STATA – GIẢI PHÁP MỚI CHO PHÂN TÍCH SỐ LIỆU NGHIÊN CỨU LÂM NGHIỆP

Bùi Mạnh Hưng, Nguyễn Thị Bích Phượng

ThS. Trường Đại học Lâm nghiệp

TÓM TẮT

Stata là phần mềm thống kê được phát triển từ năm 1985. Stata là một phần mềm đã và đang thể hiện được sức mạnh vượt trội trong phân tích số liệu nghiên cứu nói chung và trong Lâm nghiệp nói riêng. Stata có thể được ứng dụng trong nhiều lĩnh vực như: Điều tra rừng, Lâm học, khoa học đất, Quản lý bảo vệ tài nguyên rừng, Chế biến lâm sản…. Nó có thể thực hiện các phân tích từ đơn biến đến đa biến, từ định tính đến định lượng. Stata có thể cung cấp nhanh chóng các thông tin về mẫu và phân bố thực nghiệm thông qua lệnh tính toán đặc trưng mẫu, lệnh vẽ biểu đồ và xây dựng các bảng biểu. Stata còn có khả năng phân tích trong việc so sánh các mẫu quan sát và tìm ra điều kiện tốt nhất cho các đại lượng lâm nghiệp. Chương trình có thể so sánh bằng các tiêu chuẩn tham số và phi tham số, từ hai mẫu đến nhiều mẫu, từ mẫu độc lập đến mẫu liên hệ. Hơn nữa, phân tích phương sai một và hai nhân tố cũng là một trong những ưu việt của Stata. Một chức năng ưu việt nữa của Stata là phân tích tương quan hồi quy. Stata có thể phân tích tuyến tính một biến đến đa biến. Chương trình cũng có thể phân tích tương quan phi tuyến với những loại hàm rất khó thực hiện như: Gompertz, Schumacher, Koller, Verhulst-Robertso.... Vì vậy, việc khai thác và sử dụng Stata trong phân tích số liệu nghiên cứu về lâm nghiệp là thực sự cần thiết và là vấn đề đáng quan tâm.

Từ khóa: ***Đặc trưng mẫu, phân tích phương sai, phân tích số liệu, thống kê, tương quan, so sánh, Stata***

I. ĐẶT VẤN ĐỀ

Stata là phần mềm phân tích thống kê bắt đầu được phát triển vào năm 1985 bởi tập đoàn Stata (Wiki, 2012). Stata là chữ viết tắt bởi cụm từ "Statistics and data" tạm dịch là "Thống kê và số liệu". Phần mềm này được sử dụng rộng rãi để phân tích số liệu nghiên cứu trong nhiều lĩnh vực khác nhau như: kinh tế, xã hội học, y tế… (Wiki, 2012). Stata đã liên tục được phát triển, hoàn thiện và nâng cấp các chức năng và trình lệnh mới từ những phiên bản đầu tiên cho đến nay là phiên bản 12.

Stata là một phần mềm mạnh trong phân tích số liệu nói chung, nhưng vẫn còn ít được ứng dụng để phân tích số liệu nghiên cứu về lâm nghiệp. Tuy nhiên, qua quá trình khai thác, ứng dụng Stata trong phân tích số liệu, nó đã chứng minh được sức mạnh trong lĩnh vực này. Stata có thể đem lại nhiều lời giải cho các vấn đề rất khó khăn đặt ra trong phân tích số liệu lâm nghiệp. Những vấn đề là số lượng và dung lượng các mẫu thường rất lớn. Ngoài ra, trong lĩnh vực lâm nghiệp có nhiều chuyên môn sâu khác nhau như: Điều tra rừng, Lâm sinh, Quản lý tài nguyên rừng, Chế biến lâm sản… Mỗi chuyên môn sâu lại có những yêu cầu riêng khi phân tích. Tuy nhiên, với hệ thống các trình lệnh phong phú, Stata có thể được ứng dụng trong các lĩnh vực khoa học này. Nó có thể được sử dụng để phân tích đơn biến đến đa biến, từ phân tích định tính đến phân tích định lượng, từ tính toán đặc trưng mẫu đến lập phân bố thực nghiệm, so sánh các mẫu quan sát, phân tích phương sai và tương quan hồi quy.

Trong khuôn khổ của bài báo này tác giả sẽ trình bày khả năng ứng dụng Stata đáp ứng yêu cầu của người sử dụng trong công tác phân tích số liệu nghiên cứu lâm nghiệp.

II. PHƯƠNG PHÁP NGHIÊN CỨU

Kế thừa các số liệu trong lĩnh vực lâm nghiệp như: Điều tra rừng, Lâm học, Chế biến lâm sản, Quản lý tài nguyên rừng… để phục vụ cho việc chạy các mô hình phân tích khác nhau sau này.

Tham khảo tài liệu, nghiên cứu các phần mềm ứng dụng, phân tích số liệu khác có liên quan, tham khảo ý kiến chuyên gia trong lĩnh vực nghiên cứu khoa học lâm nghiệp, phân tích số liệu nghiên cứu lâm nghiệp nói riêng và các lĩnh vực khác nói chung.

Xây dựng các qui trình kỹ thuật trên phần mềm Stata để thực hiện các nội dung nghiên cứu và chạy thử các quy trình lệnh với số liệu của các mô hình đã thu thập được trong công tác ngoại nghiệp.

III. KẾT QUẢ NGHIÊN CỨU

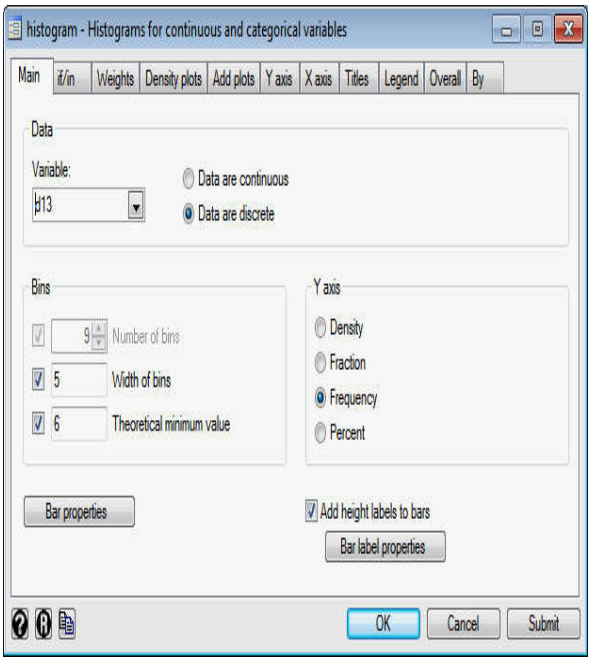
3.1. Cung cấp các thông tin về mẫu lâm nghiệp và phân bố thực nghiệm

Yêu cầu đầu tiên trong phân tích số liệu lâm nghiệp là các thông tin thống kê về các mẫu, các biến như: đường kính, chiều cao, độ ẩm, độ dày ván gỗ… và các thông tin về phân bố thực nghiệm. Những thông tin này thực sự cần thiết để các nhà nghiên cứu có cái nhìn khái quát, và những nhận định sơ bộ về các mẫu nghiên cứu. Đây cũng là cơ sở ban đầu để đề xuất các biện pháp kỹ thuật nói chung và kỹ thuật lâm sinh nói riêng, nhằm nâng cao chất lượng rừng, cải thiện hiệu quả quản lý tài nguyên rừng, hay điều chế rừng theo mục đích kinh doanh phù hợp (Nguyễn Hải Tuất và Nguyễn Trọng Bình, 2005).

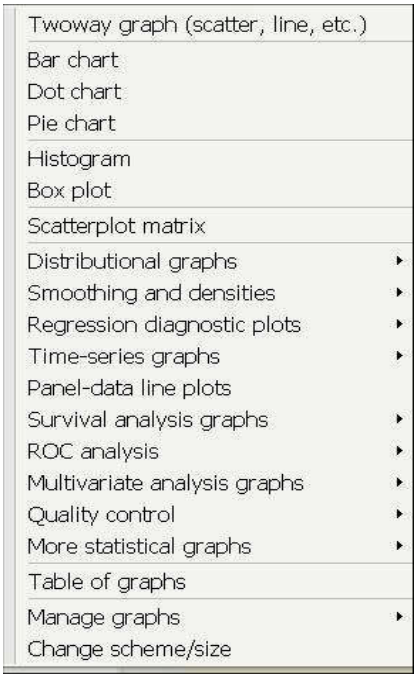
Với lệnh "Summary and descriptive statistics" (Tổng hợp và thống kê mô tả), Stata có thể nhanh chóng tính toán và xuất ra các kết quả về đặc trưng mẫu như: dung lượng mẫu, trung bình mẫu, phương sai, độ lệch, độ nhọn… Stata cung cấp cả những thông tin về vị trí, biến động và hình dạng phân bố. Stata có thể tính toán cho cả mẫu nhỏ và mẫu đã qua lập phân bố thực nghiệm.

Lệnh "Tables of summary Statistics (tables)" (Bảng thống kê) có thể dễ dàng lập các phân bố thực nghiệm cho các biến rời rạc như số cây tái sinh/ô, số sâu/cành hay số sinh viên/lớp.

Riêng với biến liên tục, Stata là một trong những phần mềm được đánh giá là ưu việt hơn cả để lập phân bố thực nghiệm. Khi sử dụng Stata chúng ta có thể dễ dàng điều chỉnh số lượng tổ được chia, cự ly tổ và các giá trị cận dưới, cận trên. Đây là một ưu điểm nổi trội của Stata, rất phù hợp với phân tích số liệu nghiên cứu lâm nghiệp.



Hình 01. ***Hộp thoại histogram để điều chỉnh số tổ, cự ly tổ để lập phân bố thực nghiệm***



Hình 02. ***Các lệnh vẽ biểu đồ trong Stata***

Lệnh "Distributional graphs" trong "Graphics" (biểu đồ) chúng ta có thể dễ dàng kiểm chứng được mức độ phù hợp của phân bố thực nghiệm theo các phân bố lý thuyết khác nhau như: phân bố chuẩn, phân bố Chi-squared, phân bố đối xứng… dựa vào phương

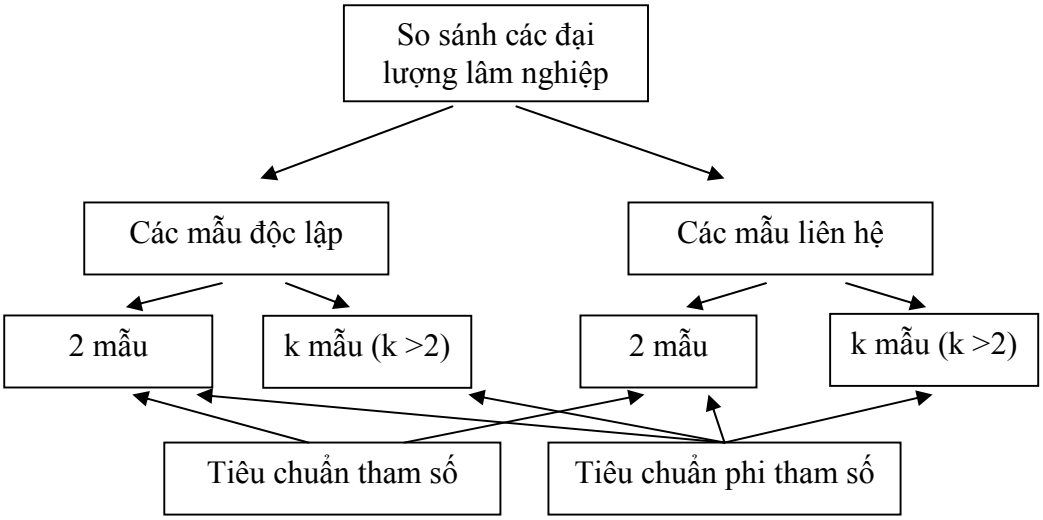
pháp biểu đồ.

Ngoài ra, Stata còn trang bị hệ thống lệnh vẽ biểu đồ như trên hình 02. Hệ thống lệnh này sẽ giúp chúng ta xây dựng nhiều loại biểu đồ khác nhau, từ biểu đồ dạng 2D, đến biểu đồ không gian 3D, từ biểu đồ dạng đường, cột đến biểu đồ tần số và chuỗi thời gian. Những biểu đồ được xây dựng từ những thông tin của các biến. Đó là những cơ sở trực quan để đề xuất và kết luận về đối tượng nghiên cứu.

3.2. Khả năng so sánh và tìm ra điều kiện tốt nhất cho các đại lượng lâm nghiệp

Trong phân tích số liệu lâm nghiệp, có rất nhiều trường hợp chúng ta phải so sánh các mẫu thu được như so sánh đường kính, chiều cao tại 2 vị trí địa hình khác nhau, độ chính xác của 2 loại dụng cụ, 2 loại thiết bị quan trắc khác nhau hay tìm hiểu sự ảnh hưởng của mật độ đến sinh trưởng cây bản địa…. Những so sánh và phân tích này nhằm cho thấy các mẫu có sự khác biệt hay không, các nhân tố có thực sự ảnh hưởng tới kết quả thí nghiệm hay không? Và công thức nào, điều kiện nào là tốt nhất cho sự sinh trưởng và phát triển của cây rừng cũng như các đại lượng lâm nghiệp khác.

Để giải quyết vấn đề này trong Stata đã có hệ thống các lệnh so sánh. Các lệnh này giúp các nhà khoa học lâm nghiệp có thể thực hiện các tiêu chuẩn thống kê một cách nhanh chóng, đơn giản. Stata cung cấp cả những tiêu chuẩn tham số và phi tham số như tiêu chuẩn U của Mann và Whitney, tiêu chuẩn Kruskal and Wallis, tiêu chuẩn Wilcoxon.... Từ việc so sánh 2 mẫu đến nhiều mẫu, từ mẫu độc lập đến mẫu liên hệ. Có thể tóm tắt các lệnh so sánh từ 2 mẫu trở lên trong Stata như sau:



Sơ đồ 01. ***Sơ đồ các lệnh so sánh 2 mẫu trở lên trong Stata***

Bên cạnh các lệnh so sánh, Stata còn mạnh trong phân tích phương sai để tìm hiểu ảnh hưởng của một hay nhiều nhân tố như: địa hình, độ dốc, độ ẩm, ánh sáng, phân bón… tới sinh trưởng và phát triển của các đối tượng nghiên cứu lâm nghiệp như cây rừng, nấm, độ cứng và dẻo của ván ép… Stata cung cấp cả các lệnh để phân tích phương sai một nhân tố, hai nhân tố không lặp và 2 nhân tố có lặp lại. Những lệnh này giúp người phân tích trả lời được câu hỏi: các nhân tố đó có thực sự ảnh hưởng tới số liệu và đối tượng nghiên cứu hay không? Ví như tuổi cây có ảnh hưởng tới hệ số biến đổi các bon hay không? Ánh sáng có ảnh hưởng tới sinh trưởng cây con trong vườn ươm hay không? Địa hình có ảnh hưởng tới sinh trưởng cây rừng hay không? Nồng độ thuốc có ảnh hưởng tới sinh trưởng của nấm hay không? Hay nhiệt độ ép có ảnh hưởng tới độ bền của ván ép hay không?

Tuy nhiên, một hạn chế nhỏ của Stata trong phân tích phương sai đó là không tích hợp tiêu

chuẩn Duncan như trong Spss. Duncan là tiêu chuẩn được sử dụng để so sánh và kiểm tra sự khác biệt giữa các nhóm (David, 2008). Dựa vào việc phân nhóm của tiêu chuẩn Duncan, nhà nghiên cứu có thể tìm được những công thức tốt nhất cho đối tượng nghiên cứu. Ví như độ tàn che tốt nhất cho sinh trưởng cây bản địa, nồng độ thuốc tốt nhất để kích thích ra rễ, nhiệt độ tốt nhất cho ván ép…. Dưới đây là ví dụ một quy trình sử dụng Stata để phân tích phương sai một nhân tố.

Quy trình phân tích phương sai một nhân tố

1. Statistics\ Linear models and related \ ANOVA/ MANOVA\ Analysis of variance and convariance

2. Trong hộp thoại **anova - Analysis of variance and covariance** khai báo **d0** vào **Dependent variable,** khai báo **cong_thuc** vào **Model**

3. Ok

Dưới đây là ví dụ kết quả phân tích phương sai để tìm hiểu ảnh hưởng của độ tàn che tới sinh trưởng đường kính của loài Gội nếp trong giai đoạn vườn ươm ở giai đoạn 3-4 tháng tuổi (Bùi Mạnh Hưng, 2011).

Bảng 01. ***Kết quả phân tích phương sai một nhân tố***

Analysis of Variance

Source	SS	df	MS	F	Prob > F
Between groups	2.07758343	3	.692527809	74.82	0.0000
Within groups	4.57257639	494	.009256228		
Total	6.65015982	497	.013380603		

Bartlett's test for equal variances: chi2(3) = 22.3961 Prob>chi2 = 0.000

Nguồn: Bùi Mạnh Hưng (2011), ĐH quốc gia Úc

3.3. Khả năng phân tích và cung cấp các thông tin về mối quan hệ giữa các đại lượng lâm nghiệp

Trong phân tích số liệu lâm nghiệp, chúng ta thường xuyên phải phân tích mối quan hệ giữa các đại lượng trong lâm nghiệp như mối tương quan giữa hệ số chuyển đổi các bon (BEF) với đường kính, chiều cao cây, phân tích mối quan hệ giữa chiều cao ngọn lửa với độ ẩm của vật rơi rụng… Dựa vào kết quả của việc phân tích này sẽ giúp các nhà nghiên cứu dễ dàng xác định được những đại lượng khó đo đếm, khó tính toán như thể tích cây, hệ số chuyển đổi sinh khối, hình số… dựa vào những đại lượng dễ đo đếm, dễ tính toán như đường kính ngang ngực, chiều cao….Khi phân tích các mối quan hệ, thường có một số câu hỏi cần phải được trả lời là: mối quan hệ đó tuân theo dạng hàm nào? và mức độ quan hệ giữa các đại lượng là chặt hay không chặt? Để giải quyết vấn đề này trong Stata có trang bị các lệnh về phân tích tương quan hồi quy; phân tích tuyến tính và phi tuyến.

Phân tích tuyến tính bằng sử dụng lệnh "Linear regression" sau đó đưa các biến độc lập (X) và phụ thuộc (Y) vào. Các thông tin về tương quan như: hệ số tương quan, các tham số của phương trình, ước lượng và kiểm tra sự tồn tại của các tham số trong tổng thể như trong bảng dưới đây:

Lệnh "Nonlinear regression least square" tạm dịch là "Tương quan phi tuyến bằng phương pháp bình phương nhỏ nhất" có thể giúp các nhà phân tích thăm dò nhanh mối quan hệ giữa các đại lượng theo một số loại hàm cơ bản cơ bản khác nhau: Linear, logistic, Gompertz và hàm Exponential. Và cho biết

mức độ liên hệ giữa các đại lượng theo các dạng hàm này. Một chức năng khác của lệnh này là nó có thể giúp phân tích theo bất kỳ một dạng hàm phi tuyến nào. Thông tin cần có là các biến, các tham số và giá trị bắt đầu của các tham số (David, 2011). Stata sẽ tự động tính toán và dừng lại khi tổng bình phương sai số dư là bé nhất và hệ số xác định là lớn nhất (David, 2011).

IV. KẾT LUẬN

Stata là một phần mềm mạnh, nó có thể được ứng dụng để phân tích số liệu nói chung và số liệu nghiên cứu lâm nghiệp nói riêng. Stata có thể giúp giải đáp được nhiều vấn đề khó khăn đặt ra khi phân tích số liệu lâm nghiệp. Stata sẽ giúp các nhà nghiên cứu tính toán và nhận được các giá trị đặc trưng mẫu một cách nhanh chóng, từ các đặc trưng về vị trí tới các đặc trưng về hình dạng phân bố. Stata cũng có thể cung cấp cho người sử dụng các lệnh về ước lượng số trung bình tổng thể, các lệnh về so sánh các mẫu. Từ so sánh 2 mẫu đến nhiều mẫu, từ so sánh mẫu độc lập đến mẫu liên hệ. Từ đó các nhà nghiên cứu có thể kết luận được các mẫu là thuần nhất hay không thuần nhất. Hơn nữa, Stata còn có khả năng ứng dụng tốt trong phân tích phương sai một và hai nhân tố, trong phân tích tương quan tuyến tính cũng như phi tuyến. Vì vậy, chương trình này sẽ đem lại những cơ sở khoa học cần thiết cho các kết luận, đề xuất để góp phần vào việc quản lý và phát triển tài nguyên rừng bền vững trong tương lai.

TÀI LIỆU THAM KHẢO

1. Bùi Mạnh Hưng (2011), *Phân tích ảnh hưởng của các nhân tố sinh thái tới sinh trưởng của hai loài cây bản địa tại miền Bắc Việt Nam*, Luận văn thạc sỹ, ĐH Quốc gia Úc, Canberra, Úc.

2. Bùi Mạnh Hưng (2012), *Nghiên cứu khai thác và ứng dụng phần mềm Stata trong phân tích số liệu nghiên cứu Lâm nghiệp*, Đề tài cấp cơ sở năm 2012, ĐH Lâm nghiệp, Hà nội, Việt Nam.

3. David, L.(2008), *Online Statistics: An Interactive Multimedia Course of Study*, University of Houston-Downtown, Texas, USA. Available from: <http://www.onlinestatbook.com/index.html> (Accessed 14 November, 2011).

4. David G., 2011, *Nonlinear*, <http://faculty.chass.ncsu.edu/garson/PA765/nonlinear.htm> (Accessed 15 November, 2011)

5. Nguyễn Hải Tuất và Nguyễn Trọng Bình (2005), *Khai thác và sử dụng Spss để xử lý số liệu nghiên cứu trong lâm nghiệp*, Nhà xuất bản Nông nghiệp, Hà nội.

6. Wiki, 2012 <http://en.wikipedia.org/wiki/Stata>, xem ngày 20/11/2012.

STATA – A NEW SOLUTION FOR FORESTRY DATA ANALYSIS

Bùi Manh Hung, Nguyen Thi Bich Phuong

SUMMARY

Stata is a statistical software which has developed from 1985. Stata has shown superior strength in data analysis in many fields, especially in forestry. Stata has been applied in many fields of forest science, such as forest inventory, silviculture, soil science, natural resources management, forest product processing and so on. It can analyze from the univariate to multi-variable and from qualitative to quantitative analysis. Stata can quickly provide information about the sample and the frequency distribution through descriptive statistics menu, graph and table menus. Stata also shows the strength of analysis to compare observed samples and find the best conditions for the forestry variables. The program can analyze by using parametric and non parametric tests, from two samples to multiple samples, from independent samples to paired samples. Moreover, Stata also has very powerful ability about analysis of variance with one or two factors. One of functional advantages of Stata is to analyze regression. It can analyze linear regression from one variable to multi-variable. Stata also can carry out non-linear correlation analysis for many difficult function types, such as Gompertz, Schumacher, Koller, Verhulst-Robertso and so on. Therefore, the use of Stata to analyze research data in forestry is really necessary and to be a concerned issue.

Key words: ***ANOVA, comparison, data analysis, descriptive statistics, regression Stata, statistics***

Người phản biện: **TS. Vũ Khắc Bảy**

Ngày nhận bài: 10/6/2013
Ngày phản biện: 28/6/2013
Ngày quyết định đăng: 20/9/2013