APPLYING LINEAR MIXED MODEL (LMM) TO ANALYZE FORESTRY DATA, CHECKING AUTOCORRELATION AND RANDOM EFFECTS, USING R

Bui Manh Hung[1]**, Bui The Doi**[2]

[1,2] *Vietnam National University of Forestry*

SUMMARY

Currently, R is showing its strengths and benefits in data analysis in general and forestry data in particular. R can perform new, difficult and complex statistical analyses such as linear mixed model, replicated point patter analysis, etc. In the forestry data analysis, checking independence between samples and random effects has not been done so far by scientists. This is a really difficult problem in forestry data analysis, because it is an important basis for choosing analytical tools later on. However, the linear mixed model (LMM) application with the support of R language, this problem has been resolved. The LMM results for the data collected from 20 plots in Kon Ka Kinh national park indicate that heteroscedasticity is occurring between the two forest types. This study does not find a significant influence of autocorrelation on the observed data. In other words, the samples are completely independent. The difference in diameter between the secondary forest and the old-growth forest is significant. Normal distribution assumption was also tested and the hypothesis is accepted. Random effects are not important in this study.

Keywords: Autocorrelation, linear mixed model, old-growth forest, R language, random effect, secondary forest.

I. INTRODUCTION

Data analysis is essential to make later decisions. In forest science, it becomes more important, because with long business cycles. Therefore, that requires more careful decisions or more precise data analysis results.

Nowadays, there are many new and modern statistical tools in order to analyze data. Linear mixed effect models are one of them. A linear mixed model (LMM) is an extension of a linear regression model. LMM is used for data which are collected and summarized in groups. Mixed models analyze the relationship between a response variable and independent variables (Mathworks, 2016). The response variable is continuous, and independent variables can be both continuous and discrete (West et al., 2015). Another different point from linear regression models is that independent variables can be categorical. Therefore, LMM can be used to compare groups in order to understand the difference between them.

The grouped data suitable for analysis by LMM includes longitudinal data, repeated measures data, multilevel data, nested designs and block designs (Pinheiro and Bates, 2000; Faraway, 2006; Wagner, 2014). There are two main parts in a mixed-effect model. They are the fixed factor and the random factor. It is critical to distinguish them in the context of LMM. There are some clues that can be used to distinguish and apply them accurately, especially for the data from nested designs. Fixed variables are categorical or continuous. In other words, fixed-effect terms are usually the conventional linear regression or analysis of variance parts (Faraway, 2006; Winter, 2013; Wagner, 2014; Mathworks, 2016). They are both factor variables (like factors in analysis of variance) with fixed values and continuous variables (like independent variables in regression models). They are the object of interest of analysts. Fixed variables are a part of research hypotheses. They are used directly to analyze the relationship with the response variable or check the differences between groups. In contrast, a random factor is a classification variable. It is a discrete variable. It is randomly sampled from a population of levels being studied (Faraway, 2006). Random variables are always higher than the analyzed individuals by at least one level, especially for nested designs. Random variables are used to check the non-randomness of the data. Therefore, it must include all the analysis units (Hung, 2016).

LMM is very strong tool in checking random effects and autocorrelation or independence between samples. Although, these are essential conditions to choose and apply relevant tools, they have not been checked before last some years, especially in forest science fields in Vietnam. That really is a problem in the past, but now it is solved by using LMM.

In recent years, R is showing its benefits and strong points in data analysis. R is a language and environment for statistical computing and graphics. R is similar to the S language and environment which was developed by John Chambers and colleagues. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, LMM, point pattern analysis…) and graphical techniques (R-project, 2016). It is powerful. R can handle complex and large data. There are many packages, so R can perform very complex analysis (Smart, 2014). R can conduct from basic to very difficult statistical practices.

For LMM, up to now, there are some packages which can be implemented to conduct LMM analysis such as lme4 and nlme. In the case of this study, the nlme package was used. Therefore, a linear mixed-effect model consists of some main components: a response variable, fixed variables, data file name, random variables and used methods (Pinheiro et al., 2016).

The above explanation has shown the urgency and necessity for applying LMM, with the support of R. This paper will present how to implement LMM, especially for nested designs and how to call commands in R in order to perform these analyses.

II. METHODOLOGY

2.1. Data collection methods

Data were collected from 20 plots in Kon Ka Kinh national park, Kon Tum province, Central Highlands of Vietnam. The location of the park is presented in the following figure. In 20 plots, there are 10 plots secondary forests (Type IIb) and 10 plots of old-growth forests (Type IV).

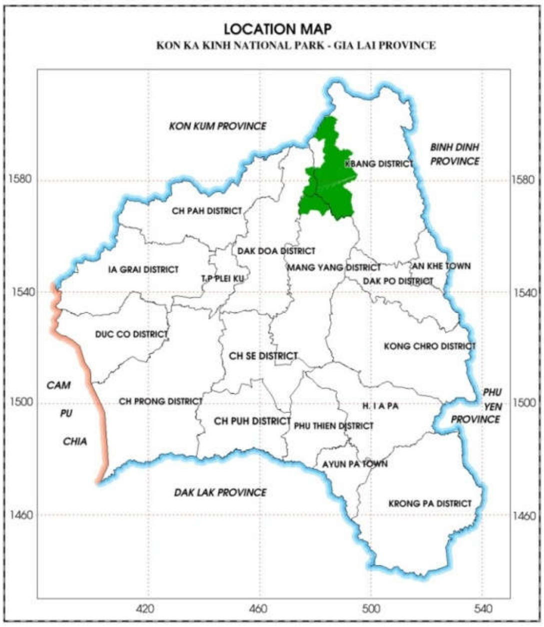


Fig. 01. Kon Ka Kinh national park location (Hung, 2016)

Stratified random sampling was applied to select plot locations (Fig. 02) (Shiver and Borders, 1996), because the forest resource is not homogeneous (Shiver and Borders, 1996).

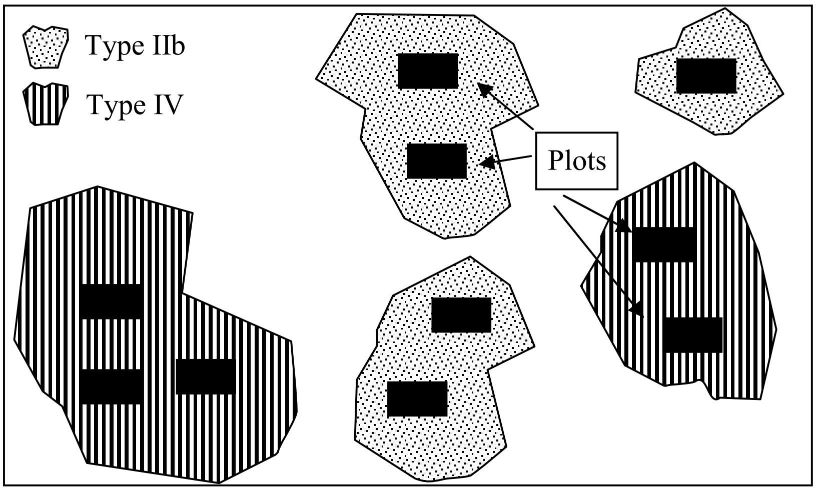


Fig. 02. Plot arrangement

2.2. Data analysis methods

2.2.1. LMM applications with this study and data arrangement

In the case of this study, the difference of the diameter of trees between two forest types (Type IIb and Type IV) needed to be checked. The averages diameter of trees were calculated in each plot. The plots are analysis units. The plots were randomly sampled in each section. The sections were also drawn randomly from the population (Kon Ka Kinh national park). The area of each section is restricted because of limited human resources, paths and living conditions in the park. The plots in the same section are often closer to others, compared to the plots in different sections. Therefore, spatial autocorrelation or non-randomness could be present between plots in the same section.

The model for diameter variable was:

$$DBH = (\beta_0) + (\beta_{1,i})\,Type + (b_j)Section + \varepsilon_{i,j}$$

Where:

β_0 = the intercept;

β_1 = a parameter estimated for fixed effects (Type);

b = a parameter for the random effect (Section);

ε_{ij} = error.

A model for height variable is very similar, but instead using a DBH by H variable.

For the next steps of analysis by using LMM, the following commands were run first to attach data files, make the Type variable a fixed factor, draw a box chart in order to gain some basic information about differences between two types, run the nlme package and a mixed model. The following commands were used for the DBH variable. The data file name is “data”.

```
attach(data)
Type = factor(Type)
boxplot(DBH ~ Type, notch = FALSE,  data = data, xlab = "Type",
        ylab = "Diameter at breast height",  boxwex = 0.5, lty=1,
col=c("green","red"))
library(nlme)
model.lme = lme(DBH ~ Type, data = data,
                 random = ~ 1|Section,
                 method ="REML")
```

2.2.2. Homoscedasticity checking

LMM uses parametric statistics, meaning that there are some assumptions that need to be satisfied. These are the homogeneity of variance, independence and normality of residuals (Pinheiro and Bates, 2000; Zuur et al., 2009).

Firstly, variance homogeneity of residuals should be examined. It is the most important assumption of linear regression and additive modeling (Zuur et al., 2009). A raw residual is the difference between an observed value and the predicted value of the response variable (West et al, 2015). To check variance homogeneity, a diagnostic plot is often used. That is a plot of raw residuals against fitted values. If there is no heteroscedasticity, the plot will display a random pattern and constant variability along the vertical axis, like the following figure (Gałecki and Burzykowski, 2013). In other words, the plot should show the same residual spread per stratum for some of the variables (Zuur et al., 2009). That means that the model equally explained the actual data at that point. Heteroscedasticity will lead to parameters with incorrect standard errors, and wrong F and t statistics, because they do not follow F and t distributions. As a consequence, the significance of the model and estimated parameters will be wrongly assessed.

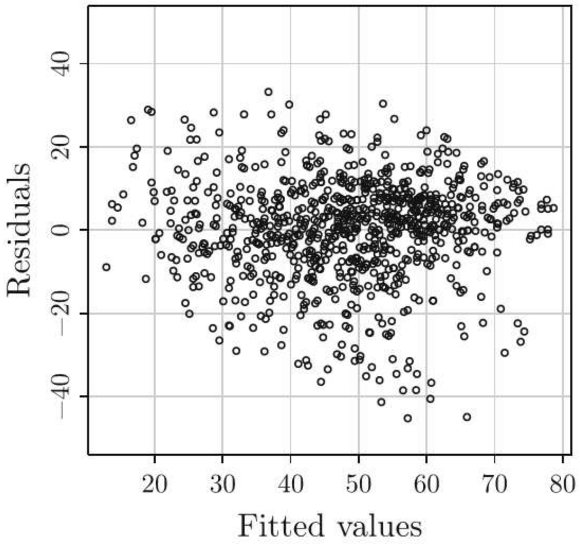


Fig. 03. A typical diagnostic plot showing no heteroscedasticity (Gałecki and Burzykowski, 2013)

To achieve that plot in R, the following command was implemented.

```
plot(model.lme, resid(., type = "n") ~
fitted(.), abline = 0, pch=19, col="red")
```

There are some ways to deal with heteroscedasticity. The first solution is data transformation, but this method should not be used, because the data will be changed and then it is not your data any more. Another better way is to use variance structure functions. In R, there are some functions to deal with heteroscedasticity such as: varExp, varIdent, varFixed and varPower functions. They will weight variances differently between groups and make variances more homogeneous. In practice, it is better to use varPower, varExp, or varConstPower, because they allow more flexibility than the varFixed (Pinheiro and Bates, 2000; Zuur et al., 2009). In this research, varExp is often applied. For example:

```
model1.lme    =    update(model.lme,
weights = varExp())
```

2.2.3. Checking autocorrelation

Independence is the second important assumption that should be satisfied when LMM is used. In this study, the data is not time-series data. Therefore, to check autocorrelation, this means investigating whether the correlation between residuals are independent or dependent on the distance between subjects (Wagner, 2014). Semi-variogram technique is used to check autocorrelation (Wagner, 2014).

Semi-variance is a half mean of the squared differences between residuals of analysis subjects with distance (d). In my case, the subjects are plots.

$$\text{Semi}-\text{variance}=\frac{1}{2}.\left(\frac{\sum_{i=1}^{n}((R_{i+d}-R_i)^2)}{n}\right)$$

In which:

R_i is residual at i;

R_{i+d} is residual at i+d;

d is distance, n is sample size.

Smaller semi-variance leads to stronger correlation of residuals and makes the semi-variogram closer to 1. Semi-variograms will provide researchers with answers to the following questions: does autocorrelation exist in the data? And from which distance does the autocorrelation not exist? The following command was called.

```
plot(Variogram(model1.lme, form = ~X+Y,
resType = "n"), smooth = TRUE, cex = 1.5)
```

2.2.4. Checking normal distribution of the residuals

Normal distribution is another assumption should be satisfied in order to use LMM. This assumption is less important than previous ones, and it is not a serious problem. The normality assumption is not needed if the sample size is large enough (Zuur et al., 2009). Taking mean values from samples and using them in LMM can meet this assumption because of the central limit theory. If the number of replicates is large, the distribution of mean values tends toward normality (Zar, 1999), regardless of the sample distribution. However, if the number of replicates is not much, normality checking should not be based on histogram of raw data. Instead, a model should be applied and after that inspect residuals (Zuur et al., 2009).

There are some ways to check and test the normality of residuals. The first method is QQplot. QQplot is the quantile-quantile plot. That is an exploratory graphical device used to check the validity of a normal distribution assumption for residuals. In this plot, the quantiles of ordered residuals are plotted against the corresponding values for the standard normal distribution. If all the scatter points are close to the reference line, we can say that the dataset follows the normal distribution (Gałecki and Burzykowski, 2013). This method is intuitive and graphical. The following command was run in R to generate QQ plot for normality checking.

```
qqnorm(model2.lme, abline = c(0,1),
pch=19, lty=1, col="red", main =
"QQplot to check normality")
```

Another way to test the normality assumption of the residual is the Shapiro-Wilk test. This test work best if the dataset size is less than 50 (Zar, 1999). In order to run the Shapiro-Wilk test in R for normality checking, the following command was applied for the best model. For example:

```
data$residual = residuals(model2.lme)
shapiro.test(data$residual)
```

2.2.5. Model selection and information summary

After establishing models or improving models in the above steps, the update command in R was run to update the model information. The ML (maximum likelihood) method was applied to compare models. REML should not be used for model comparison, especially models with different fixed effects (Faraway, 2006). For example:

```
model1a.lme     =    update(model1.lme,
method = "ML")
model2a.lme     =    update(model2.lme,
method = "ML")
```

To select the better model, anova was called in R (Pinheiro and Bates, 2000; Faraway, 2006). For instance:

```
anova(model1a.lme, model1b.lme)
```

In this research models are not nested, so the use of information criteria is a possible solution (Gałecki and Burzykowski, 2013). Akaike's information criterion (AIC) is again a basis to choose the better model. The formula of AIC is presented in detail in 4.3.1.5c section. The better model is one that has a smaller AIC (Wagenmakers and Farrell, 2004 and Osman et al., 2012) and simpler structure. The p-value will show an answer for a question: are the models significantly different?

After conducting all the above steps, the best model was selected. The model information was obtained by using commands as follows:

```
summary(model2.lme)
```

III. RESULTS AND DISCUSSION

Linear mixed effect model analysis was applied to check the diameter and height difference among two forest types. Not only fixed effects, but also random effects are checked.

3.1. Box plots for the diameter variable

Linear mixed effect models were built for analysis. Type is fixed factor and Section is random factor. Box charts for the diameter variable were generated and the results are as follows.

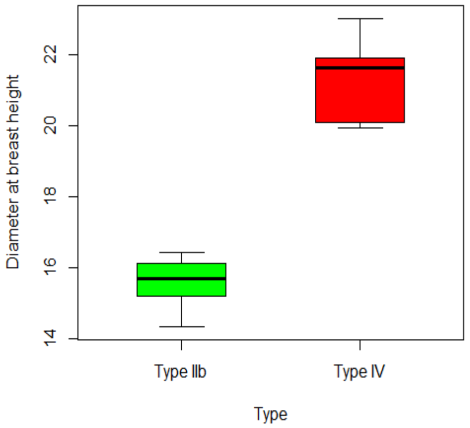


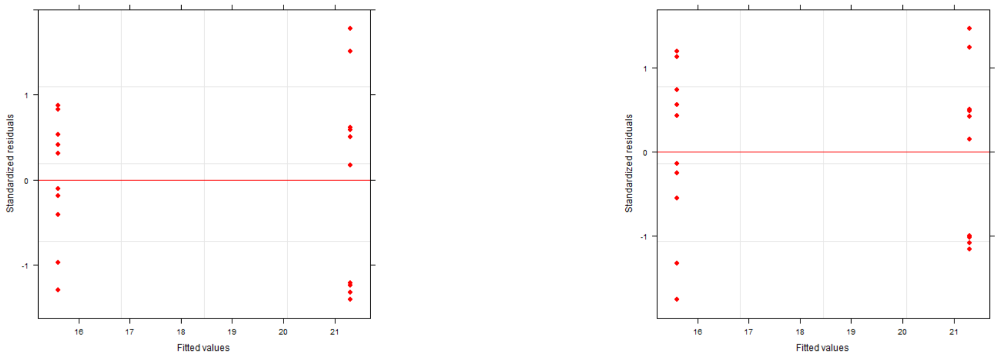
Fig. 04. Boxplots for variables

Figure 04 indicates that the diameter of Type IIb is smaller than Type IV. In addition, the variation for the variable in young forests is smaller.

3.2. Model analysis and adaptation

a. Homoscedasticity checking

Homogeneity of variances was checked.



a. For diameter before adapting

b. For diameter after adapting

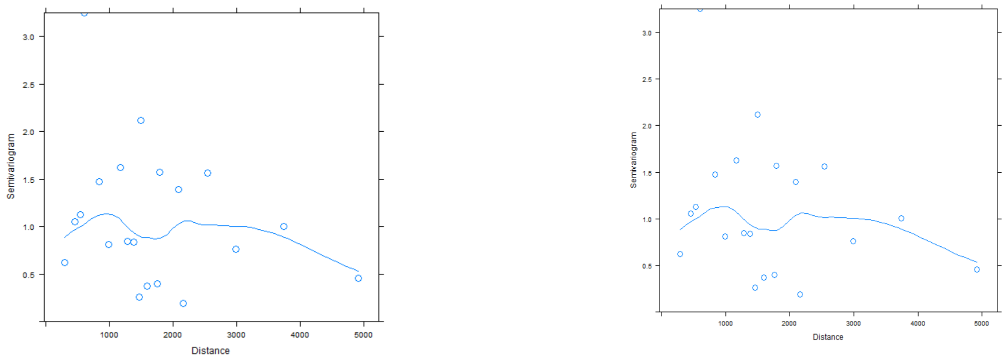
Fig. 05. Homoscedasticity adaptation

Figure 05 shows results of improvements in term of homoscedasticity by using the varExp function. The varExp function was applied for model1. These charts indicate that the improvement is not clear for the diameter. This again is proven by the results of anova so as to select a better model. The p-value is greater than 0.05. Therefore, the model is still selected for the diameter.

```
        Model df      AIC      BIC    logLik   Test  L.Ratio p-value
model1a     1  4 61.5374 65.52033 -26.76870
model1b     2  5 61.1045 66.08317 -25.55225 1 vs 2 2.432896  0.1188
```

b. Autocorrelation checking

After that, better models were used for autocorrelation checking. For the diameter variable, "Model2" is improved by using the corExp function. It is presented in the following figure.



a. For diameter before adapting

b. For diameter after adapting

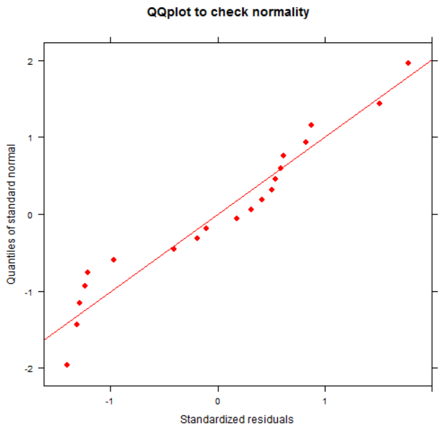
Fig. 06. Spatial autocorrelation improvement

Based on these graphs, the influence of autocorrelation is not obvious, especially for the diameter variable. Additionally, the adaptation for autocorrelation is not considerable. The following anova comparison results also prove this. The p-values are larger than 0.05. As a result, model is still selected for the diameter.

```
        Model df      AIC      BIC   logLik   Test     L.Ratio p-value
model2a     1  4 61.5374 65.52033 -26.7687
model2b     2  5 63.5374 68.51606 -26.7687 1 vs 2 5.59325e-10       1
```

c. Normality checking

Normality of residual distribution is one assumption of LMM as mentioned in the previous chapter. It was also checked by using qqplot and the Shapiro-Wilk test.



a. For diameter

Fig. 07. QQplot for normal distribution checking

The normal distribution assumption is satisfied (Figure 07). This is more obvious and convincing by the Shapiro-Wilk test results as follows, because the p-value for both is larger than 0.05.

```
     Shapiro-Wilk normality test
data:  data$residual
W = 0.9277, p-value = 0.1395
```

3.3. Model parameter estimation

Parameter estimation results of the best model for the diameter is as follows.

```
Linear mixed-effects model fit by REML
 Data: data
      AIC      BIC    logLik
  62.68532 66.24681 -27.34266

Random effects:
 Formula: ~1 | Section
        (Intercept)  Residual
StdDev: 2.546154e-05 0.9725602

Fixed effects: DBH ~ Type
              Value Std.Error DF  t-value p-value
(Intercept) 15.59883 0.3075506 16 50.71956       0
TypeType IV  5.70758 0.4349422 16 13.12262       0
 Correlation: 
           (Intr)
TypeType IV -0.707

Standardized Within-Group Residuals:
      Min        Q1       Med        Q3       Max 
-1.4050646 -1.0323011  0.2434091  0.5955569  1.7798280 

Number of Observations: 20
Number of Groups: 3
```

The results show that the random effect is not important as the standard deviation of the intercept is very small. The average diameter for Type IIb is 15.599 cm, while that of Type IV is greater by 5.708 cm. The diameter is significantly different between the two forest types.

IV. CONCLUSION

To provide more solid evidence of the difference in size of trees between two forest types, and to examine influence of autocorrelation to the collected data, linear mixed effect models were used. Assumptions were checked. Results show that heteroscedasticity is occurring. The variance of the old-growth forest is larger. This is improved significantly by using the varExp function, as can be seen in Figure 05. This study does not find a significant influence of autocorrelation on the observed data. Semivariograms show this. The normality assumption is accepted by checking model residuals. Therefore, linear mixed models are vital to compare the tree size among types. The results confirm that the tree size difference is statistically significant. The average diameter of the old-growth forest is greater than the regenerating forest by 5.71 cm. In addition, the random factor (section) is not really important to the diameter. The reasons for the effect of section are soil and side conditions. The plots in identical sections are often closer to others, so soil and side conditions are more similar. That will lead to non-randomness between plots in the same section, if the autocorrelation occurs.

The previous studies, especially in Vietnam, often use Excel and SPSS or Stata. However, currently, R is showing advantages in scientific data analysis in general, forestry data in particular. Again, the first advantage is a free program, so license cost is not the issue. Moreover, analysts can add, install and use many packages for specific targets. R is also very useful when researchers would like to repeat analysis. And finally, R can perform analysis that Excel and SPSS cannot conduct or hard to conduct, for example, linear mixed effect model or point pattern analysis.

Final point we would like to emphasize that linear mixed effect model application is really necessary and important for forestry data analysis nowadays. Numerous previous and current forestry research often use statistical tests or analysis of variance to compare groups or samples (Chapman and Chapman, 1997; Cao and Zhang, 1997; Gebrehiwot et al., 2004; Rad et al., 2009). However, the downside of these statistical tools is that they cannot check effects random factors and spatial autocorrelation among individuals or samples. Therefore, linear mixed effects models can be used to overcome the above disadvantages (Pinheiro and Bates, 2000; Wagner, 2014; West et al., 2015).

ACKNOWLEDGEMENTS

Firstly, we would like to send special thanks to Prof. Dr. Sven Wagner (Chair of Silviculture, Institute of Silviculture and Forest protection, Faculty of Environmental Sciences, Dresden University of Technology). We would like to thank him for the teaching, instructions, useful advice for applying Linear mixed model and using R. We also want to thank Ms.C. Nguyen Van Hoan (Director of Kon Ka Kinh national park) and our study team there for allowing us to set up plots and collect data in the park. Thank you very much for your support and nice time in there. We also would to send special thanks to Prof. Dr. Tran Van Chu, the Rector of Vietnam National University of Forestry and the Faculty of Forestry for giving a office to work and analyze data during that time.

REFERENCES

1. Hung, B. M. (2016). *Structure and restoration of natural secondary forests in the Central Highlands, Vietnam*. Chair of Silviculture, Institute of Silviculture and Forest protection, Faculty of Environmental Sciences, Dresden University of Technology. Doctor thesis.

2. Faraway, J. J. (2006). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*, Chapman & Ha ll/CRC, Taylor & Fancis Group, 6000 Broken Sound Parkway NW, Suite 300, Boca Raton, Florida, USA.

3. Pinheiro, J., et al. (2016). *Package 'nlme'*, The R foundation. Available from: https://cran.r-project.org/web/packages/nlme/nlme.pdf (Accessed 11 August, 2016).

4. Pinheiro, J. e. C. and D. M. Bates (2000). *Mixed-Effects Models in S and S-PLUS*, Springer-Verlag, New York, USA.

5. Shiver, B. D. and B. E. Borders (1996). *Sampling techniques for forest resources inventory*, John Wiley & Sons, Inc. Canada.

6. Wagner, S. (2014). *Linear mixed models (LMM)*, Institute of Silviculture and Forest protection, Faculty of Environmental science, Dresden University of Technology, Tharandt, Germany.

7. West, B. T., et al. (2015). *Linear Mixed Models: A practical Guide Using Statistical Software*, CRC Press, Taylor & Francis Group, 6000 Broken Sound Parkway NW, Suite 300, Boca Raton, Florida, USA.

8. Zuur, A. F., et al. (2009). *Mixed Effects Models and Extensions in Ecology with R*, Springer, USA.

ỨNG DỤNG MÔ HÌNH TUYẾN TÍNH HỖN HỢP ĐỂ PHÂN TÍCH SỐ LIỆU LÂM NGHIỆP, KIỂM TRA TỰ TƯƠNG QUAN VÀ ẢNH HƯỞNG NGẪU NHIÊN BẰNG NGÔN NGỮ R

Bùi Mạnh Hưng[1]**, Bùi Thế Đồi**[2]

[1,2]*Trường Đại học Lâm nghiệp*

TÓM TẮT

Hiện nay, R đang chứng minh được những thế mạnh và ưu điểm nổi trội của mình trong phân tích số liệu nghiên cứu nói chung và số liệu lâm nghiệp nói riêng. R có thể thực hiện được những phân tích thống kê mới, khó và phức tạp như: mô hình tuyến tính hỗn hợp, phân tích phân bố không gian cây rừng lặp… Trong phân tích số liệu lâm nghiệp, kiểm tra tính độc lập giữa các mẫu và ảnh hưởng ngẫu nhiên chưa từng được thực hiện từ trước đến nay bởi các nhà khoa học. Đây thực sự là một vấn đề khó khăn trong phân tích số liệu lâm nghiệp, bởi lẽ nó là cơ sở để lựa chọn các công cụ phân tích phù hợp sau này. Tuy nhiên, với việc ứng dụng mô hình tuyết tính hỗn hợp (LMM) với sự hỗ trợ của ngôn ngữ R, vấn đề này đã được giải quyết. Kết quả ứng dụng LMM cho số liệu được thu thập từ 20 ô tiêu chuẩn tại vườn quốc gia Kon Ka Kinh cho thấy rằng: tính bất đồng về phương sai đã tồn tại giữa hai trạng thái rừng. Ảnh hưởng của tự tương quan không rõ rệt tới biến đường kính. Hay nói cách khác các mẫu là hoàn toàn độc lập. Sự khác biệt về đường kính giữa rừng thứ sinh và rừng già là thực sự rõ rệt. Điều kiện phân bố chuẩn cũng được kiểm tra và cho thấy giả thuyết được chấp nhận. Ảnh hưởng ngẫu nhiên không đáng kể trong nghiên cứu này.

Từ khóa: Ảnh hưởng ngẫu nhiên, mô hình tuyến tính hỗn hợp, ngôn ngữ R, rừng già, rừng thứ sinh, tự tương quan.

Received	**: 07/3/2017**
Revised	**: 17/3/2017**
Accepted	**: 31/3/2017**