PHƯƠNG PHÁP THỐNG KÊ MÔ PHỎNG GẦN ĐÚNG CHO MÔ HÌNH NHIỀU YẾU TỐ ĐẦU RA

Trần Ngọc Sơn, Nguyễn Văn Đức, Trần Quang Hoàng Anh*

***Tóm tắt**: Bài báo nghiên cứu hệ thống phức tạp có nhiều yếu tố đầu ra. Những hệ thống phức tạp chịu sự tác động của nhiều yếu tố, vấn đề đặt ra là cần thiết phải đánh giá tầm quan trọng của từng yếu tố và phân tích sự ảnh hưởng của những yếu tố đó đến hệ thống, từ đó, xây dựng mô hình toán để phân tích cũng như dự báo sự phát triển của hệ thống. Tác giả đưa ra quy trình sử dụng phương pháp toán thống kê để nghiên cứu hệ thống phức tạp. Tác giả đề xuất phương pháp xây dựng mô hình gần đúng cho hệ thống nhiều yếu tố đầu ra trên nền tảng sử dụng dạng mở rộng của thuật toán bình phương tối thiểu. Đồng thời, tác giả đệ trình những phương pháp kiểm tra tính tương thích của mô hình để xem xét chất lượng cũng như độ tin cậy của mô hình vừa xây dựng.*

Từ khóa: Phân tích hồi quy, Mô hình nhiều yếu tố đầu ra, Thuật toán bình phương tối thiểu, Tiêu chuẩn Bayes.

1. MỞ ĐẦU

Vấn đề mô phỏng hệ thống hiện đại đòi hỏi cần phải dự báo sự phát triển hệ thống. Một trong những phương pháp quan trọng để thực hiện phân tích và dự báo đó là sử dụng phương pháp toán học. Phương pháp toán học có khả năng tính toán toàn diện sự tác động của nhiều yếu tố khác nhau đến kết quả của dự báo, tăng độ chính xác và tăng tốc độ phân tích cho dự báo.

Dựa vào số lượng các yếu tố đầu ra, ta có thể phân chia thành 2 loại mô hình chính: Mô hình một yếu tố đầu ra, và Mô hình nhiều yếu tố đầu ra. Mô hình một yếu tố đầu ra đã được nghiên cứu trong nhiều tài liệu, có thể kể đến những tác giả như: N. Dreiper, H. Smith, A.B. Uspenskii, V.U. Burmin, E.V. Markova, J. Johnson và các tác giả khác [1-5].

Mô hình nhiều yếu tố đầu ra là mô hình đồng thời quan sát một vài yếu tố đầu ra. Có nhiều mô hình có thể sử dụng để mô tả trạng thái của đối tượng nghiên cứu. Tuy nhiên, phương pháp và thuật toán mô hình hóa nhiều yếu tố đầu ra vẫn chưa được nghiên cứu một cách toàn diện.

Những mô hình hồi quy nhiều yếu tố đầu ra truyền thống có điểm đặc trưng là các hàm số trong những phương trình hồi quy giống nhau, ngoài ra các mô hình này không nghiên cứu sự tương quan giữa các yếu tố đầu ra. Vì vậy, mục đích của bài báo này là phát triển phương pháp thống kê cho mô phỏng gần đúng trong trường hợp đồng thời quan sát nhiều yếu tố đầu ra.

2. PHƯƠNG PHÁP THỐNG KÊ CHO MÔ PHỎNG GẦN ĐÚNG

Trên cơ sở các tài liệu đã giới thiệu, tác giả xây dựng quy trình phương pháp thống kê cho mô phỏng gần đúng của hệ thống nhiều yếu tố đầu ra. Quy trình này phù hợp để dự báo, phân tích những hệ thống phức tạp vì nó cho phép nghiên cứu sự tác động của nhiều yếu tố tới hệ thống được mô hình hóa. Các bước của quy trình được thể hiện như trong hình 1.

Những bước quan trọng nhất trong quy trình trên là xây dựng (bước 5) và kiểm định tính tương thích (bước 6) của mô hình mô phỏng gần đúng dựa trên dữ liệu thống kê. Dưới đây là các bước cụ thể trong quy trình.

a. Bước 1: Đặt vấn đề

Đây là bước đầu trong phân tích hệ thống bao gồm những nhiệm vụ cơ bản như: Phân tích những khó khăn gặp phải, liệt kê những những nhiệm vụ, phân tích cấu trúc của hệ thống và đưa ra những mục tiêu chung ban đầu khi phân tích hệ thống.

b. Bước 2: Tổng hợp những biến đầu vào và đầu ra của hệ thống

Đây là bước liệt kê tất cả những yếu tố tác động lên hệ thống.

c. Bước 3: Đặt vấn đề cho mô hình gần đúng phức tạp

Trong bước này hệ thống ban đầu sẽ được phân tích chi tiết hơn, và sẽ được xem xét, đặt vấn đề phù hợp với mô hình gần đúng nào trên nền tảng kết quả thống kê thực nghiệm.

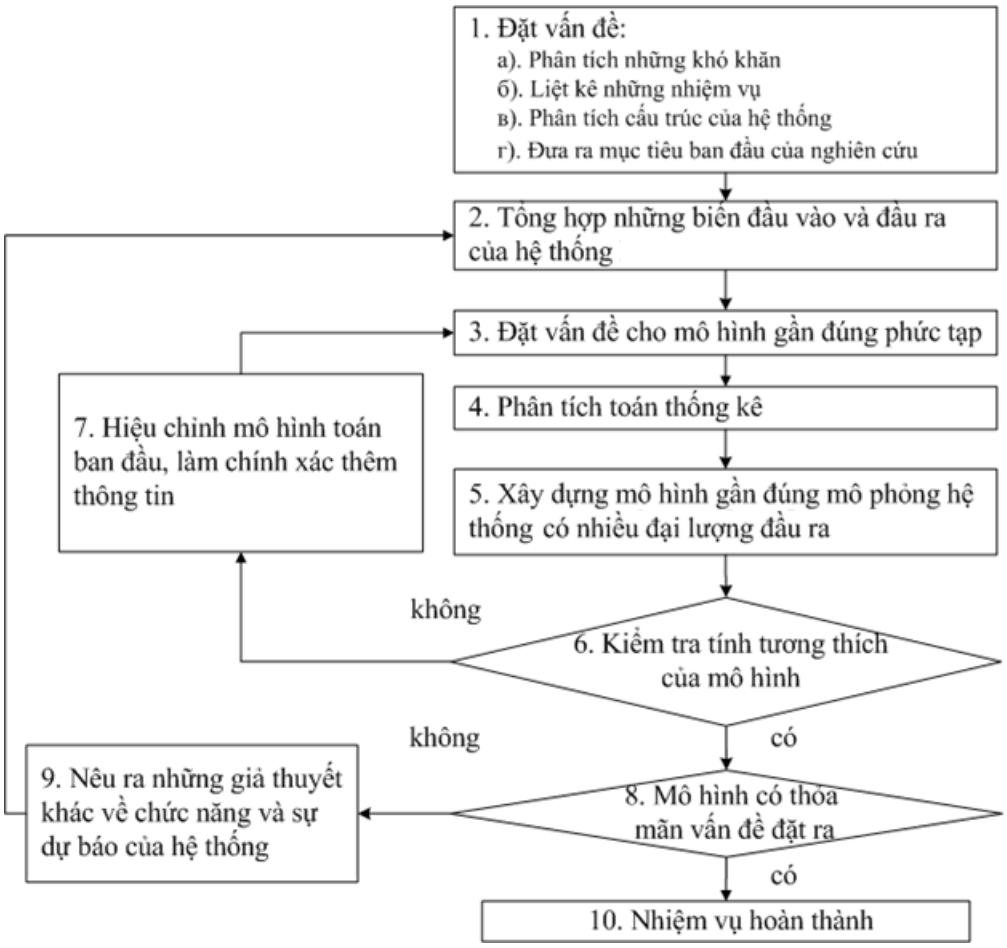
d. Bước 4: Phân tích toán thống kê

Sử dụng phương pháp toán thống kê để mở ra những khả năng phân tích hệ thống phức tạp. Cụ thể trong quá trình mô phỏng có thể sử dụng phương pháp toán thống kê để lựa chọn cấu trúc cho mô hình, hay nói cách khác là lựa chọn những biến có giá trị để đưa vào phân tích.

Để thực hiện nhiệm vụ trong trường hợp mô hình hồi quy đa biến có thể kể đến một vài phương pháp như phương pháp hồi quy từng bước và phương pháp Bayes.

d.1. Phương pháp hồi quy từng bước

Mục tiêu của phương pháp hồi quy từng bước [6-8] là lựa chọn từ các biến đầu vào để được một tập hợp những biến có ý nghĩa hơn, tương quan nhiều hơn với những yếu tố đầu ra. Thông thường quá trình này được thực hiện trên cơ sở sử dụng hệ số F-test, t-test hay những hệ số khác. Những cách sử dụng hồi quy từng bước bao gồm:



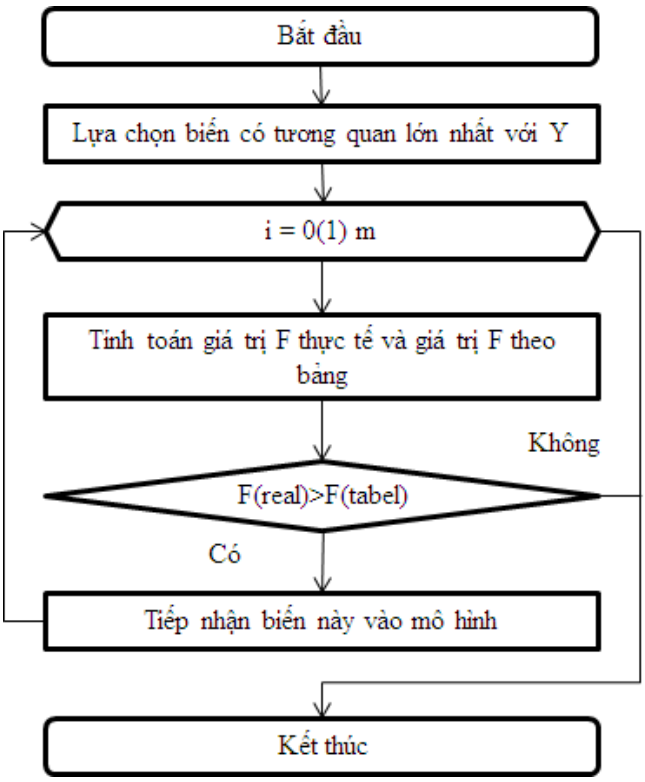
***Hình 1**. Quy trình phương pháp thống kê cho mô phỏng gần đúng.*

- Lựa chọn tiến: Ban đầu phương trình hồi quy không chứa biến nào. Những biến sẽ được tiếp nhận lần lượt nếu như chúng thỏa mãn một điều kiện đã xác định trước. Thứ tự tiếp nhận biến là mức độ quan trọng của biến đó đối với các yếu tố đầu ra (sơ đồ phương pháp được thể hiện như trong hình 2).

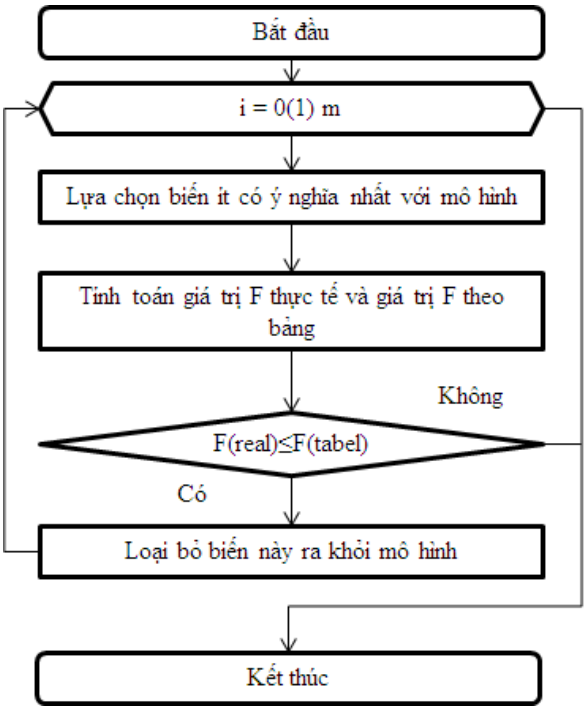
- Loại bỏ lùi: Ban đầu tất cả các biến sẽ được đưa vào phương trình hồi quy. Sau đó theo thứ tự những biến này sẽ bị loại bỏ khỏi mô hình theo một tiêu chí thích hợp (sơ đồ phương pháp được thể hiện như trong hình 3).

- Lựa chọn từng bước: Đây là cách thức kết hợp 2 phương pháp trên. Trong từng giai đoạn lựa chọn tiến sẽ đồng thời loại bỏ biến.

Trong thực tế, phương pháp hồi quy từng bước có một số hạn chế như việc không đưa ra phương trình hồi quy tối ưu với mô hình số lượng biến đầu vào lớn. Nguyên tắc tương quan giữa những biến đầu vào sẽ khiến cho những biến quan trọng có thể



***Hình 2**. Sơ đồ phương pháp lựa chọn tiến.*



***Hình 3**. Sơ đồ phương pháp loại bỏ lùi.*

không được đưa vào phương trình. Để có được cấu trúc mô hình tối ưu, cần phải tính toán những trường hợp, trong đó phân tích tất cả những khả năng kết hợp. Tuy nhiên, phương pháp hồi quy từng bước sẽ có hiệu quả cao với những mô hình có số lượng thống kê lớn hơn nhiều số lượng biến đầu vào.

d.2. Phương pháp Bayes

Phương pháp này có tên là BMA (Bayesian Model Average) [9-11] là thuật toán lựa chọn mô hình nhiều biến. Trong đó mỗi mô hình có một trọng số, trọng số này là BIC (Bayesian Information Criterion). BIC là tiêu chuẩn lựa chọn mô hình từ tập hợp mô hình tham số, mô hình này phụ thuộc vào số lượng tham số. Để đánh giá mô hình này ta sử dụng phương pháp ước lượng hợp lý cực đại, giá trị này có thể tăng lên khi thêm những tham số mới. Tiêu chuẩn Bayes cho phép giải quyết bài toán với số lượng tham số lớn, đưa ra hệ số phạt khi tăng số lượng tham số của mô hình. Tiêu chuẩn này gần giống với tiêu chuẩn thông tin Akaike, chỉ khác là giá trị phạt nghiêm ngặt hơn khi tăng số lượng tham số của mô hình.

Giả sử ta có: $X = \{x_i\}_{i=1}^{n}$ là một bộ phận của mẫu, trong đó từng thành phần đặc trưng cho biến $x_i = (x_{i1},...,x_{ik})$. Khi đó, tiêu chuẩn thông tin Bayes sẽ được tính theo công thức:

$$BIC = -2\ln(L) + k\ln(n),$$

trong đó, L là giá trị cực đại của hàm số hợp lý của mẫu quan sát với số lượng tham số cho trước.

Trong trường hợp mô hình hồi quy tuyến tính tiêu chuẩn được thể hiện thông qua SSE là tổng bình phương của số dư:

$$BIC = n\ln\frac{SSE}{n} + k\ln(n).$$

Từ những mô hình được xem xét, ta sẽ chọn mô hình có giá trị tiêu chuẩn Bayes nhỏ hơn. Tiêu chuẩn Bayes phụ thuộc vào số lượng tham số và tổng bình phương số dư của mô hình. Thay đổi biến phụ thuộc và tăng số lượng các biến sẽ làm thay đổi giá trị tiêu chuẩn Bayes.

e. Bước 5: Xây dựng mô hình mô phỏng gần đúng

Dưới đây là mô hình nhiều yếu tố đầu ra quan trọng trong mô hình mô phỏng:

$$y_i = \eta(x_i, \Theta) + \varepsilon_i, \ (i = \overline{1,n}), \tag{1}$$

trong đó:

$x_i^T = (x_{1i},...,x_{ki})$ là đại lượng độc lập hay những yếu tố đầu vào;

$y^T = (y_1,...y_\ell)$ là yếu tố phụ thuộc (đầu ra);

n là số lần quan sát; l là số lượng biến đầu ra;

$\Theta^T = (\theta_1,...\theta_m)$ là tham số chưa biết;

$\eta^T(x_i,\Theta) = \left(\eta_1(x,\Theta),...,\eta_\ell(x,\Theta)\right)$ là hàm số cho trước;

ε_i là sai số ngẫu nhiên, tuân theo những tiêu chuẩn sau:

$$E[\varepsilon_i] = 0,\ E[\varepsilon_i \varepsilon_i^T] = d(x_i),\ E[\varepsilon_i \varepsilon_j] = 0,\ i \neq j,\ \left|d(x_i)\right| \neq 0$$

trong đó, E là giá trị kỳ vọng.

Nhiệm vụ của vấn đề đặt ra là phải xác định giá trị tham số cho mô hình được thể hiện trong công thức (1). Để tìm giá trị này tác giả sử dụng biến thể của thuật toán bình phương tối thiểu như sau:

$$\min_\theta S(\theta) = \min_\theta \sum_{j=1}^{l}\sum_{k=1}^{l}\sum_{i=1}^{n} \omega_{jki}(y_{ji} - \eta_j(x_i,\theta))(y_{ki} - \eta_k(x_i,\theta)), \qquad (2)$$

trong đó: $\omega_i = d^{-1}(x_i)$ là trọng số của mô hình.

Trong trường hợp tuyến tính theo tham số, ta có:

$$\eta(x,\Theta) = F^T(x)\Theta,$$

trong đó:

$$F(x) = \|f_1(x),...f_\ell(x)\| = \begin{Vmatrix} f_{11}(x) & ... & f_{1\ell}(x) \\ ... & ... & ... \\ f_{m1}(x) & ... & f_{m\ell}(x) \end{Vmatrix}$$

Giá trị tham số tuyến tính tốt nhất có dạng:

$$\hat{\Theta} = \mathrm{M}^{-1}Y, \qquad (3)$$

trong đó:

$$\mathrm{M} = n^{-1}\sum_{i=1}^{n} F(x_i)\omega_i F^T(x_i),\ Y = n^{-1}\sum_{i=1}^{n} F(x_i)\omega_i y_i,\ \omega_i = d^{-1}(x_i).$$

hoặc dưới dạng chi tiết:

$$M = (M_{jk}),\ j,k = \overline{1,\ell},\ Y^T = (Y_1,...,Y_\ell),$$

trong đó:

$$M_{jk} = \sum_{i=1}^{n} \omega_{jki} f_j(x_i) f_k^T(x_i),$$

$$Y_j = \sum_{i=1}^{n} \sum_{k=1}^{l} \omega_{jki} y_{ki} f_j(x_i).$$

f. Bước 6: Kiểm tra tính tương thích của mô hình

Sau khi tìm được giá trị tham số của mô hình gần đúng, từng phương trình sẽ được kiểm tra tính tương thích. Để đạt được điều này có thể sử dụng những tiêu chuẩn như tiêu chuẩn Student, tiêu chuẩn Khi bình phương, tiêu chuẩn Fisher…, giống như mô hình hồi quy đơn thuần.

Để kiểm định tính tương thích của mô hình nhiều yếu tố đầu ra tác giả đề xuất sử dụng 3 tiêu chuẩn sau:

f.1. Sai số gần đúng

Để xem xét độ chính xác của mô hình, ta sử dụng giá trị tương tự với sai số gần đúng như công thức (4). Sai số gần đúng là độ lệch trung bình của những giá trị thực tế và giá trị được tính toán theo mô hình [3]:

$$\overline{A} = \frac{1}{n} \sum_{i=1}^{n} \frac{sum \mid y_i - \hat{y}_i \mid}{sum(y_i)}.100\%, \quad (4)$$

trong đó: y_i là giá trị thực tế; $\hat{y}_i$ là giá trị theo mô hình.

Chỉ nên sử dụng mô hình để phân tích và dự báo nếu có sai số nhỏ hơn 15%. Nếu sai số nhỏ hơn 5% thì mô hình có độ chính xác cao.

f.2. Hệ số xác định

Để xem xét chất lượng của mô hình ta sử dụng đại lượng tương tự như hệ số xác định như công thức (5). Hệ số xác định là đại lượng thể hiện phần trăm biến động của yếu tố đầu ra được giải thích bởi các yếu tố đầu vào [3]:

$$R^2 = 1 - \frac{\sum tr E_i^T \omega_i E_i}{\sum tr (E_i^0)^T \omega_i E_i^0}, \quad (5)$$

trong đó: tr là vết của ma trận, $E_i = \hat{y}_i - y_i$, $E_i^0 = y_i - \bar{y}$, $\bar{y}$ là giá trị trung bình.

Hệ số xác định càng gần với 1 thì chất lượng mô hình càng tốt.

f.3. Tính ổn định của tham số

Khi xem xét độ ổn định của giá trị tham số của mô hình nhiều yếu tố đầu ra có thể chia ra làm 2 trường hợp:

- Trường hợp dữ liệu được tổng hợp trong thời gian dài: Ta chia nhỏ cơ sở dữ

liệu và kiểm tra độ ổn định của tham số thông qua những mô hình nhỏ này. Nếu những tham số thu được có khuynh hướng không ổn định thì việc sử dụng mô hình xây dựng trên dữ liệu đầy đủ sẽ không đáng tin cậy.

- Trường hợp dữ liệu được tổng hợp trong thời gian ngắn: Ta chia dữ liệu ra làm 2 phần, sau đó sử dụng 1 phần để xây dựng mô hình dự báo, phần còn lại để kiểm tra tính đúng đắn của mô hình. Như vậy, có thể tính toán chất lượng của mô hình dự báo trên dữ liệu đầy đủ. Nếu mô hình tìm được không có độ chính xác cao chúng ta bước sang bước 7.

g. Bước 7: Hiệu chỉnh mô hình

Hiệu chỉnh lại mô hình toán và làm chính xác thêm thông tin, sau đó quay lại Bước 3 để đặt lại vấn đề cho mô hình gần đúng. Nếu mô hình có chất lượng tốt, tương thích với giá trị thống kê ta chuyển sang bước 8.

h. Bước 8: Kiểm tra mô hình

Kiểm tra xem mô hình có thỏa mãn những vấn đề nhiệm vụ đã được đặt ra ở bước 1 hay không. Nếu không ta chuyển sang bước 9.

i. Bước 9: Nêu ra những giả thuyết khác về chức năng và sự dự báo của hệ thống

Từ những giả thuyết mới này ta sẽ tổng hợp và lựa chọn lại những yếu tố đầu vào và đầu ra cho mô hình (quay lại bước 2).

3. KẾT LUẬN

Trong phạm vi bài báo tác giả đã đề xuất phương pháp nghiên cứu mô hình gần đúng của hệ thống, trong đó quan sát đồng thời nhiều yếu tố đầu ra. Tác giả đã sử dụng dạng biến thể của thuật toán bình phương tối thiểu để xác định giá trị tham số của mô hình gần đúng dựa vào kết quả thống kê, đồng thời đệ trình những phương pháp để kiểm tra tính tương thích của mô hình dựa vào giá trị như sai số gần đúng và hệ số xác định.

TÀI LIỆU THAM KHẢO

[1]. N. Dreiper, G. Smit, "*Applied regression analysis*", 2nd ed. Russian, Moscow, Book 1 (1986), pp. 366; Book 2 (1987), pp. 351.

[2]. L.N. Ezhova, "*Econometrics: The initial course with the probability theory and mathematical statistics basics*", Baikal State University Economics and Law Publ. (2008), pp. 287.

[3]. J. Johnson, "*Methods of econometrics*", Russian, Moscow, Statistika Publ.

(1980), pp. 444.

[4]. A.B. Uspenskii, B.V. Fedorov, "*Computational aspects of the method of least squares in the analysis and design of regression experiments*", Moscow State University Publ. (1975), pp. 168.

[5]. E.B. Маркова, "*Планирование эксперимента в условиях неоднородностей*", E.B. Маркова, A.H. Лисенков. M.: Наука (1973), pp. 220.

[6]. R.R Hocking, "*Criteria for selection of a subset regression: which one should be used?*", Technometrics. **Vol. 14** (1972), pp. 967-970.

[7]. R.R Hocking, "*The analysis and selection of variables in linear regression*", Biometrica, **Vol. 32**, No. 2 (1976), pp. 1-49.

[8]. C.H.A Li, "*Sequential method for screening experimental variables*", Journal of the American Statistical Association, **Vol. 57**, No. 298 (1962), pp. 455-477.

[9]. J.A. Hoeting, D. Madigan, A.E. Raftery, C.T. Volinsky, "*Bayesian Model Averaging: A Tutorial*", Statistical Science, **Vol. 14**, No. 4 (1999), pp. 382-417.

[10]. P.J. Brown, "*Bayes model averaging with selection of regressors*", Journal of the Royal Statistical Society, Part 3 (2002), pp. 519-536.

[11]. A.E. Raftery, "*Bayesian Model Selection in Social Research*", Sociological Methodology, **Vol. 25** (1995), pp. 111-163.

ABSTRACT

APPROXIMATED STATISTICAL APPROACH FOR MULTIPLE OUTPUT MODELS

In this article, the authors focus on complex systems, which have many outputs. Complex systems are affected by many factors, and the issue is that, it is necessary to evaluate the importance of each factor and analyze the effect of those factors on the systems, from which to build mathematical model for analysis as well as predicting the development of the systems. The authors propose a procedure using statistical methods to study complex systems. The authors then propose an approximated modeling approach for multi-factor systems based on the use of the expansion form of the least squares algorithm. At the same time, the authors also propose methods to validate the compatibility as well as the reliability of the constructed model.

Keywords: Regression analysis, Multi-output model, Least squares algorithm, Bayesian information criterion.

Nhận bài ngày 22 tháng 02 năm 2017
Hoàn thiện ngày 10 tháng 4 năm 2017
Chấp nhận đăng ngày 01 tháng 5 năm 2017

Địa chỉ: Trung tâm 586, Cục Công nghệ thông tin

*Email: newsv2004@gmail.com