, \xi) = -n\log\sigma - (1+1/\xi)\sum_{i=1}^{n}\log(1+\xi y_i/\sigma) \qquad (7)$$

provided that $(1+\xi y_i/\sigma) > 0$ for i=1,…,n

The maximum likelihood procedures can also be utilized to estimate the GPD parameters, given the threshold (Lei and Qiao, 2010).

Estimates and confidence intervals with the maximum likelihood estimate of z_N for the N-year return level is defined by

$$z_N = u + \frac{\sigma}{\xi}\left[(Nn_y\zeta_u)^{\xi} - 1\right], \; for\, \xi \neq 0,$$

$$z_N = u + \sigma \log(Nn_y\zeta_u), \; for\, \xi = 0, \tag{8}$$

where N is the level expected to be exceeded once every N years. If there are n_y observations per year, this corresponds to the m-observation return level, where $m = N \times n_y$. ζ_u is the probability of an individual observation exceeding the threshold u. The $\hat{\zeta}_u$ is the maximum likelihood estimate of ζ_u. This has a natural estimator of $\hat{\zeta}_u = \frac{k}{n}$, the sample proportion of points exceeding u. Furthermore, by the delta method,

$$Var(\hat{z}_m) \approx \nabla z_m^{T} V \nabla z_m, \tag{9}$$

where

$$\nabla z_m^{T} = \left[\frac{\partial z_m}{\partial \zeta_u}, \frac{\partial z_m}{\partial \sigma}, \frac{\partial z_m}{\partial \xi}\right]$$

$$= [\sigma m^{\xi}\zeta_u^{\xi-1}, \xi^{-1}\{(m\zeta_u)^{\xi} - 1\}, -\sigma\xi^{-2}\{(m\zeta_u)^{\xi} - 1\} + \sigma\xi^{-1}(m\zeta_u)^{\xi}\log(m\zeta_u)],$$

evaluated at $(\hat{\zeta}_u, \hat{\sigma}, \hat{\xi})$ (Coles, 2001).

The second objective uses the bivariate BM and POT to analyze the relationship between soybean oil and palm oil prices, crude oil and palm oil prices.

2.4 Bivariate Block Maxima

This method is concerned with parametric and non-parametric cases. This thesis chooses the parametric models. A brief summary of bivariate BM is given below:

Let (X, Y) denote a bivariate random vector representing the component-wise maxima of an i.i.d. sequence over a given period of time. Under the appropriate conditions the distribution of (X, Y) can be approximated by a bivariate extreme value distribution (BEVD) with c.d.f. G. The BEVD is determined by its two univariate margins G_1 and G_2 respectively, which are necessarily EVD, and by its Pickands dependence function A (Rakonczai & Tajvidi, 2010).

$$G(x,y)=\exp\left\{\log(G_1(x)G_2(y))\times A\left(\frac{\log(G_2(y))}{\log(G_1(x)G_2(y))}\right)\right\} \qquad (10)$$

A(W) is responsible for capturing the dependence structure between the margins and determines only up to the condition that it is convex, passes through the points (0,1), (1,1) and (1/2,1/2) binds the upper left and right corners. The properties of function A are (1) A(w) is convex, (2) max{(1 − w), w} ≤ A(w) ≤ 1 and (3) $A(0) = A(1) = 1$. Rakonczai and Tajvidi, (2010) explained in their paper that the lower bounds in the second item of the properties of A corresponds to the complete dependence $G(x,y) = \min\{G1(x),G2(y)\}$, while the upper bound corresponds to (complete) independence $G(x,y) = G1(x)G2(y)$.

The Parametric Bivariate Extreme Value Distributions have nine models (Stephenson, 2011) as follow:

The logistic distribution function with parameter dep = r is

$$G(x,y)=\exp[-(x^{\frac{1}{r}}+y^{\frac{1}{r}})^r] \qquad (11)$$

where $0<r\le 1$. The independence case corresponds to $r=1$. For $r\to 0$, we get complete dependence.

The asymmetric logistic distribution function with parameter dep = r and asy = (t_1,t_2) is

$$G(x,y)=\exp\left\{-(1-t_1)x-(1-t_2)y-[(t_1x)^{\frac{1}{r}}+(t_2y)^{\frac{1}{r}}]^r\right\} \qquad (12)$$

where $0<r\le 1$ and $0\le t_1,t_2\le 1$. When $t_1=t_2=1$, the model reduces to the symmetric logistic model. Independence is obtained by $r=1$ together with either $t_1=0$ or $t_2=0$. Complete dependence is obtained in the limit when $t_1=t_2=1$ and $r\to 0$.

The Husler-Reiss (HR) distribution function with parameter dep = r is

$$G(x,y)=\exp(-x\Phi\{r^{-1}+\tfrac{1}{2}r[\log(x/y)]\}-y\Phi\{r^{-1}+\tfrac{1}{2}r[\log(y/x)]\}) \qquad (13)$$

where Φ(.) is the standard normal distribution function and $r > 0$. Independence is obtained in the limit as $r \to 0$. Complete dependence is obtained as r tends to ∞.

The negative logistic distribution function with parameter dep = r is

$$G(x, y) = \exp\left\{-x - y + [x^{-r} + y^{-r}]^{-1/r}\right\} \tag{14}$$

where $r > 0$. Independence is obtained in the limit as $r \to 0$. Complete dependence is obtained as r tends to ∞.

The asymmetric negative logistic distribution function with parameters dep = r and asy = (t_1, t_2) is

$$G(x, y) = \exp\{-x - y + [(t_1 x)^{-r} + (t_2 y)^{-r}]^{-1/r}\} \tag{15}$$

where $r > 0$ and $0 < t_1, t_2 \le 1$. When $t_1 = t_2 = 1$, the model reduces to the negative logistic model. Independence is obtained in the limit as either r, t_1 or t_2 approaches zero. Complete dependence is obtained in the limit when $t_1 = t_2 = 1$ and r tends to infinity.

The bilogistic distribution function with parameters α and β is

$$G(x, y) = \exp\left\{-xq^{1-\alpha} - y(1-q)^{1-\beta}\right\} \tag{16}$$

where $q = q(x, y; \alpha, \beta)$ is the root of the equation

$(1-\alpha)x(1-q)^{\beta} - (1-\beta)yq^{\alpha} = 0, 0 < \alpha, \beta < 1.$

when $\alpha = \beta$ the bilogistic model is equivalent to the logistic model with dependence parameter dep = α = β. Independence is obtained as α = β approaches to one, and when one of α, β is fixed and the other approaches to one. Complete dependence is obtained in the limit as α = β approaches to zero.

The negative bilogistic distribution function with parameters α and β is

$$G(x, y) = \exp\left\{-x - y + xq^{1+\alpha} + y(1-q)^{1+\beta}\right\} \tag{17}$$

where $q = q(x, y; \alpha, \beta)$ is the root of the equation

$(1+\alpha)xq^{\alpha} - (1+\beta)y(1-q)^{\beta} = 0, \alpha > 0, \beta > 0.$

when $\alpha = \beta$ the negative bilogistic model is equivalent to the negative logistic model with dependence parameter dep = $1/\alpha = 1/\beta$. Independence is obtained as α = β tends to ∞, and when one of α, β is fixed and the other tends to ∞. Complete dependence is obtained in the limit as α = β approaches to zero.

The Coles-Tawn distribution function with parameters α>0 and β>0 is

$$G(x,y) = \exp\{-x[1 - Be(q;\alpha+1,\beta)] - yBe(q;\alpha,\beta+1)\} \quad (18)$$

where $q = \alpha y/(\alpha y + \beta x)$ and $\mathrm{Be}(q;\alpha,\beta)$ is the beta distribution function evaluated at q with shape1= α and shape2 = β. Independence is obtained as α = β approaches to zero, and when one of α, β is fixed and the other approaches to zero. Complete dependence is obtained in the limit as α = β tends to ∞.

The asymmetric mixed distribution function with parameters α and β has a dependence function with the cubic polynomial form is shown below

$$A(w) = \beta w^3 + \alpha w^2 - (\alpha+\beta)w + 1 \quad (19)$$

where $\alpha \ge 0$ and $\alpha + 3\beta \ge 0$, $\alpha + \beta \le 1$ and $\alpha + 2\beta \le 1$. These constraints imply that β lies in the interval [-0.5,0.5] and that α lies in the interval [0,1.5], though α can only be greater than one if β < 0. Complete dependence cannot be obtained. Independence is obtained when both parameters are zero.

2.5 Bivariate Threshold Exceedances

There are at least two ways of defining exceedances in higher dimensions. In the first definition, a distribution is fitted to the observations $\{(x,y) | (x,y) > (u_x,u_y)\}$ where u_x and u_y are suitable thresholds for each margin. Second definition aims to fit a distribution to $\{(x,y) | (x,y) \nless (u_x,u_y)\}$ where (u_x,u_y) is defined as before. These distributions will be called Type I and Type II bivariate generalized Pareto distributions (BGPD), respectively (Coles & Tawn, 1991), (Coles, 2001).

In this thesis, the strength of the dependence between variables is estimated by fitting joint exceedances to bivariate extreme value distribution using BGPD type I. From univariate GPD, the details for approximating the tail of X by

$$G(x) = 1 - \eta_u \left(1 + \xi \frac{x-u}{\sigma}\right)^{-\frac{1}{\xi}}, x \ge u \quad (20)$$

$\eta_u = P(X > u)$

Suppose $(x_1, y_1),\ldots,(x_n, y_n)$ are independent realizations of a random variable (X,Y) with joint distribution function $F(x, y)$ on regions of the from $x > u_x, y > u_y$, for large enough u_x and u_y. The marginal distributions of F each have an approximation of equation (16), with respective parameter sets $(\eta_x, \sigma_x, \xi_x)$ and $(\eta_y, \sigma_y, \xi_y)$ (Coles, 2001). The finding of this equation can approximate the tail of X and Y for $x > u_x, y > u_y$ with $G(x : \eta_x, \sigma_x, \xi_x)$ and $G(y : \eta_y, \sigma_y, \xi_y)$, respectively. The Bivariate Generalized Pareto Distributions (BGPD) type I is

$$G(x, y) = \exp\{-V(x, y)\}, x > 0, y > 0 \quad (21)$$

The specified models of BGPD type I have nine models. The definition and equation of each model is given in bivariate block maxima.

Finally, the third objective adopts the extreme value copula to find the dependence structure between the return on palm oil future price in the future markets.

2.6 Extreme Value Copulas

Copulas have become the attention multivariate modeling in various fields. A copula is a function that links together univariate distribution functions to from a multivariate distribution function (Patton, 2007). The relevance of copulas stems from a famous result by Sklar (1959) (cited in Segers, 2005). For simplicity, we confined it to the bivariate case. Let X and Y be the stochastic behavior of two random variables with respective marginal cdf's F(x) and G(y) is appropriately described with joint distribution function

$$H(x,y) = P(X \le x, Y \le y) \quad (22)$$

and marginal distribution functions

$$F(x) = P(X \le x), G(y) = P(Y \le y) \quad (23)$$

Since F(x) and G(y) are uniformly distributed between 0 and 1, then the joint distribution function C on $[0,1]^2$ for all (x,y) ϵ R^2 such that:

$$H(x,y) = C(F(x), G(y)) \quad (24)$$

where C is called the copula associated with X and Y which couples the joint distribution H with it margins. Equation (20) is equivalent to $H(F^{-1}(u),G^{-1}(v))$ =

C(u,v) as a consequence of the Sklar's Theorem, where u = F(x), v= G(y) are marginal distributions of X,Y. The implication of the Sklar's Theorem is that, after standardizing the effects of margins, the dependence between X and Y is fully described by the copula (Lu, et al, 2008). A comprehensive overview of the copulas properties can be referred to the work by Nelsen (1999). This thesis combines the copula construction with the extreme value theory.

The extreme value copula family is used to represent the Multivariate Extreme Value Distribution (MEVD) by the uniformly distributed margins. Consider a bivariate sample (X_i,Y_i), i=1,….,n. Denote component-wise maxima by $M_n = \max(X_1,\ldots,X_n)$ and $N_n = \max(Y_1,\ldots,Y_n)$. The object of interest is the vector of component-wise block maxima: $M_c = (M_n, N_n)'$. The bivariate extreme distribution H can be connected by an extreme value copula (EV copula) C_o: (Segers, 2005)

$$H(x,y) = C_o(F(x;\mu_1,\sigma_1,\xi_1), G(y;\mu_2,\sigma_2,\xi_2)) \qquad (25)$$

Where $\mu_i, \sigma_i \xi_i$ are GEV parameters and F(x) and G(y) are GEV margin. By Sklar's Theorem, the unique copula C_o of H is given by

$$C_o(u^t, v^t) = C_o^t(u,v), t > 0 \qquad (26)$$

The EV copula has more family. In this thesis, the two family applied are Gumbel and HuslerRiess. (Cited in Lu et al., 2008)

Gumbel copula:

$$C(u,v) = \exp(-[(-\ln u)^r + (-\ln v)^r]^{\frac{1}{r}}) \qquad (27)$$

The independence copula is obtained in the limit as r = 1, and complete dependence is obtained in the limit as r = ∞.

HuslerReiss copula:

$$C(u,v) = \exp\left\{ -\tilde{u}\,\Phi\left(\frac{1}{r} + \frac{1}{2} r \ln\left(\frac{\tilde{u}}{\tilde{v}}\right)\right) - \tilde{v}\,\Phi\left(\frac{1}{r} + \frac{1}{2} r \ln\left(\frac{\tilde{v}}{\tilde{u}}\right)\right) \right\} \qquad (28)$$

Where $\tilde{u} = -\ln u, \tilde{v} = -\ln v$ and Φ is the standardized normal distribution. The independence copula is obtained in the limit as r = 0, and complete dependence is obtained in the limit as r = ∞. For the estimation of copulas parameters, we used Exact

Maximum Likelihood method (EML): the parameters for margins and copula are estimated simultaneously, see Yan (2007) for details.