Tạp chí Khoa học Đại học Cần Thơ
Số chuyên đề: Khoa học tự nhiên
ISSN 1859-2333 | e-ISSN 2815-5599

DOI:10.22144/ctujos.2024.330

MỘT SỐ HỆ SỐ PHÂN TÍCH PHƯƠNG SAI GIẢM THIỂU SỐ THAM SỐ TRONG MÔ HÌNH THỐNG KÊ

Trần Văn Lý[1*], Nguyễn Thị Ngọc Mỹ[2], Nguyễn Thị Như Ý[2], Nguyễn Thị Cẩm Hường[2], Nguyễn Thị Tuyết Nhi[3] và Lê Thị Minh Thư[3]

[1]*Khoa Khoa học tự nhiên, Trường Đại học Cần Thơ*
[2]*Lớp Toán ứng dụng – K47, Trường Đại học Cần Thơ*
[3]*Lớp cao học Lý thuyết xác suất và thống kê toán học – K29, Trường Đại học Cần Thơ*
**Tác giả liên hệ (Corresponding author): tvly@ctu.edu.vn*

Thông tin chung (Article Information)

Nhận bài (Received): 08/04/2024
Sửa bài (Revised): 24/06/2024
Duyệt đăng (Accepted): 27/07/2024

Title: *Some variance analysis coefficients for minimizing the number of parameters in statistical models*

Author(s): *Tran Van Ly[1*], Nguyen Thi Ngoc My, Nguyen Thi Nhu , Nguyen Thi Cam Huong, Nguyen Thi Tuyet Nhi and Le Thi Minh Thu*

Affiliation(s): *Can Tho University*

TÓM TẮT

Trong bài báo này, các hệ số và chỉ số có thể sử dụng để đánh giá mức độ tác động của các biến đầu vào trong mô hình thống kê được xem xét. Hệ số quan trọng và chỉ số độ nhạy là hai định lượng được thực nghiệm trên mô hình hồi quy đa thức, trong đó tiếp cận Monte Carlo được sử dụng để tính toán các chỉ số độ nhạy.

Từ khoá: *Chỉ số nhạy, hệ số quan trọng, hệ số xác định, hồi quy đa thức, mô phỏng Monte Carlo*

ABSTRACT

In this paper, coefficients and indices that can be used to assess the impact of input variables in statistical models are considered. The importance coefficient and sensitivity indices are two quantities that are performed on the polynomial regression model, in which the Monte Carlo approach is used to calculate the sensitivity indicators.

Keywords: *Sensitivity indice, coefficient of importance, coefficient of determination, polynomial regression, Monte Carlo method*

1. GIỚI THIỆU

Trong nghiên cứu phát triển khoa học công nghệ, có nhiều trường hợp cần phải thực nghiệm khảo sát trên mô hình có số biến tham số rất lớn. Tuy nhiên đối với từng trường hợp quan tâm, khảo sát cụ thể thì thực tế chỉ có một số biến tham số có vai trò tác động chính. Chẳng hạn trong việc nghiên cứu phát triển hệ thống hỗ trợ lái xe nâng cao ADAS (Advanced Driver Assistance Systems), các mô hình thực nghiệm được thử nghiệm có số lượng lên đến hàng chục thậm chí hàng trăm các biến tham số có liên quan (Zhao et al, 2017). Thế nhưng đánh giá hiệu quả của hệ thống trên mỗi một kịch bản thử nghiệm cụ thể nào đó thì sự tác động của các tham số có các mức độ cao thấp khác nhau. Do đó để tối ưu hóa các mô hình nghiên cứu, những cơ sở để loại bỏ những tham số có mức tác động yếu trong mô hình cần được nghiên cứu.

Bài báo này trình bày các nghiên cứu về các chỉ số định lượng nhận được qua tính toán tỷ lệ phương sai đầu ra do một biến đầu vào ngẫu nhiên gây ra. Bên cạnh chỉ số quen thuộc hay dùng như hệ số tương quan đánh giá mức độ tác động liên hệ của các biến thống kê đối với hàm mục tiêu nghiên cứu, hệ số xác định của mô hình thường được dùng để đánh giá chất lượng các mô hình xấp xỉ, phương pháp phân tích độ nhạy (Sobol, 2001), có thể được áp dụng trực tiếp như một công cụ xử lý hậu kỳ để phân tích sự đóng góp của từng biến đầu vào đối với sự phân tán của các phản hồi giá trị của biến đầu ra

trong mô hình thống kê. Để định lượng sự đóng góp này, các phương pháp dựa trên phương sai rất phù hợp.

Các mẫu thực nghiệm nhận được từ mô phỏng Monte Carlo theo dạng thuật toán Gibbs quét tuần tự (Brémaud, 1999; Levine & Casella, 2006; Rubinstein & Kroese, 2017).

2. CÁC CHỈ SỐ ĐO ĐỘ NHẠY CỦA CÁC BIẾN ĐẦU VÀO MÔ HÌNH

2.1. Mô hình phân tích phương sai

Xét I^n là một siêu lập phương đơn vị cấp n với I là khoảng đơn vị $[0; 1]$, $x = (x_1, x_2, \dots, x_n)$ là điểm thuộc I^n, ký hiệu $dx = dx_1 dx_2 \dots dx_n$. Một hàm khả tích $u(x)$ trên I^n được nghiên cứu dưới dạng

$$u(x) = u_0 + \sum_{s=1}^{n} \sum_{i_1<\dots<i_s}^{n} u_{i_1\dots i_s}(x_{i_1}, \dots, x_{i_s}), (1)$$

trong đó $1 \le i_1 < \dots < i_s \le n$.

Định nghĩa 2.1. *Mô hình (1) được gọi là một dạng phân tích phương sai của $u(x)$ nếu*

$$\int u_{i_1\dots i_s}(x_{i_1}, \dots, x_{i_s}) dx_k = 0, \ x_k = x_{i_1}, \dots, x_{i_s}.$$

Nếu $u(x)$ bình phương khả tích thì mọi số hạng $u_{i_1\dots i_s}(x_{i_1}, \dots, x_{i_s})$ trong (1) cũng bình phương khả tích và ta có

$$\int u^2(x)\, dx - u_0^2 =$$

$$\sum_{s=1}^{n} \sum_{i_1<\dots<i_s}^{n} \int u_{i_1\dots i_s}^2 dx_{i_1} \dots dx_{i_s}.$$

Các hằng số $V = \int u^2\, dx - u_0^2$, $V_{i_1\dots i_s} = \int u_{i_1\dots i_s}^2 dx_{i_1} \dots dx_{i_s}$ được gọi là các phương sai.

2.2. Chỉ số độ nhạy của một nhóm biến

Xét một nhóm m các biến tùy ý ($1 \le m \le n - 1$):

$$y = (x_{k_1}, \dots, x_{k_m}), \ 1 \le k_1 < \dots < k_m \le n,$$

và z là tập $n - m$ các biến còn lại, tức là $x = (y, z)$.

Đặt $K = (k_1, \dots, k_m)$, phương sai tương ứng với tập được xác định bởi

$$V_y = \sum_{s=1}^{n} \sum_{(i_1<\dots<i_s)\in K}^{n} V_{i_1\dots i_s}.$$

Với V_z là phương sai tương ứng với tập z thì được V_y^{tot} xác định theo hiệu dưới đây được gọi là phương sai tổng tương ứng với tập y:

$$V_y^{tot} = V_y - V_z.$$

Định nghĩa 2.2. *Các tỷ lệ $S_y = \frac{V_y}{V}$, $S_y^{tot} = \frac{V_y^{tot}}{V}$ lần lượt được gọi là chỉ số độ nhạy và chỉ số độ nhạy tổng của tập y.*

Chú ý là $S_y^{tot} = 1 - S_z$.

2.3. Tiếp cận Monte Carlo

Định lý 2.1. $V_y = \int u(x)u(y,t)dxdt - u_0^2$ (2).

Chứng minh. Tích phân trong **Định lý 2.1** có thể được biến đổi như sau

$$\int u(x)u(y,t)dxdt$$

$$= \int dy \int u(y,z)dz \int u(y,t)dt$$

$$= \int dy \left[\int u(y,z)dz\right]^2.$$

Áp dụng (1) ta viết được

$$u(y,z) = u_0 + \sum_{s=1}^{n} \sum_{(i_1<\dots<i_s)\in K}^{n} u_{i_1\dots i_s}(x_{i_1}, \dots, x_{i_s}).$$

Sau khi bình phương và lấy tích phân theo $dy = dx_{k_1} \dots dx_{k_m}$ ta được

$$\int u(x)u(y,t)dxdt$$

$$= u_0^2 + \sum_{s=1}^{n} \sum_{(i_1<\dots<i_s)\in K}^{n} V_{i_1\dots i_s} = u_0^2 + V_y \quad \square$$

Một công thức tương tự như trong (2) có thể được viết cho V_z:

$$V_z = \int u(x)u(t,z)dxdt - u_0^2.$$

Như vậy việc tính toán S_z và $S_y^{tot} = 1 - S_z$ có thể nhận được từ các tích phân $\int u(x)dx$, $\int u^2(x)dx$, $\int u(x)u(y,t)dxdt$, $\int u(x)u(t,z)dxdt$.

Từ đây phương pháp Monte Carlo có thể được sử dụng. Hai biến ngẫu nhiên ξ và ξ' phân phối đều trong I^n được xem xét và đặt $\xi = (\eta, \zeta)$, $\xi' = (\eta', \zeta')$. Ở mỗi mô phỏng Monte Carlo, các mô hình $u(\eta, \zeta)$, $u(\eta', \zeta)$ và $u(\eta, \zeta')$ được tính. Sau N mô phỏng các ước lượng Monte Carlo cần thiết bao gồm

$$\frac{1}{N}\sum_{i=1}^{N} u(\xi_i) \xrightarrow{P} u_0, \ \frac{1}{N}\sum_{i=1}^{N} u^2(\xi_i) \xrightarrow{P} V + u_0^2,$$

$$\frac{1}{N}\sum_{i=1}^{N} u(\xi_i)u(\eta_i, \zeta'_i) \xrightarrow{P} V_y + u_0^2,$$

$$\frac{1}{N}\sum_{i=1}^{N} u(\xi_i)u(\eta'_i, \zeta_i) \xrightarrow{P} V_z + u_0^2.$$

Trong đó, ký hiệu $\xrightarrow{P}$ là chỉ cho dạng hội tụ theo xác suất.

3. CÁC HỆ SỐ ĐÁNH GIÁ MỨC ĐỘ QUAN TRỌNG CÁC BIẾN TRONG MÔ HÌNH HỒI QUY ĐA THỨC

Xét một hàm đa thức dưới dạng

$$\mathbf{p}^T(\mathbf{x}) = \lfloor 1\ x_1\ x_2\ \dots\ x_1x_2\ \dots\ x_1^2\ \dots\ x_2^2\ \dots \rfloor.$$

Giá trị biến đầu ra $\mathbf{y}_i$ mô hình hồi quy đa thức tại một tập $\mathbf{x}_i$ của các biến đầu vào $\mathbf{X}$ được xác định bởi tổng của giá trị xấp xỉ $\hat{y}_i$ và sai số ϵ_i:

$$y_i(\mathbf{x}) = \hat{y}_i(\mathbf{x}) + \epsilon_i = \mathbf{p}^T(\mathbf{x})\boldsymbol{\beta} + \epsilon_i,$$

trong đó $\boldsymbol{\beta}$ là véc tơ của các hệ số hồi quy. Các hệ số này được ước lượng bởi lời giải bình phương bé nhất

$$\widehat{\boldsymbol{\beta}} = (\mathbf{P}^T\mathbf{P})^{-1}\mathbf{P}^T\mathbf{y},$$

trong đó $\mathbf{P}$ là ma trận giá trị mẫu của các biến đầu vào, $\mathbf{y}$ là véc tơ các giá trị tương ứng của các biến đầu ra.

3.1. Hệ số xác định

Hệ số xác định (CoD) R^2 có thể sử dụng để đánh giá chất lượng xấp xỉ của mô hình hồi quy đa thức

$$R^2 = 1 - \frac{SSE}{SST},$$

trong đó $SST = \sum_{i=1}^{N}(y_i - \bar{y})^2$ là tổng các bình phương biến thiên biến đầu ra, $SSE = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$ là tổng các biến thiên sai số của mô hình và y_i $(i = 1,2, \dots, N)$ là các giá trị mẫu của biến đầu ra với trung bình $\bar{y}$.

Nếu CoD càng gần 1 thì phép xấp xỉ đa thức có chất lượng càng tốt. Một nhược điểm của phép đo CoD là giới hạn phù hợp của nó đối với hồi quy đa thức, hồi quy tuyến tính. Đối với các mô hình xấp xỉ cục bộ khác, như nội suy Kriging, phép đo này có thể bằng hoặc gần bằng 1 nhưng chất lượng xấp xỉ vẫn kém.

3.2. Hệ số quan trọng

Trong mô hình hồi quy đa thức đã nêu ở trên, hệ số quan trọng $\mathrm{CoI}(x_i, y)$ của biến tương ứng với biến đầu vào được định nghĩa bởi

$$\mathrm{CoI}(x_i, y) = R^2_{y,\mathbf{x}} - R^2_{y,\mathbf{x}_{\sim\mathbf{i}}},$$

trong đó $R^2_{y,\mathbf{x}}$ là CoD của mô hình đầy đủ với tất cả các biến đầu vào, $R^2_{y,\mathbf{x}_{\sim\mathbf{i}}}$ là CoD của mô hình thu gọn trong đó loại bỏ mọi số dạng tuyến tính, các số hạng tương tác và các số hạng bậc 2 có chứa x_i của mô hình đa thức, $\mathbf{x}_{\sim\mathbf{i}}$ là ký hiệu cho tập tất cả các biến đầu vào ngoại trừ biến x_i.

Nếu một biến có tầm quan trọng thấp thì CoI của nó gần bằng không, vì mô hình hồi quy đa thức đầy đủ và thu gọn (loại bỏ các số hạng chứa biến đó) có chất lượng tương tự.

Dựa trên một giá trị tối thiểu CoI_{min}, chỉ những biến có hệ số quan trọng lớn hơn giá trị này mới được xem xét trong mô hình xấp xỉ cuối cùng. Thường giá trị CoI_{min} được lấy trong khoảng từ 1% đến 9% (Brémaud, 1999).

3.3. Hệ số tiên lượng

Hệ số tiên lượng CoP (Sobol, 2001) được định nghĩa bởi

$$\mathrm{CoP} = 1 - \frac{SSE_{test}}{SST_{test}},$$

trong đó SST_{test} là tổng các bình phương biến thiên biến đầu ra kiểm tra, SSE_{test} là tổng các biến thiên sai số của mô hình kiểm tra.

Kết hợp với các chỉ số độ nhạy được nêu ở Mục 2, chúng ta có thể sử dụng kết hợp với CoP để đánh giá mức độ quan trọng của các biến đầu vào. Chỉ số độ tiên lượng của biến đầu vào x_i được tính bởi

$$\mathrm{CoP}(x_i) = \mathrm{CoP}^{optimal} \times S_{OP}^{tot}(x_i),$$

trong đó $\mathrm{CoP}^{optimal}$ là hệ số tiên lượng của mô hình được xây dựng tối ưu, $S_{OP}^{tot}(x_i)$ là chỉ độ nhạy của biến đầu vào x_i được tính tương ứng trên mô hình tối ưu của giá trị $\mathrm{CoP}^{optimal}$. Chỉ số này có thể được sử dụng để đánh giá mức độ quan trọng của biến đầu vào x_i trong mô hình dùng để tiên lượng, dự báo. Biến đầu vào có chỉ số này càng lớn thì mức độ đóng góp của nó vào chất lượng xấp xỉ và dự báo của mô hình càng cao.

4. ÁP DỤNG THỰC NGHIỆM

4.1. Mô hình thực nghiệm

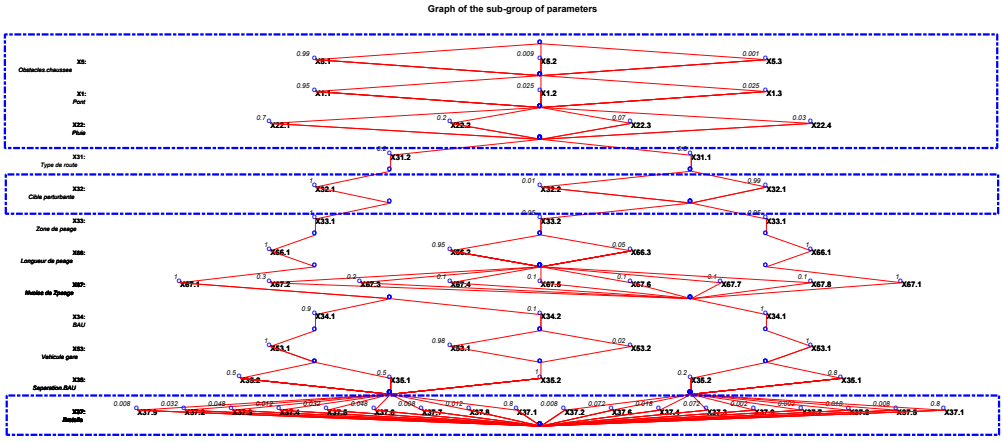
Xét một nhóm gồm 16 tham số độc lập, là các tham số đơn giản nhất trong mô hình mô phỏng các kịch bản thẩm định hệ thống hỗ trợ lái xe nâng cao ADAS sử dụng cho xe tự hành trình (xe tự lái). Các tham số X5, X1, X22, X31, X32, X33, X66, X67, X34, X53, X35, X37 được xét trong mô hình vectơ ngẫu nhiên X có 12 biến tham số X = (X5, X1,…, X37). Các tham số thành phần này có các trạng thái được mã hóa như trong Bảng 1 với các xác suất tương ứng được giới thiệu ở Hình 1.

Các trạng thái có xác suất xảy ra nhỏ của các tham số (biến) được mã hóa bởi các giá trị lớn nên xác suất để hàm tổng của những biến này S(X) = X5 + X1 + ⋯ + X37 nhận giá trị lớn sẽ rất nhỏ. Dạng hàm tổng này có thể được sử dụng làm hàm mục tiêu (biến đầu ra) để khảo sát các vấn đề liên quan đến sự kiện hiếm (sự kiện có xác suất xảy ra rất bé,

thường hiểu là nhỏ hơn 10^{-6}). Dạng hàm tổng này được sử dụng để xây dựng mô hình hồi quy đa thức trên các biến liên quan đến tổ hợp sự kiện hiếm,… Qua đó đánh giá mức độ đóng góp của các biến đầu vào đối với mô hình xấp xỉ theo các hệ số quan trọng và hệ số tiên lượng của chúng.

Bảng 1. Mã hóa các trạng thái các tham số thành phần của vectơ $\mathbf{X} = (\mathbf{X5}, \mathbf{X1}, \ldots, \mathbf{X37})$

Ký hiệu	Tham số	Mã hóa các trạng thái
X5	*Obstacles.chaussee*	X5.1, X5.2, X5.3
X1	*Pont*	X1.1, X1.2, X1.3
X22	*Pluie*	X22.1, X22.2, X22.3, X22.4
X31	*Type.de.rue*	X31.1, X31.2
X32	*Cible.perturbante*	X32.1, X32.2
X33	*Zone.de.peage*	X33.1, X33.2
X66	*Longueur.de.peage*	X66.1, X66.2, X66.3
X67	*Nvoies.de.Zpeage*	X67.1, X67.2, X67.3, X67.4, X67.5, X67.6, X67.7, X67.8
X34	*BAU*	X34.1, X34.2
X53	*Vehicule.gare*	X53.1, X53.2
X35	*Separation.BAU*	X35.1, X35.2
X37	*Bretelle*	X37.1, X37.2, X37.3, X37.4, X37.5, X37.6, X37.9, X37.8, X37.9

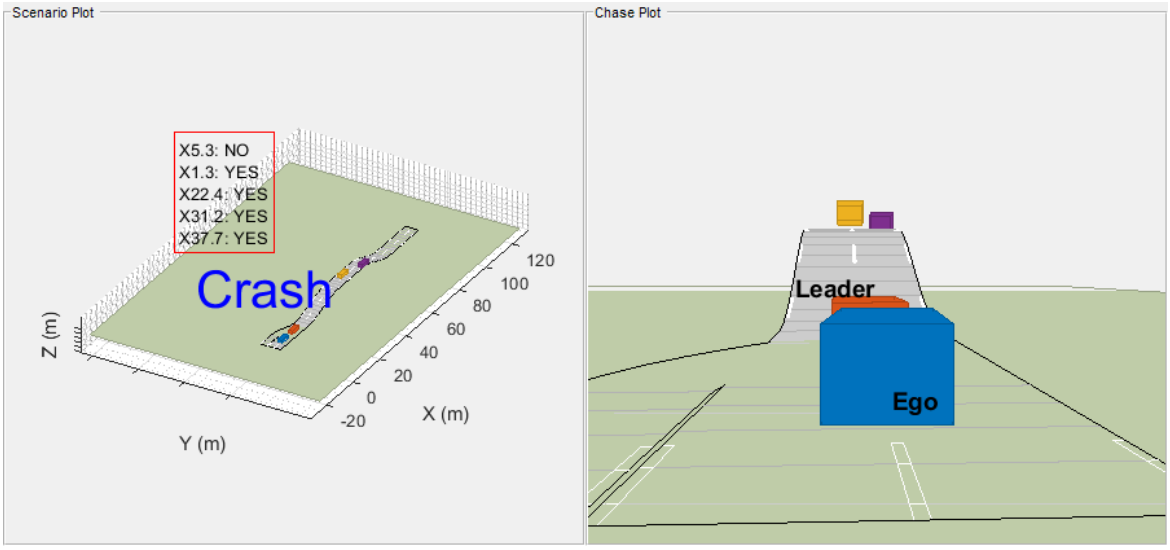


Hình 1. Phân bố xác suất các trạng thái của các thành phần véctơ $\mathbf{X} = (\mathbf{X5}, \mathbf{X1}, \ldots, \mathbf{X37})$

Mô hình hồi quy đa thức được xây dựng trên các biến liên quan đến tổ hợp sự kiện hiếm {X5.3-*xuất hiện chướng vật*, X1.3-*Xuất hiện đầu cầu*, X22.4-*Có mưa to*, X31.2-*Dạng đường nhanh*, X37.7-*Xuất hiện đường nhánh nhập làn*} (Hình 2) có thể dẫn đến rủi ro tai nạn va chạm của xe tự hành trình **Ego** thử nghiệm ADAS với xe đi trước **Leader**. Qua đó đánh giá mức độ đóng góp của các biến đầu vào đối với mô hình xấp xỉ theo các hệ số quan trọng và hệ số tiên lượng của chúng.

4.2. Kết quả thực nghiệm

Sử dụng mô phỏng Monte Carlo khởi tạo theo thuật toán lấy mẫu Gibbs quét tuần tự, mẫu mô phỏng được tạo có kích thước N = 8000 cho vectơ X = (X5, X1, …, X37) theo cấu trúc phân phối ở Hình 1, mẫu tổ hợp các biến (X5, X1, X22, X31, X37) được trích ra để xây dựng mô hình hồi quy đa thức cho hàm tổng. Kết quả tính toán hệ số quan trọng và chỉ số độ nhạy của các biến trong tổ hợp này được giới thiệu trong Bảng 2.



Hình 2. Kịch bản xuất hiện tổ hợp hiếm có thể dẫn đến va chạm của xe tự hành Ego

Bảng 2. Hệ số quan trọng và chỉ số độ nhạy của các biến trong tổ hợp (X5, X1, X22, X31, X37)

Biến Xi	X5	X1	X22	X31	X37
Hệ số quan trọng CoI(Xi)	0,000005	0.034647	0.159876	0.000006	0.789876
Chỉ số độ nhạy CoP(Xi)	0,0442	0,0494	0,1219	0,0441	0,4556

Các kết quả nhận được đánh giá theo hệ số quan trọng và theo chỉ số độ nhạy của các biến là rất tương đồng. Theo hệ số quan trọng, nếu các biến có hệ số nhỏ hơn $\mathrm{CoI}_{min} = 5\%$ (trong khoảng thông thường từ 1% đến 9%) bị loại thì chỉ có hai biến X22 và X37 được giữ lại xây dựng cho mô hình tối ưu phục vụ nghiên cứu. Hai biến này cũng có chỉ số độ nhạy cao hơn khác biệt. Đối với các mô hình phi tuyến, khi mà hệ số quan trọng không được tin tưởng cao (thường chỉ phù hợp tốt đối với mô hình đa thức và tuyến tính) thì ta có thể tiếp cận qua chỉ số độ nhạy (Sobol, 2001) để loại bớt các biến tham số có ít tác động trong mô hình, xây dựng mô hình xấp xỉ tốt nhất.

LỜI CẢM ƠN

Bài báo được thực hiện qua đề tài được tài trợ bởi Trường Đại học Cần Thơ, Mã số TSV2024-22, và sự hỗ trợ dữ liệu, kĩ thuật tính toán của Armines-Pháp thông qua đề tài mã số TCN2023-10. Các tác giả xin trân trọng cảm ơn các hỗ trợ rất quý báu này.

TÀI LIỆU THAM KHẢO

Brémaud, P. (1999). Markov chains – Gibbs Fields, Monter Carlo Simulation and Queues. Spinger – New York. https://doi.org/10.1007/978-1-4757-3124-8

Levine, R. A. and Casella, G. (2006). Optimizing random scan Gibbs samplers. *Journal of Multivariate Analysis, 97*(10), 2071-2100. https://doi.org/10.1016/j.jmva.2006.05.008

Rubinstein, R. Y. & Kroese, D. P. (2017). Simulation and the Monte Carlo method (Wiley Series in Probability and Statistics). John Wiley & Sons, Inc., Hoboken, NJ. https://doi.org/10.1002/9781118631980

Sobol, I.M. (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation, 55*(1-3), 271–280.

Zhao, D., Lam, H., Peng, H., Bao, S., LeBlanc, D. J., Nobukawa, K., & Pan, C. S. (2017). Accelerated Evaluation of Automated Vehicles Safety in Lane-Change Scenarios Based on Importance Sampling Techniques. *IEEE Transactions on Intelligent Transportation Systems*, *18*(3), 595-607. https://doi.org/10.1109/TITS.2016.258220