MATEMATIKA, 2018, Volume 34, Number 1, 113–123

Extreme Modelling of Missing *Acacia Mangium* Height Data

[1]M. B. Adam*, [2]N. Norazman and [3]M. R. Mohamad Kasim

[1]Institute of Mathematical Research, Universiti Putra Malaysia, Malaysia

[2]KPJ Healthcare University College, Malaysia

[3]Faculty of Forestry, Universiti Putra Malaysia, Malaysia

*Corresponding author: pmbakri@gmail.com

Article history
Received: 14 November 2017
Received in revised form: 27 November 2017
Accepted: 4 December 2017
Published on line: 1 June 2018

Abstract Logging activity is one of the most important activities for tropical countries including Malaysia, as it produces quality trees for papers. One of the important tree species is the *Acacia Mangium* which it produces a soft tree for papermaking enterprises. The papers are exported to Europe and countries which have high demand for paper due to the rapid development of the printing industry. Thus we analyzed the height for individual trees. We investigate the maximum height of the trees from 1990 to 2006 and we fit the data using extreme value model. Some of the data are missing and three imputation methods we used to solve this problem.

Keywords *Acacia Mangium* Height; Extreme Value Theory; Imputation.

Mathematics Subject Classification 62**P**12

1 Introduction

One of the most important disciplines within statistics in dealing with extreme values over the last 50 years is Extreme Value Theory, EVT. The EVT method is unique in describing rare events more than the (coverage) events. Usually, extreme value analysis requires estimation of the probability of events that are more extreme than any other observation [1]. EVT deals with statistical problems concerning of heavy tail distribution [2] and the existing data is used in the estimations of design parameters which lead to the construction of a specified probability [3]. In pursuing national development and improving living standards of the people, economic activities and development projects in a country often ignore the environmental issues. Forest is one of the victims as a result of the implementation of these developments. A lot of research has been conducted to deal with this problem and one of the approaches is using the EVT method. This method provides a draft in estimating the anticipated feature in tree growth pattern using historical data. Extreme value theory originated from Fisher and Tippet described the behaviour of the maximum of independent and identically distributed (i.i.d.) random variables in 1928. Hydrologists, environmentalist, scientists and statisticians use extreme value concept

to analyze varieties of extreme data. Extreme value model is important because in some areas the interest is not only on estimating some population central characteristics (e.g. average rainfall, average temperature, average wind speed and sports modelling) but also on estimating the minimum or the maximum values for these central characteristics [2, 4–6].

2 Materials and Methods

The data used in this paper is obtained from *Acacia Mangium* plantation at Segaliud Lokan Project, Sabah. The data consist of 20 permanent sample plots covering five types of random-level range deployed in four blocks at each level [7]. This data is measured in block with spacing for every tree is 2.0 m x 2.0 m. Twenty five trees from each block are chosen randomly. All trees were observed annually from 1990 (when the tree ages either 4 or 5 years old) to 2000 and the variables of the trees recorded were their height (maximum height is 35m) and the circumstance (maximum diameter is 35cm) of each tree whether they are alive or dead [8]. Only maximum height data are available for further analysis. The issue of missing data had been observed right from the beginning of the project due to the dead trees and further deteriorated where many more trees died after 2000. Furthermore there is another case of missing tree monitoring record which is in 1998. The challenge in this research is to replace the missing data for 1998 i.e. first stage and also to replace the missing data for death tree each year from 1990 to 2000 i.e. second stage. Imputation methods have been explored by researchers to replace the missing value [9, 10].

There are three classes of the extreme value distribution mentioned in Extremal Types Theorem [1] for types I, II and III that can be widely known as the Gumbel, Fréchet and Weibull families. The Gumbel family when shape parameter is equal to zero ($\xi = 0$), the Fréchet family when the shape parameter is greater than zero ($\xi > 0$) and the Weibull family when the shape parameter is less than zero ($\xi < 0$) [1].

- Gumbel

$$G(z) = \exp\left\{-\exp\left[-\left(\frac{z-\mu}{\sigma}\right)\right]\right\}.$$

- Fréchet if $\xi = \alpha^{-1}$ with $\alpha > 0$

$$G(z) = \begin{cases} 0, & \text{z}\leq \mu, \\ \exp\left\{-\left[\frac{z-\mu}{\sigma}\right]^{-\alpha}\right\}, & \text{z}> \mu. \end{cases}$$

- Weibull if $\xi = -\alpha^{-1}$ with $\alpha > 0$

$$G(z) = \begin{cases} \exp\left\{-\left[-\frac{z-\mu}{\sigma}\right]^{\alpha}\right\}, & \text{z}< \mu, \\ 1, & \text{z}\geq \mu. \end{cases}$$

By reformulation of all three families, we can get a single family of models that is the Generalized Extreme Value (GEV) as in Equation 1. The distribution of the maxima would

converge to a member of the Generalized Extreme Value (GEV) family of distributions if it is under the same affine transformation converges.

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

defined on $\{z : 1+\xi(z-\mu)/\sigma > 0\}$, where $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$. Here μ is location parameter, σ is a scale parameter and ξ is a shape parameter.

2.1 Maximum Likelihood Extreme

Maximum likelihood estimation method (MLE) is a popular statistical method for estimating the values of parameters [2, 11]. By using this method, the unknown parameters of a model are inferred on the basis of historical data and it is also unique for its adaptability to model change. Although the model is modified and the estimating equations are changed, the underlying methodology remains unchanged.

Let $z_1, \ldots, z_m$ be independently and identically distributed, i.i.d. variables having the GEV distribution, when $\xi \neq 0$ the log-likelihood for the GEV parameters is

$$\begin{aligned} L(\mu,\sigma,\xi) = &- m\log\sigma - \left(1+\frac{1}{\xi}\right)\sum_{i=1}^{m}\log\left[1+\xi\left(\frac{z_i-\mu}{\sigma}\right)\right] \\ &- \sum_{i=1}^{m}\left[1+\xi\left(\frac{z_i-\mu}{\sigma}\right)\right]^{-1/\xi} \end{aligned} \tag{2}$$

where

$$1+\xi\left(\frac{z_i-\mu}{\sigma}\right) > 0, \quad \text{for} \quad i = 1, \ldots, m. \tag{3}$$

The maximum likelihood estimate with respect to the entire GEV family will take a lead when maximization of (2) with respect to the parameter vector (μ, σ, ξ). For any given dataset, the maximization is straightforward but there is no analytical solution using standard numerical optimization algorithms.The likelihood will become zero and the likelihood equals to negative infinity when the parameter combinations in (3) is violated when at least one of the observed data falls beyond end-point of the distribution [1].

In this study, the aim is to model the data within each year and also to model the growth of maximum height across the years from 1990 to 2000. The focus parameter is μ as it will give the everage height of the Acacia Mangium trees.

2.2 Imputation

Missing data can lead to highly inaccurate results in certain cases. In addition, there are various types of imputation methods which can lead to accurate results [10], [12], [13]. The issue of missing data had been observed right from the beginning of the project due to the dead trees and further deteriorated where many more trees died after 2000. Furthermore there is another case of missing tree monitoring record which is in 1998. Many Imputation methods

have been explored by researchers to deal with the missing value [9]. In this work we apply Mean Imputation, Maximum Imputation, Average Mean Imputation and Random Imputation to solve this problem. Two stages of the imputation have been used:

1. For the first stage, some of the missing data in 1998 are imputed using Mean Imputation. This method is applied to the trees that has no missing height measures in 1997 and 1999. The missing 1998 height measure for each tree is replaced by the mean of its observed 1997 and 1999 height values. See Figure 1. By taking the mean from the existing pair data before and after the missing year, we still can maintain the trend characteristic of the data. The mean of two value of data also similar to the median of two value of data. Considering only the available data in 1997 and in 1999 is used to replace data in 1998 lead to the nearest, accurate and logical value of the missing data.

2. In the second stage, all the other missing data are imputed using the three imputation methods, Maximum Imputation, Average Mean Imputation and Random Imputation. This stage is implemented to get a set of all data across years in order to obtain the GEV parameter estimates within year from 1990 to 2000. This stage will provid and fill all the missing data with the imputated values and give "a complete set" of *Acacia Mangium* height data for further analysis.

2.2.1 Maximum Imputation

In maximum imputation, the highest observed height value of the trees for each year is used to impute all the missing data. The results is shown in Figure 2. The resulting completed data set is then used to estimate the EVT model parameters which is location, μ, scale σ, and shape, ξ using MLE.

2.2.2 Average Mean Imputation

In this method, we impute the missing data by the average of the observed yearly data The result is shown in Figure 3. When $\{t_{n_j}\}$ is the remaining and available data set of that year then

$$x_j = \frac{\{t_1, t_2, \ldots, t_{n_j}\}}{n_j}$$

where x_j is the average of the data and n_j is the number of observations in that j year respectively. Remember that for $n_j = 2$, the average is equal to medium value. The resulting completed data set is then used to estimate the model parameters stated in Section 2.3.

2.2.3 Random Imputation

With this method, we impute the missing value in a particular year chosen at random from a uniform distribution

$$x \sim \text{Uniform}\,(a_x, b_x)$$

where a_x is the observed minimum value and b_x is the observed maximum value in that particular year.

The resulting completed data set is shown in Figure 4. Similarly the EVT parameters which are location μ, scale σ and shape ξ are estimated from this completed data set.

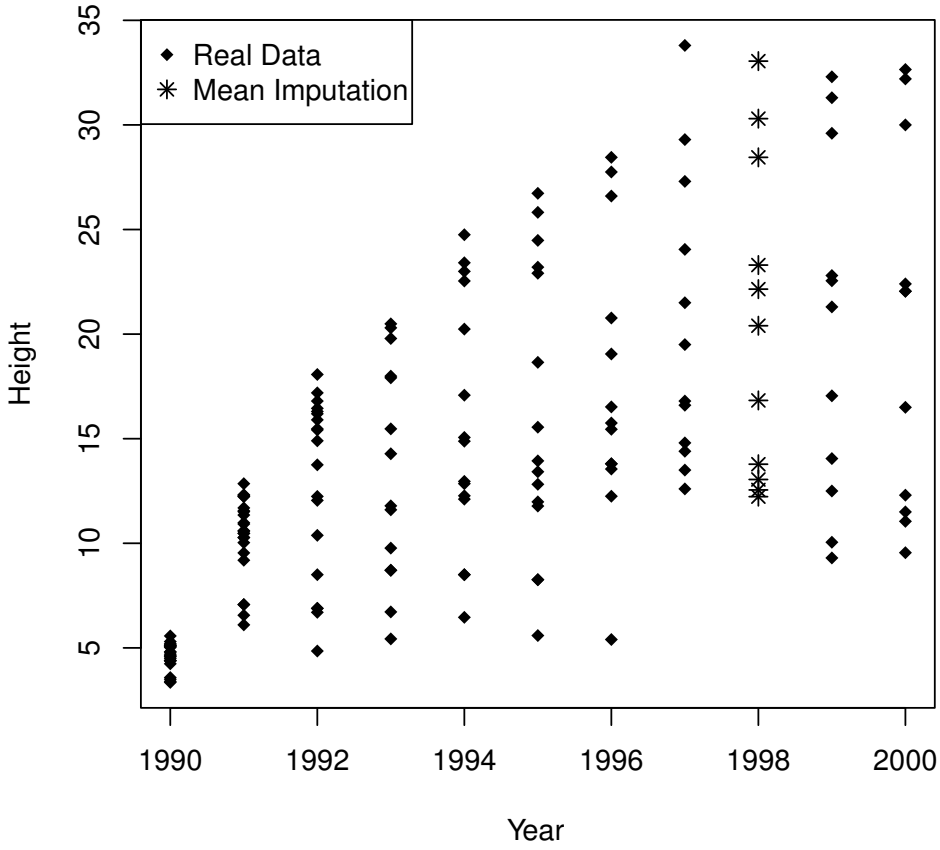

Figure 1: Completed Data using Mean Imputation Method for 1998 Missing Data in the First Stage

3 Results

Tables 1, Table 2, Table 3 and Table 4 show the years in the first column, location, scale, shape parameters in the second, third and fourth column repectively and the last column is the number of trees that are still alive at the corresponding years. The number of trees that are alive is denotated by n^l. The Mean Imputation for the first stage and the Random, Average Mean and Maximum Imputations have been carried out for the height *Acacia Mangium* height data from 1990 until 2000 for the second stage.

Table 1, Table 2 and Table 3 show the location parameter estimate values increase as the years increases and it is also proportional to the dead trees. Results for the shape estimates are not stable as its consist of regular ($\xi > -0.5$) and non-regular ($\xi < -0.5$)) also the changing of the sign of ξ value of estimates. For example in 1997 (0.376) and 1998 (0.607) were positive values instead of negative values in other years. For standard error in 1998, it gives the biggest standard error, $\text{se}(\mu) = 2.7$, $\text{se}(\sigma) = 5.67$ and $\text{se}(\xi) = 6.296$ compared to other years.

Table 2 shows an increasing then decreasing with small differences for the location estimates

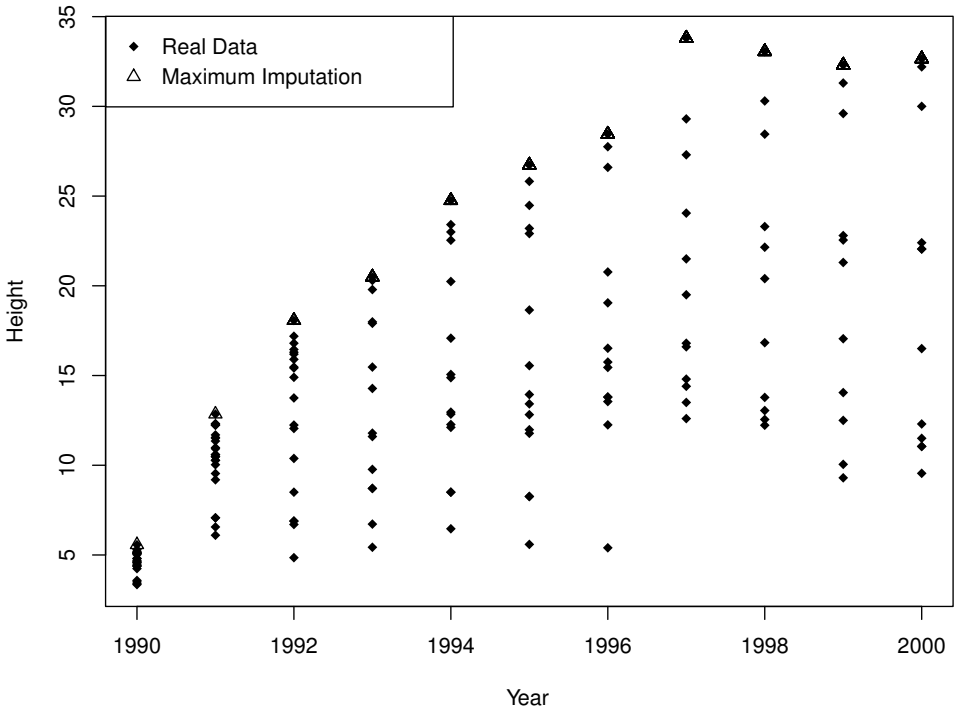

Figure 2: Completed Data Set using Maximum Imputation Method from 1990 to 2000

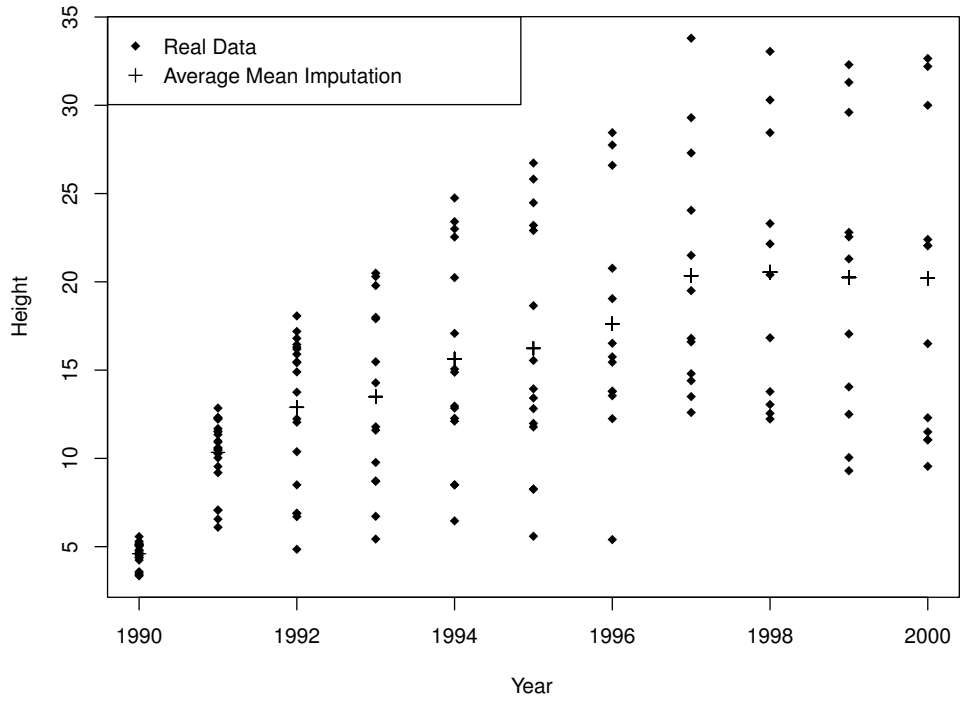

Figure 3: Completed Data using Average Mean Imputation Method from 1990 to 2000

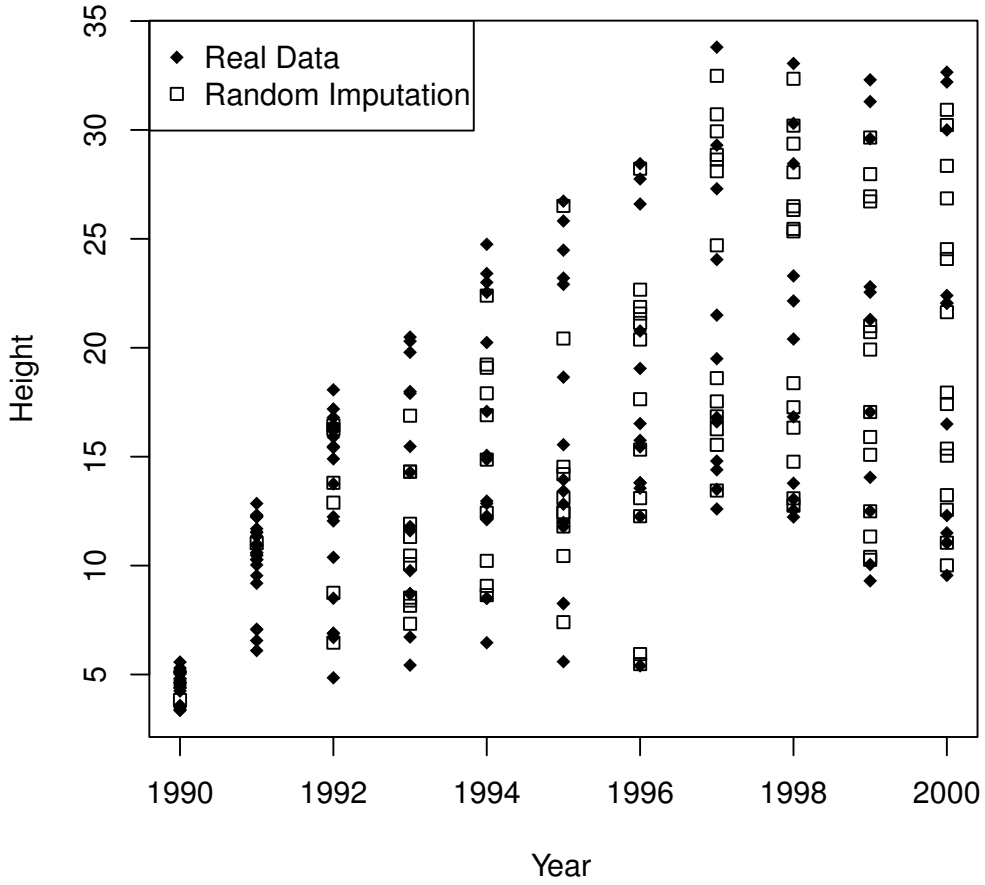

Figure 4: The Completed Data Set using Random Imputation Method from 1990 to 2000

Table 1: Parameters Estimates using GEV with the Data Height

Year	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	n^l
1990	4.5 (0.1)	0.66 (0.11)	−0.566 (0.146)	24
1991	10.1 (0.1)	2.05 (0.39)	−0.726 (0.460)	24
1992	12.6 (0.1)	4.59 (1.06)	−0.819 (1.109)	20
1993	13.8 (0.2)	4.57 (1.06)	−0.133 (1.109)	14
1994	13.8 (0.4)	5.77 (1.75)	−0.360 (2.000)	15
1995	14.0 (0.4)	6.62 (2.13)	−0.331 (2.396)	15
1996	15.2 (0.3)	6.20 (1.44)	−0.246 (1.975)	13
1997	16.4 (0.5)	4.06 (1.40)	0.376 (1.565)	12
1998	15.4 (2.7)	4.26 (5.67)	0.607 (6.296)	11
1999	17.8 (0.6)	8.11 (3.19)	−0.382 (3.478)	11
2000	16.9 (0.7)	7.56 (3.52)	−0.190 (3.523)	11

Table 2: GEV Parameters Estimates using Average Mean Imputation Method for Height Data

Year	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	n^l
1990	4.5 (0.1)	0.65 (0.11)	−0.551 (0.146)	24
1991	10.1 (0.1)	2.00 (0.39)	−0.704 (0.450)	24
1992	12.3 (0.2)	3.99 (0.78)	−0.671 (0.903)	20
1993	12.2 (0.2)	3.83 (0.64)	−0.323 (0.895)	14
1994	14.0 (0.1)	4.34 (0.71)	−0.251 (1.008)	15
1995	14.3 (0.2)	5.01 (0.83)	−0.240 (1.169)	15
1996	15.9 (0.1)	4.62 (0.68)	−0.226 (1.040)	13
1997	18.4 (0.1)	3.89 (0.62)	−0.076 (0.891)	12
1998	18.6 (0.1)	4.43(0.66)	−0.156(0.973)	11
1999	18.3 (0.1)	5.21 (0.80)	−0.232 (1.187)	11
2000	18.1 (0.1)	5.28 (0.82)	−0.203 (1.208)	11

Table 3: GEV Parameters Estimates using Random Imputation Method for Height Data

Year	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	n^l
1990	4.4 (0.1)	0.67 (0.11)	−0.550 (0.145)	24
1991	10.1 (0.1)	2.00 (0.37)	−0.714 (0.430)	24
1992	12.2 (0.1)	4.60 (0.93)	−0.767 (1.013)	20
1993	10.6 (0.3)	3.87 (0.76)	−0.103 (0.985)	14
1994	14.1 (0.2)	5.35 (1.05)	−0.373 (1.303)	15
1995	12.8 (0.2)	5.23 (0.96)	−0.051 (1.276)	15
1996	15.6 (0.2)	7.00 (1.28)	−0.419 (1.641)	13
1997	18.6 (0.5)	5.59 (1.65)	−0.468 (1.769)	12
1998	19.9 (0.3)	7.38 (1.53)	−0.468 (1.769)	11
1999	16.5 (0.3)	6.69 (1.57)	−0.173 (1.879)	11
2000	17.1 (0.3)	6.76 (1.50)	−0.183 (1.833)	11

in the years of 1990-2000. In 1994, the number of trees that are alive gained one tree from 14 to 15 trees. All the shape estimates are negative values with small the standard errors values. For example, in all years, the value of average location estimates is only from 0.1 until 0.2.

Table 2 and Table 3 present quite similar estimates for all parameters, but in some years the value is not very stable where it decreases slightly and then increased again in the following year as in 1992 (12.2) and 1993 (10.6) where it decreased but increased again in the next year, 1994 (14.1) for Random Imputation in Table 3. But in 1999 and 2000 they continue to decrease. For the shape estimates in Table 3, the values are negative which are similar to Table 2, 1997 (−0.017) and 1998 (−0.468).

For Maximum Imputation method, we can see the result in Table 4. Only in 1990, 1991 and 1996 that give the estimate value for all three parameters. Both parameters estimates, location and scales parameter increases as years increases. The shape parameter estimates give negative values even other years cannot give parameter estimates.

Table 4: GEV Parameters Estimates using Maximum Imputation Method for Height Data (the dashed is because the process of optimization is not happen. Convergency is not happening when maximizing the GEV likelihood function)

Year	μ	σ	ξ	n^l
1990	4.5	0.69	−0.610	23
1991	10.3	2.16	−0.827	23
1992	-	-	-	19
1993	-	-	-	13
1994	-	-	-	14
1995	-	-	-	14
1996	25.0	4.24	−1.248	12
1997	-	-	-	11
1998	-	-	-	10
1999	-	-	-	10
2000	-	-	-	10

Table 5: The Approximate Total Production of Tree, $\mu \times n^l \times$ diameter, $\times 100$ cm^3

Year	Diameter (cm)	First Stage	Second Stage Random	Second Stage Average Mean
1990	5	540	528	540
1991	14	3394	3394	3394
1992	23	5842	5612	5658
1993	26	5023	3858	4441
1994	29	6003	6134	6090
1995	30	6300	5760	6435
1996	32	6323	6490	6614
1997	33	6494	7366	7286
1998	34	5760	**7443**	6956
1999	35	6853	6353	7035
2000	36	6692	6772	7168

3.1 Production Prediction

Hedge et al. [8] give the rough estimate of the diameter growth for *Acacia Mangium* i.e. Max 5cm for the first 5 years, approximately 9.4 cm after 2 years and declining after 7 to 8 years. In this study, the value of diameter growths that have been considered from 1990 to 2000 are 5 cm, 14 cm, 23 cm, 26 cm, 29 cm, 30 cm, 32 cm, 33 cm, 34 cm, 35 cm and 36 cm. The Observed Maximum Value Imputation has been excluded because it not suitable for modelling maximum height data. Table 5 calculates the rough volume for the *Acacia Mangium* trees.

4 Discussion

This research proposed a new method for imputing missing extreme data. Three methods have been proposed which is Observed Maximum Value Imputation, Average Mean Imputation and Random Imputation. The results show there is an increasing pattern in location estimation for each model by years as the number of dead trees increased. Among all methods tested, we found that the method Average Mean Imputation method give the lowest root mean square error (0.339) compared to the other methods. For location estimators, standard error for Average Mean Imputation in Table 2 gives the lowest range (0.1-0.2) compared to the other models, followed by Random Imputation in Table 3 (0.1-0.5) and first stage imputation in Table 1 (0.1-2.7). As for the shape parameters, the lowest is Average Mean Imputation (0.1-1.2), Random Imputation (0.1-1.8) and the highest in the first stage imputation (0.1-6.2).

Table 4 shows that Maximum Imputation method failed to give the values of parameter estimation as it is not suitable for EVT because of Maximum Imputation by nature using the maximum value, creating two peaks of local maximum in either histogram or density plot (not shown). The feature to capture the extreme pheomena is not available in the data i.e. the data is non degenerated and not full fill the max-stable properties of extreme distribution.

The 1998 replacement data from the first stage mean imputation shows high standard errors for all parameters, se(μ) = 2.7, se(σ) = 5.67 and se(ξ) = 6.296. From the replacement for all missing data from second stages, the Average Mean Imputation give se(μ) = 0.1, se(σ) = 0.66 and se(ξ) = 0.973 and Random Imputation, se(μ) = 0.3, se(σ) = 1.53 and se(ξ) = 1.769 i.e. lower standard errors than first stage imputation. By implementing the second stages of imputations better models have been fitted but for the completed data size.

In term of the productivity of producing log, in general *Acacia Mangium* tree can be cut after 1994 where more than 6000 cm^3 being produced, i.e. after the tree age of more than nine years old. The production is more than 7000 cm^3 for 1997 and 1998 for Random Imputation method and 1997, 1999 and 2000 for the Average Mean Imputation method where the later method gave the best fitting model for the maximum height data.

5 Conclusion

The imputation methods can be one of the statistical solution in solving the missing extreme *Acacia Mangium* height data. Ignoring the missing data will lead to a less data and difficulty in analysing *Acacia Mangium* height data.The proposed of two stages Average Mean Method gives the best, nearest and accurate model compared to the other two methods in the research. In term of harvesting time of getting the log, the recommend time is after the age of tree is more than 10 years old.

Acknowledgements

The authors would like to thanks the Forest Research Centre, Sepilok, Sabah, East Malaysia for *Acacia Mangium* data. Authors also would like to thanks all the referees for the suggestions and recommendations for this article.

References

[1] Coles, S. *An Introduction to Statistical Modelling of Extreme Values.* New York: Springer-Verlag. 2001.

[2] Adam, M. B. *Extreme Value Modelling of Sports Data.* PhD Thesis, Lancaster University, UK. 2007.

[3] Xiuyun, S. Extreme value modeling of hydrological floods: A case study. *Statistics Canada.* 2005. 31(2): 139–149.

[4] Abidin, Z., Adam, M. and Midi, H. The goodness-of-fit test for gumbel distribution: A comparative study. *MATEMATIKA*. 2012. 28(1): 35–48.

[5] Adam, M. B. and Tawn, J. A. Modelling record times in sport with extreme value methods. *Malaysian Journal of Mathematical Sciences*. 2016. 10: 1–21.

[6] Adam, M. B. and Tawn, J. A. Bivariate extreme analysis of olympic swimming data. *Journal of Statistical Theory and Practice*. 2012. 6: 510–523.

[7] Kamziah, A., Kimber, A. and Lapongan, J. A parametric model for the interval censored survival times of acacia mangium plantation in a spacing trial. *Applied Statistics*. 2006. 33: 1067–1074.

[8] Hegde, M., Palanisamy, K. and Yi, J. S. *Acacia Mangium Willd.* - a fast growing tree for tropical plantation. *Journal of Forest Science*. 2013. 29(1): 1–14.

[9] Rebecca, R. and Roderick, J. A review of hot deck imputation for survey non-respone. *International Statistical Review*. 2010. 78(1): 40–64.

[10] Rubin, D. *Multiple Imputation for Nonresponse in Surveys.* New York:Wiley. 1987.

[11] Amin, N. M., Adam, M. and Aris, A. Extreme value analysis for modeling high pm10 level in johor bahru. *Jurnal Teknologi*. 2015. 76(1): 171–179.

[12] Wayne, A. and Kim, J. Hot deck imputation for the response model. *Statistics Canada.* 2005. 31(2): 139–149.

[13] Krishnamoorthy, K., Mallick, A. and Mathew, T. *Model-based imputation approach for data analysis in the presence of non-detects.* Oxford: Oxford University Press. 2009.