Tạp chí Khoa học Trường Đại học Cần Thơ
Phần A: Khoa học Tự nhiên, Công nghệ và Môi trường
website: ctujsvn.ctu.edu.vn

DOI:10.22144/ctu.jvn.2022.011

PHÂN TÍCH TIN CẬY ĐỐI VỚI CHUỖI DỮ LIỆU MÔ PHỎNG CÓ TÍNH LẶP VÀ TÍNH CHU KỲ

Trần Văn Lý[1*], Bùi Phong Quy[2], Nguyễn Lê Phúc Anh[2], Trang Thị Hiền[2] và Phan Thị Mỹ Tiên[3]

[1]*Khoa Khoa học Tự nhiên, Trường Đại học Cần Thơ*
[2]*Lớp cao học Lý thuyết xác suất và thống kê toán học – K26, Trường Đại học Cần Thơ*
[3]*Lớp Toán ứng dụng – K44, Trường Đại học Cần Thơ*
**Người chịu trách nhiệm về bài viết: Trần Văn Lý (email: tvly@ctu.edu.vn)*

Thông tin chung:
Ngày nhận bài: 01/09/2021
Ngày nhận bài sửa: 28/10/2021
Ngày duyệt đăng: 26/02/2022

Title:
Reliable analysis for simulation data chains having repeatability and seasonality

Từ khóa:
Chuỗi Markov, dữ liệu mô phỏng, kiểm định độc lập, kiểm định phù hợp, mô hình Markov ẩn

Keywords::
Goodness-of-fit test, hidden Markov model, independent test, Markov chain, simulation data

ABSTRACT

Presented in the paper are the two types of reliability analysis including the goodness-of-fit test for transition probabilities and the independent test for simulation data chain. These tests are applied for the simulation data chain having repeatability and seasonality. The simulation data chain is assumed that the Markov chain of order 1 and having stationary transition probabilities. Based on some existing research results on statistical inferences for Markov chains, the two statistics chi-square were established for using in the two mentioned tests. The illustration application is realized on the simulation data chain of the daily clear index sequence generated from the hidden Markov model having four states.

TÓM TẮT

Kiểm định về sự phù hợp với các xác suất chuyển và kiểm định về tính độc lập của chuỗi dữ liệu mô phỏng là hai dạng phân tích tin cậy được đề cập trong bài báo. Các kiểm định này được áp dụng đối với chuỗi dữ liệu mô phỏng có tính lặp và tính chu kỳ. Chuỗi dữ liệu mô phỏng được giả thiết là chuỗi Markov bậc nhất có các xác suất chuyển ổn định. Dựa vào một số kết quả nghiên cứu đã có về những suy luận thống kê trên các chuỗi Markov, hai dạng thống kê χ^2 – bình phương được xây dựng để sử dụng trong hai kiểm định đã nêu. Áp dụng minh họa được thực hiện trên chuỗi dữ liệu mô phỏng của dãy chỉ số sáng hàng ngày được khởi tạo từ mô hình Markov ẩn với bốn trạng thái.

1. GIỚI THIỆU

Trên nền tảng phát triển của công nghệ số, các mô hình ngẫu nhiên ngày càng được ứng dụng mạnh mẽ trong nhiều lĩnh vực như trong toán học (phép thử thống kê, lý thuyết thông tin phục vụ đám đông và đối sách), trong vật lý (vật lý hạt nhân, địa vật lý, quang học và khí tượng, khí quyển) cùng nhiều lĩnh vực khác như sinh học, hóa học, kỹ thuật quân sự và kinh tế. Các ứng dụng chủ yếu là để nắm bắt, quan sát, nghiên cứu các hiện tượng trong tự nhiên, kỹ thuật hay kinh tế xã hội và cả trong các nghiên cứu cơ bản thông qua các mô phỏng dưới các dạng cụ thể như các hình ảnh, các chuyển động, các phản ứng hay những phép thử, thử nghiệm,… Những hình thức ứng dụng khác biệt này có thể được xem xét, nghiên cứu ở một dạng chung là *các dữ liệu mô phỏng* được khởi tạo ban đầu từ các mô hình. Các cấu trúc mô hình và các phương pháp khởi tạo dữ liệu mô phỏng đã được nghiên cứu rất nhiều. Ngoài những mô hình cơ bản như Markov, Markov ẩn, các mô hình được xây dựng theo chuyển động Brown

(Elliott et al., 2010)…, còn có rất đa dạng các mô hình được xây dựng trong rất nhiều lĩnh vực khác nhau. Đối với các phương pháp khởi tạo dữ liệu mô phỏng, xuất phát từ phương pháp Monte Carlo (Rubinstein & Kroese, 2017), nhiều nhánh được phát triển rất phong phú như nhóm các thuật toán tạo mẫu Gibbs (Levine & Casella, 2006) hay khởi tạo mô phỏng theo phương pháp MCMC (Markov Chain Monte Carlo), …

Tuy nhiên, vấn đề phân tích sự tin cậy và suy luận kiểm chứng các tính chất và các điều kiện phù hợp của các chuỗi dữ liệu mô phỏng chưa được nghiên cứu phong phú lắm, mặc dầu những vấn đề này đã được chú ý đến từ trước đây rất lâu. Ví dụ như Cramér (1946) có các phân tích, nghiên cứu cơ bản về các suy luận chung thường gặp, hay Bartlett (1951) nghiên cứu về các suy luận trên chuỗi mô phỏng theo mô hình xác suất; còn các phân tích suy luận trên các dãy mô phỏng Markov có thể tìm thấy ở Anderson and Goodman (1957) và Billingsley (1961). Mặc dầu vậy, số lượng các nghiên cứu về nhóm vấn đề này đến nay là chưa nhiều, chưa phong phú tương xứng với sự phát triển đa dạng của các cấu trúc mô hình, các phương pháp, thuật toán khởi tạo mô phỏng.

Cũng có nhiều trường hợp ứng dụng dữ liệu mô phỏng, trong đó tính tin cậy của các dữ liệu khởi tạo nội tại đã được đảm bảo từ chính các phương pháp mô phỏng sử dụng; hay đã có các nghiên cứu lý thuyết liên quan nhưng người sử dụng lại thiếu sót không sử dụng để luận giải, suy luận cho chặt chẽ, giúp cho các ứng dụng của mình được thuyết phục hơn theo các tiêu chuẩn thống kê phù hợp.

Bài báo trình bày một ứng dụng mà trong đó một số kết quả nghiên cứu lý thuyết liên quan được sử dụng để rút ra các suy luận kiểm định về sự tin cậy của dữ liệu mô phỏng, đảm bảo dữ liệu mô phỏng mang tải được các tính chất của cấu trúc mô hình. Ưu thế của phương pháp nghiên cứu dựa trên mô phỏng là có thể khởi tạo và sử dụng số lượng rất lớn các chuỗi dữ liệu mô phỏng để thử nghiệm và thẩm định các thiết bị mới, các phát triển ứng dụng mới (trên thực tế hầu như không thể nào có được cỡ mẫu dữ liệu thực lớn được như vậy để sử dụng thử nghiệm, thẩm định). Cách thức khởi tạo được xem xét ở đây (và cũng là cách được sử dụng rất phổ biến) là sử dụng cấu trúc mô hình để mô phỏng lần lượt một số lượng lớn các chuỗi dữ liệu để phục vụ các mục đích nghiên cứu. Phương pháp tiếp cận này phù hợp với các ứng dụng cần phân tích thử nghiệm trên các chuỗi mô phỏng cần quan sát lặp lại và cần phải sử dụng nhiều lần những chuỗi dữ liệu có tính chu kỳ. Trên cơ sở sử dụng một số suy luận thống kê có liên quan, phạm vi bài báo đề cập đến hai kiểm định trên các chuỗi dữ liệu mô phỏng có tính lặp và tính chu kỳ. Các chuỗi mô phỏng được giả thiết là các chuỗi Markov bậc nhất và là các chuỗi Markov ổn định. Xuất phát từ mục đích nghiên cứu của các ứng dụng, kiểm định thứ nhất xem xét liệu các chuỗi mô phỏng có đảm bảo tính ổn định, có phù hợp của cấu trúc chuyển đổi của mô hình hay không và kiểm định thứ hai sẽ kiểm tra tính độc lập của chuỗi dữ liệu được khởi tạo.

2. MÔ HÌNH

Giới hạn bài báo là xem xét một mô hình ngẫu nhiên để khởi tạo các chuỗi Markov bậc nhất có không gian trạng thái $\Omega = \{1,2,\dots,m\}$. Đặt $t = 0,1,2,\dots,n$ là các chỉ số thời gian quan sát, $p_{ij}(t)$ là xác suất chuyển từ trạng thái i ở thời điểm $t-1$ đến trạng thái j ở thời điểm $t (i,j = 1,2,\dots,m; t = 0,1,2,\dots,n)$. Chuỗi Markov được giả thiết là có xác suất chuyển ổn định, tức là $p_{ij}(t) = p_{ij}$ với mọi $t = 0,1,\dots,n$.

Đặt $x_0, x_1, \dots, x_n$ là mẫu mô phỏng được khởi tạo từ mô hình với các xác suất chuyển p_{ij} và các xác suất ban đầu $p_i = \mathbb{P}(x_0 = i)$, $i,j = 1,2,\dots,m$. Nếu mẫu mô phỏng nhận dãy các giá trị trạng thái tương ứng là $i(0), i(1), \dots, i(n)$ thì xác suất đồng thời của mẫu cụ thể này được tính bởi $p_{i(0)}p_{i(0)i(1)}p_{i(1)i(2)} \cdots p_{i(n-1)i(n)}$.

Trong dãy $i(0), i(1), \dots, i(n)$, gọi $n_{ij}(t)$ là số lần chuyển từ trạng thái i tại thời điểm $t-1$ sang trạng thái j ở thời điểm $t (i,j = 1,2,\dots,m; t = 1,\dots,n)$. Đặt $n_{ij} = \sum_{t=1}^{n} n_{ij}(t)$, xác suất của chuỗi trạng thái này có thể được tính bởi

$$p_{i(0)}p_{i(0)i(1)}p_{i(1)i(2)} \cdots p_{i(n-1)i(n)} = p_{i(0)} \prod_{ij} p_{ij}^{n_{ij}}. \quad (2.1)$$

Ma trận $(n_{ij})_{m\times m}$ được gọi là ma trận tần số chuyển.

Với $k = 1,2,\dots,m$, đặt $n_k = \sum_{i=1}^{m} n_{ki}$ và $\tilde{n}_k = \sum_{i=1}^{m} n_{ik}$. Hai tổng này hầu như là như nhau, duy chỉ có trường hợp $k = i(0)$ hoặc $k = i(n)$ thì chúng lệch nhau một đơn vị, tức là $n_k - \tilde{n}_k = \delta_{ki(0)} - \delta_{ki(n)}$ $(k = 1,2,\dots,m)$ và lưu ý thêm

$$\sum_{ij} n_{ij} = \sum_i n_i = \sum_j \tilde{n}_j = n. \quad (2.2)$$

3. KIỂM ĐỊNH SỰ PHÙ HỢP CỦA MẪU MÔ PHỎNG VỚI CÁC XÁC SUẤT CHUYỂN

Theo các thành phần đã nêu trong mô hình, ứng với mỗi trạng thái $i = 1,2,\dots,m$ thì có thể xem

$(n_{i1}, n_{i2}, \dots, n_{im})$ là một bộ các tần số mẫu tương ứng với các xác suất $(p_{i1}, p_{i2}, \dots, p_{im})$ trong một mẫu quan sát có kích thước n_i, trong đó $n_i = \sum_j^m n_{ij}$ và $\sum_j^m p_{ij} = 1$.

Đặt $\eta_{ij} = \frac{n_{ij} - n_i p_{ij}}{\sqrt{n_i}}$ và $\eta_i = (\eta_{i1}, \eta_{i2}, \dots, \eta_{im})$, $i, j = 1,2, \dots, m.$ (3.1)

Xét theo các điều kiện thống kê thông thường với các quan sát mẫu độc lập, khi cỡ mẫu quan sát n lớn để đảm bảo các cỡ mẫu thành phần n_i đủ lớn, thì các véctơ ngẫu nhiên $\eta_i = (\eta_{i1}, \eta_{i2}, \dots, \eta_{im})$, $i = 1,2, \dots, m$, sẽ xấp xỉ phân phối chuẩn với cấu trúc hiệp phương sai $cov(\eta_{ij}, \eta_{il}) = \delta_{jl} p_{ij} - p_{ij} p_{il}$, trong đó $\delta_{jl} = 1$ nếu $j = l$ và $\delta_{jl} = 0$ nếu $j \neq l$.

Khẳng định này cũng đúng khi $x_0, x_1, \dots, x_n$ là một mẫu mô phỏng Markov. Điều này được thể hiện qua Định lý 3.1 dưới đây.

Định lý 3.1.Nếu $x_0, x_1, \dots, x_n$ là chuỗi Markov ổn định với các xác suất chuyển p_{ij}thì véctơ ngẫu nhiên $m \times m$ chiều $\eta = (\eta_{ij})_{i,j=1,2,\dots,m}$ (η_{ij} được xác định theo (3.1)) sẽ có phân phối hội tụ về phân phối chuẩn khi $n \to \infty$ với ma trận hiệp phương sai $(\lambda_{ij,kl})$, trong đó

$\lambda_{ij,kl} = \delta_{ik}(\delta_{jl} p_{ij} - p_{ij} p_{il})$, $i, j, k, l = 1,2, \dots, m.$ (3.2)

Đã có những phiên bản khác nhau chứng minh cho Định lý 3.1 mà tiêu biểu có thể tham khảo trong Billingsley (1961) hay trong Bartlett (1951).

Hệ quả 3.1 dưới đây được suy ra trực tiếp từ sự xấp xỉ phân phối chuẩn của $\eta = (\eta_{ij})_{i,j=1,2,\dots,m}$ trong Định lý 3.1.

Hệ quả 3.1. Nếu $x_0, x_1, \dots, x_n$ là chuỗi Markov ổn định với các xác suất chuyển $p_{ij} (i, j = 1,2, \dots, m)$ thì các biến ngẫu nhiên $\frac{n_{ij} - n_i p_{ij}}{\sqrt{n_i p_{ij}(1 - p_{ij})}}$sẽ xấp xỉ phân phối chuẩn tắc và do đó các thống kê dưới đây sẽ có phân phối χ^2 với $m - 1$ bậc tự do:

$$S_{1i} = \sum_{j=1}^m \frac{(n_{ij} - n_i p_{ij})^2}{n_i p_{ij}}, \quad i = 1,2, \dots, m. \quad (3.3)$$

Thực ra các số hạng trong tổng S_{1i}chỉ xuất hiện giới hạn trong điều kiện $p_{ij} > 0$ và nếu gọi d_i là số các xác suất p_{ij} $(j = 1,2, \dots, m)$ dương thì S_{1i} có phân phối χ^2 với $d_i - 1$ bậc tự do $(i = 1,2, \dots, m)$. Từ đây suy ra rằng thống kê dưới đây sẽ có phân phối χ^2 với $d - m$ bậc tự do:

$$S_1 = S_{11} + S_{12} + \dots + S_{1m} = \sum_{i,j=1}^m \frac{(n_{ij} - n_i p_{ij})^2}{n_i p_{ij}}, \quad (3.4)$$

trong đó $d = \sum_{i=1}^m d_i$. Thống kê này có thể ứng dụng để kiểm định giả thiết H_1 về "sự phù hợp của một chuỗi Markov $x_0, x_1, \dots, x_n$ với các xác suất chuyển $p_{ij} (i, j = 1,2, \dots, m)$".

4. KIỂM ĐỊNH TÍNH ĐỘC LẬP CỦA CHUỖI DỮ LIỆU MÔ PHỎNG

Một vấn đề đặt ra khi sử dụng thống kê (3.4) là không phải lúc nào cũng biết được các xác suất chuyển $p_{ij} (i, j = 1,2, \dots, m)$. Trong trường hợp chưa xác định được các xác suất chuyển này, giả thiết rằng các xác suất này phụ thuộc vào một tập gồm r tham số $\theta = (\theta_1, \theta_2, \dots, \theta_r)$ và được viết dưới dạng $p_{ij}(\theta)$ $(i, j = 1,2, \dots, m)$. Ước lượng hợp lý cực đại cho θ được thực hiện từ hàm log-likelihood của xác suất (2.1):

$$L(\theta) = p_{i(0)} \sum_{ij} n_{ij} \log p_{ij}(\theta). \quad (4.1)$$

Ước lượng $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_r)$ nhận được từ các phương trình

$$\frac{\partial L(\theta)}{\partial \theta_k} = 0 \Leftrightarrow \sum_{ij} \frac{n_{ij}}{p_{ij}(\theta)} \frac{\partial p_{ij}(\theta)}{\partial \theta_k} = 0, \quad k = 1,2, \dots, r. \quad (4.2)$$

Sự tồn tại xấp xỉ nghiệm của các phương trình này được chỉ ra ở Định lý 4.1 dưới đây.

Định lý 4.1. Giả thiết rằng các xác suất $p_{ij}(\theta)$ $(i, j = 1,2, \dots, m)$ có các đạo hàm riêng liên tục đến cấp 2 theo tham số $\theta \in \Theta$ và ma trận D cấp $m^2 \times r$ với các phần tử $\frac{\partial p_{ij}(\theta)}{\partial \theta_k} (i, j = 1,2, \dots, m, \ k = 1,2, \dots, r)$ có hạng r trên khắp tập Θ. Mẫu quan sát $x_0, x_1, \dots, x_n$ được giả thiết là chuỗi Markov được mô phỏng với các xác suất chuyển $p_{ij}(\theta)$ $(i, j = 1,2, \dots, m)$. Khi đó tồn tại ước lượng $\hat{\theta} \in \Theta$ được tính từ mẫu quan sát này là nghiệm của (4.2) sao cho $\hat{\theta}$ sẽ hội tụ theo xác suất về tập giá trị đúng θ.

Chứng minh của Định lý 4.1 có thể được dựa theo các điều kiện tối ưu thông thường qua các ứng dụng của đạo hàm đến cấp 2, như phương pháp đã dùng trong mục 30.6 của Cramér (1946); hoặc có thể chứng minh đơn giản hơn nếu giả thiết thêm rằng các xác suất $p_{ij}(\theta)$ $(i, j = 1,2, \dots, m)$ có đạo hàm liên tục đến cấp 3, như trong mục 7 của Billingsley (1961).

Trong trường hợp các điều kiện của Định lý 4.1 thỏa mãn và các xác suất $p_{ij}(\theta)$ $(i,j = 1,2,\dots,m)$ đều dương, cùng với Định lý 3.1 thì các thống kê dưới đây sẽ phân phối theo luật χ^2 với $m(m-1)-r$ bậc tự do:

$$S_2 = \sum_{i,j=1}^{m} \frac{(n_{ij}-n_i p_{ij}(\hat{\theta}))^2}{n_i p_{ij}(\hat{\theta})}. \quad (4.3)$$

Từ Định lý 4.1 có thể rút ra Hệ quả 4.1 dưới đây, có thể được áp dụng để kiểm định giả thiết H_2 rằng "một dãy dữ liệu mô phỏng $x_0, x_1, \dots, x_n$từ một cấu trúc Markov ổn định là một dãy độc lập". Tức là kiểm định rằng dãy mô phỏng có các xác suất chuyển $p_{ij} = p_j$không phụ vào trạng thái liền trước i $(i = 1,2,\dots,m)$ ở lần khởi tạo liền trước.

Hệ quả 4.1. Trong Định lý 4.1 lấy $r = m-1$, xét vector tham số $\theta = (\theta_1, \theta_2, \dots, \theta_{m-1})$ với các thành phần dương và có tổng nhỏ hơn 1. Đặt $p_{ij}(\theta) = \theta_j$ với $j < m$ và $p_{im}(\theta) = 1 - \sum_{j=1}^{m-1} \theta_j$. Khi đó các phương trình trong (4.2) có nghiệm là $\hat{\theta}_j = \frac{\tilde{n}_j}{n}, j = 1,2,\dots,m-1.$

Chứng minh. Nếu đặt $p_{im}(\theta) = \theta_m$thì các tham số phải thỏa mãn $\sum_{j=1}^{m} \theta_j = 1$ và hàm log-likelihood ở (4.1) được viết lại khi này bởi

$$L(\theta, \theta_m) = p_{i(0)} \sum_{ij} n_{ij} \log \theta_j, \quad j = 1,2,\dots,m. \quad (4.4)$$

Đặt λ là tham số nhân tử Lagrange, điều kiện cực đại sẽ được xác định theo hàm hỗ trợ dưới đây

$$l(\theta, \theta_m) = p_{i(0)} \sum_{ij} n_{ij} \log \theta_j + \lambda\left(\sum_{j=1}^{m} \theta_j - 1\right), \quad j = 1,2,\dots,m. \quad (4.5)$$

Điểm dừng cực đại được tìm từ hệ

$$\begin{cases} \frac{\sum_{i=1}^{m} n_{ij}}{\theta_j} + \lambda = 0 \\ \sum_{j=1}^{m} \theta_j - 1 = 0 \end{cases}. \quad (4.6)$$

Với lưu ý theo (2.2) là $\tilde{n}_j = \sum_{i=1}^{m} n_{ij}$ và $\sum_j \tilde{n}_j = n$, từ hệ (4.6) rút ra được $\theta_j = \frac{\tilde{n}_j}{n}, j = 1,2,\dots,m.$ ■

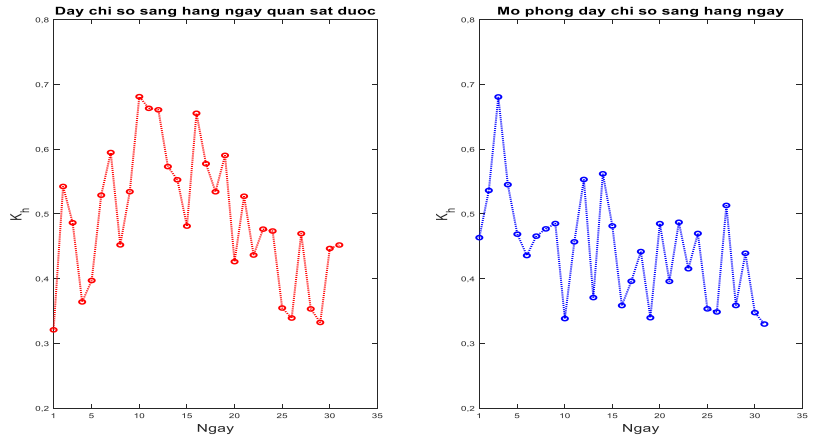
Từ Hệ quả 4.1, khi giả thiết H_2 đúng thì các xác suất chuyển $p_{ij} = p_j$ sẽ được ước lượng bởi $\hat{p}_j = \frac{\tilde{n}_j}{n}, j = 1,2,\dots,m$. Khi đó, thống kê S_2 được viết lại dưới dạng

$$S_2 = \sum_{i,j=1}^{m} \frac{(n_{ij}-n_i \tilde{n}_j/n)^2}{n_i \tilde{n}_j/n}, \quad (4.7)$$

và S_2 sẽ có phân phối χ^2 với $m(m-1)-(m-1) = (m-1)^2$ bậc tự do. Thống kê này có thể được ứng dụng để kiểm định giả thiết H_2.

5. ÁP DỤNG

Mô hình áp dụng ở đây là mô hình Markov ẩn được dùng để mô hình hóa dãy chỉ số sáng hàng ngày đã được giới thiệu bởi Lý (2016). Các nghiên cứu liên quan đã chỉ ra rằng các dãy chỉ số sáng ở một vị trí địa lý có phân phối xác suất khác nhau ở những tháng khác nhau trong năm. Nếu không có những biến động khí tượng đặc biệt gì thì phân phối chỉ số sáng của một tháng nào đó trong năm có đặc tính mùa vụ, lặp lại giống nhau hàng năm. Để có thể khảo sát, ước lượng được tốt sự phân phối chỉ số sáng của một tháng cụ thể trong năm thì cần có một số lượng rất lớn những dãy dữ liệu quan sát của tháng này. Điều này đòi hỏi phải có thời gian chờ thu thập mẫu quan sát hàng chục, hàng trăm năm và rõ ràng là rất khó có thể có được đối với những nghiên cứu ở mức độ thông thường.



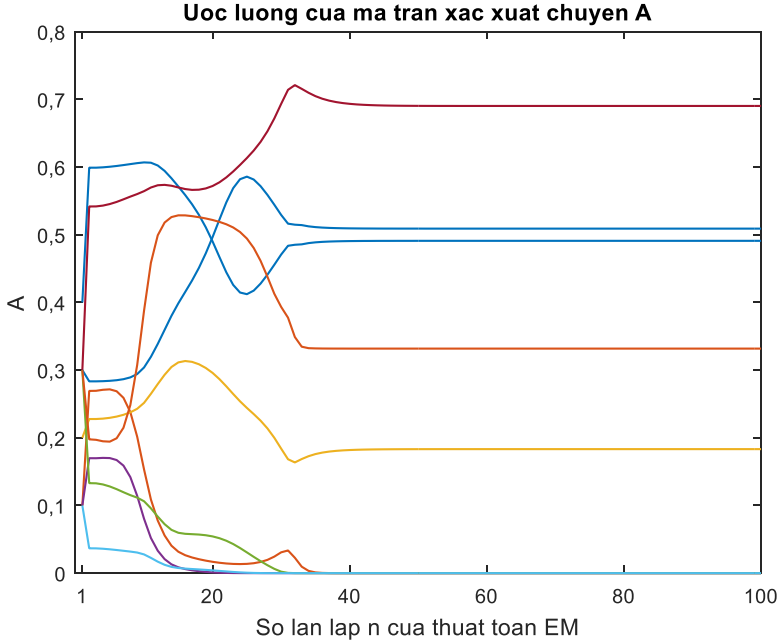
Hình 1. Dãy dữ liệu quan sát (a) và một dãy dữ liệu mô phỏng (b) của chỉ số sáng hàng ngày của tháng 7 năm 2021 ở thành phố Cần Thơ

Áp dụng cho mô hình Markov ẩn 4 trạng thái được đề xuất bởi Lý (2016), dãy dữ liệu quan sát chỉ số sáng hàng ngày của tháng 7 năm 2021 được cung cấp bởi Trạm khí tượng thành phố Cần Thơ (Hình 1(a)) được sử dụng, các véctơ trung bình μ, độ lệch chuẩn σ và ma trận xác suất chuyển A nhận được sau 100 bước lặp của thuật toán EM (Expectation Maximization) lần lượt được trình bày dưới đây:

$$\mu = (0{,}5403;\ 0{,}6652;\ 0{,}3437;\ 0{,}4669),$$

$$\sigma = (0{,}0492;\ 0{,}0096;\ 0{,}0147;\ 0{,}0426),$$

$$A = \begin{pmatrix} 0{,}6902 & 0{,}1832 & 0 & 0{,}1266 \\ 0{,}4910 & 0{,}5090 & 0 & 0 \\ 0 & 0 & 0{,}3317 & 0{,}6683 \\ 0{,}1353 & 0 & 0{,}2930 & 0{,}5717 \end{pmatrix}.$$



Hình 2. Tiến trình ước lượng của ma trận xác suất chuyển *A*

Hình 2 thể hiện tiến trình ước lượng để thu được các phần tử của ma trận xác suất chuyển A qua 100 bước lặp của thuật toán EM.

Mô hình Markov được sử dụng với các tham số ước lượng được, một số lượng lớn các dãy mô phỏng có cùng phân phối chu kỳ với dữ liệu quan sát đã được khởi tạo để ước lượng hàm mật độ xác suất của chỉ số sáng một số tháng trong năm ở thành phố Cần Thơ. Tuy nhiên, ứng dụng này của Tran (2016) đã không thực hiện kiểm định tin cậy các dãy mô phỏng phù hợp với cấu trúc chuyển đổi phân phối xác suất của mô hình. Điều này sẽ được hiện minh họa dưới đây.

Với các tham số nhận được ở trên, một dãy $n = 31$ giá trị mô phỏng cho dãy chỉ số sáng của tháng 7 ở thành phố Cần Thơ đã được khởi tạo (Hình 1(b)). Với mức ý nghĩa $\alpha = 0{,}05$ dãy mô phỏng này sẽ được sử dụng để kiểm định hai giả thiết sau đây:

H_1: "Dãy dữ liệu mô phỏng phù hợp với các xác suất chuyển trong ma trận A", tức là có thể sử dụng mô hình để tạo ra các dãy dữ liệu mô phỏng tin cậy rằng chúng có cấu trúc chuyển đổi như phân phối thực tế của dữ liệu thực. Kiểm định này được thực hiện dựa vào thống kê $\chi^2 = S_1$ (theo (3.4)) có phân phối χ^2 với $d - m = 6$ bậc tự do (trong đó $d = 10$ là số các xác suất chuyển dương trong ma trận A, $m = 4$ là số trạng thái của mô hình). Giá trị tới hạn của kiểm định là $\chi^2_{0{,}05} = 11{,}0705$.

H_2: "Dãy gồm các dữ liệu được mô phỏng độc lập". Thống kê có phân phối χ^2 với $(m-1)^2 = 9$ bậc tự do $\chi^2 = S_2$ (trong (4.3)) sẽ được sử dụng trong kiểm định này. Giá trị tới hạn của kiểm định là $\chi^2_{0{,}05} = 16{,}9190$.

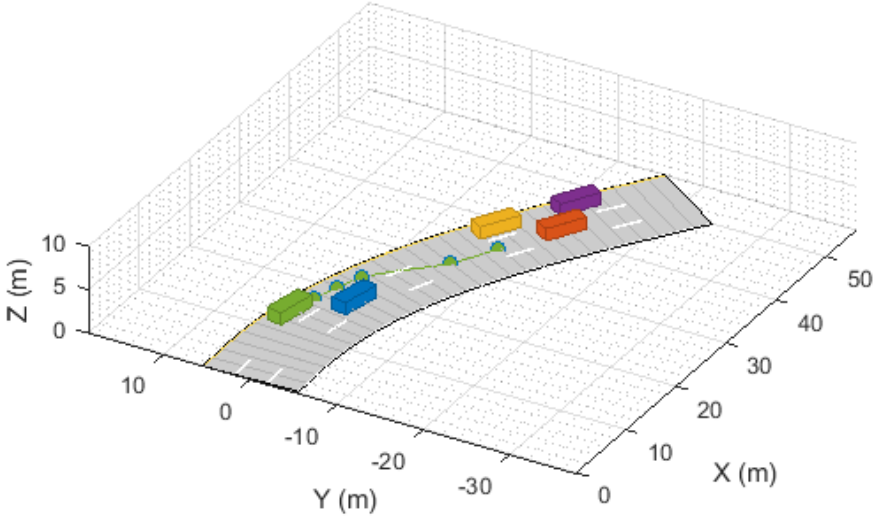
Các kết quả kiểm định được trình bày ở Bảng 1. Kết quả này cho thấy dãy mô phỏng là phù hợp với cấu trúc mô hình. Giả thiết H_2 bị bác bỏ có nghĩa là các khởi tạo trong mẫu mô phỏng là có mối liên hệ chuyển đổi phụ thuộc nhau (theo cấu trúc phụ thuộc của chuỗi Markov); còn việc chấp nhận giả thiết H_1 cho thấy là sự chuyển đổi này là phù hợp với cấu

trúc xác suất chuyển đổi của mô hình Markov được thiết lập. Những kết quả phân tích suy luận này cho phép nói rằng chuỗi mô phỏng là tin cậy, đảm bảo mang tải được các tính chất phân phối xác suất cấu trúc của mô hình; từ đó có thể tin cậy sử dụng các dãy dữ liệu mô phỏng để ước lượng hàm mật độ xác suất của chỉ số sáng ở một số tháng được quan sát.

Bảng 1. Kết quả kiểm định tin cậy ($\alpha = 0,05$) các giả thiết

Giả thiết kiểm định	Giá trị thống kê χ^2	Giá trị tới hạn $\chi^2_{0,05}$	Kết luận
H_1	4,9743	**11,0705**	Chấp nhận
H_2	23,0090	**16,9190**	Bác bỏ

Trong kết quả ứng dụng trên, giả thiết H_2 cho rằng "Dãy gồm các dữ liệu được mô phỏng độc lập" bị bác bỏ là kết quả có thể đoán được. Bởi vì lưu ý rằng cả hai kiểm định đã nêu đều xét trên giả thiết chuỗi mô phỏng là chuỗi Markov bậc nhất, mà thông thường thì ở chuỗi Markov bậc nhất các lần khởi tạo liền kề là có phụ thuộc nhau. Tuy nhiên, chuỗi Markov bậc nhất được hiểu là chuỗi mà lần khởi tạo hiện tại *chỉ có thể* phụ thuộc vào lần khởi tạo liền trước, *tất cả* các lần khởi tạo trước đó nữa đều không liên quan. Theo nghĩa như vậy, ở một mẫu mô phỏng khởi tạo từ chuỗi Markov bậc nhất thì lần khởi tạo hiện tại cũng có thể "mất trí nhớ" (độc lập hay "độc lập nhẹ") luôn cả với lần khởi tạo liền trước. Khi đó giả thiết H_2 có khả năng được chấp nhận. Trường hợp này cần xem xét tiếp theo là chuỗi Markov chỉ "độc lập nhẹ" hay là độc lập hoàn toàn (có thể xem là một "chuỗi Markov suy biến", mất trí nhớ hoàn toàn). Đây là trường hợp mà nhiều ứng dụng rất quan tâm. Ví dụ, chuỗi mô phỏng tạo ra từ thuật toán Gibbs quét tuần tự (Lý và ctv., 2020) là chuỗi Markov bậc nhất và như vậy, theo tính chất của phép tạo mẫu Gibbs quét tuần tự, chuỗi dữ liệu được khởi tạo là không độc lập. Khi ứng dụng thuật toán này để khởi tạo các tham số trong hoạt cảnh ban đầu của những trường hợp thử nghiệm ADAS (Advanced Driver Assistance Systems) dùng cho xe tự hành trình trong tình huống có một xe phía trước thực hiện chuyển làn đột ngột (Hình 3), yêu cầu đặt ra là các khởi tạo ban đầu này phải độc lập với nhau. Các tình huống này được mô phỏng lặp lại nhiều lần để thử nghiệm nên những dãy dữ liệu khởi tạo có tính lặp. Để các tình huống thử nghiệm là độc lập thì các khởi tạo ban đầu phải độc lập. Giải pháp đưa ra khi này là xây dựng trường ngẫu nhiên Markov với các nhóm tham số độc lập cực đại và sử dụng thuật toán Gibbs tuần tự để khởi tạo mẫu mô phỏng cho các tham số ban đầu trên những nhóm tham số cực đại này. Sau đó, giả thiết H_2 sẽ được kiểm định để đảm bảo tin cậy rằng tính độc lập được thỏa mãn theo yêu cầu đặt ra.



Hình 3. Mô phỏng tình huống chuyển làn xảy ra trước xe tự hành trình

6. KẾT LUẬN

Các suy luận kiểm định tin cậy như đã xem xét trong bài viết là rất cần thiết trước khi sử dụng các dãy dữ liệu mô phỏng trong các ứng dụng công nghiệp. Ngoài hai dạng kiểm định được nêu ở đây, còn có nhiều dạng kiểm định tin cậy khác cũng cần xem xét chặt chẽ khi sử dụng ứng dụng đối với các

loại dãy dữ liệu mô phỏng nói chung. Đặc biệt là những suy luận tin cậy trên các chuỗi mô phỏng Markov, chẳng hạn như kiểm định về bậc Markov của chuỗi dữ liệu, kiểm định các dãy mô phỏng độc lập nhau, kiểm định rằng chuỗi Markov có nhiều hơn một tập trạng thái, … là những dạng có thể được quan tâm nghiên cứu mở rộng.

LỜI CẢM TẠ

Các tác giả xin trân trọng cảm ơn Trung tâm Khí tượng thành phố Cần Thơ đã cung cấp dữ liệu thực nghiệm sử dụng trong bài báo; trân trọng cảm ơn sự tài trợ của Đại học Mines-ParisTech thông qua dự án EUFRA00121NCTN; trân trọng cảm ơn sự tài trợ của Trường Đại học Cần Thơ thông qua đề tài cấp Trường của sinh viên mã số TSV2021-59.

TÀI LIỆU THAM KHẢO

Anderson, T. W. & Goodman, L. A. (1957). Statistical inference about Markov Chains. *The Annals of Mathematical Statistics, 28*(1), 89-110. https://doi.org/10.1214/aoms/1177707039

Bartlett, M. S. (1951). The frequency goodiness of fit test for probability chain. *Proc. Camb. Phil. Soc., 47*, 86-95. https://doi.org/10.1017/S0305004100026402

Billingsley, P. (1961). Statistical Methods in Markov Chains. *The Annals of Mathematical Statistics, 32*(1), 12-40. https://doi.org/10.1214/aoms/1177705136

Cramér, H. (1946). Mathematical Methods of Statistics. *Princeton University Press*. https://doi.org/10.1515/9781400883868

Elliott, J. R., Aggoun, L. & Moore, J. B. (2010). *Hidden Markov Models: Estimation and control*, Springer.

Levine, R. A., & Casella, G. (2006). Optimizing random scan gibbs samplers. *Journal of Multivariate Analysis*, 97, 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

Lý, T. V. (2016). Stochastic modeling for daily clearness index sequence in can tho city. *Tạp chí Khoa học Trường Đại học Cần Thơ, 2*, 90-99. https://doi.org/10.22144/ctu.jen.2016.005

Lý, T. V., Thịnh, N. T., Phú, N. D. T., Phô, T. Đ., & Trọng, T. V. (2020). Sử dụng thuật toán entropy chéo và chọn mẫu gibbs để ước lượng xác suất sự kiện hiếm. *Tạp chí Khoa học Trường Đại học Cần Thơ, 56* (Số CĐ Tự nhiên), 46-53. https://doi.org/10.22144/ctu.jsi.2020.092