

Nghiên cứu khả năng ứng dụng phần mềm R trong thống kê và dự báo nhu cầu vận tải

■ **ThS. TRẦN NGUYỄN HÀ^(*); ThS. NGUYỄN TRUNG ANH; KS. ĐỖ ĐỨC THẮNG**

Viện Chiến lược và Phát triển Giao thông vận tải (Bộ Giao thông vận tải)

Email: ^(*)hatn.tdsi@gmail.com

TÓM TẮT: Hiện nay, việc dự báo nhu cầu vận tải chủ yếu dựa trên các phần mềm dự báo nhu cầu giao thông như JICA STRADA, VISUM, VISIM. Tuy nhiên, việc mô hình hóa được mạng giao thông cần rất nhiều công sức và nỗ lực từ khảo sát lưu lượng, nhu cầu hiện tại, thiết lập mạng O-D, tính toán phân bổ lưu lượng cho các phương thức, điều chỉnh theo các điều kiện thực tế. Nghiên cứu giới thiệu tổng quan về phần mềm R cũng như những kinh nghiệm ứng dụng phần mềm R trong phân tích dữ liệu giao thông trên thế giới, từ đó trình bày quy trình tính toán và ứng dụng tính toán khối lượng vận tải hành khách đường bộ bằng phần mềm R. Cuối cùng, nghiên cứu đánh giá độ tin cậy cũng như những khả năng ứng dụng phần mềm R phối hợp với các phần mềm dự báo truyền thống để đạt được những hiệu quả nhất định như làm sạch dữ liệu đầu vào, rút ngắn thời gian dự báo trong những trường hợp cụ thể.

TỪ KHÓA: Phần mềm R, dự báo nhu cầu vận tải, phân tích dữ liệu, vận tải hành khách.

ABSTRACT: Currently, transportation demand forecasting is mainly based on traffic demand forecasting software such as JICA STRADA, VISUM, VISIM. However, modeling a traffic network requires a lot of effort from traffic survey, current demand, O-D network setup, calculating traffic distribution for modes, adjusting according to actual conditions. The study introduces an overview of R software as well as experiences in applying R software in analyzing traffic data worldwide. Since, the calculation process and application of calculating the volume of road passenger transport using R software is presented. Finally, the study evaluates the reliability as well as the possibilities of applying R software in coordination with traditional forecasting software to achieve certain effects such as filtering input data and shortening forecast time in specific cases.

KEYWORDS: R software, travel demand forecasting, data analysis, passenger transport.

1. ĐẶT VẤN ĐỀ

Trong bối cảnh nhu cầu vận tải ngày càng tăng cao, các phương thức vận tải (đường bộ, đường sắt, đường thủy nội địa, hàng hải, hàng không) được đa dạng hóa các loại hình vận tải. Hiện nay, tình trạng một số số liệu thu thập có mức độ biến động và độ sai lệch tương đối cao, nếu không có các công cụ phân tích chuyên sâu sẽ ảnh hưởng đến việc đánh giá mức độ tin cậy của dữ liệu, có hay không sử dụng mẫu số liệu đó trong phân tích, tính toán và dự báo.

Khoa học thống kê và phân tích dữ liệu với dữ liệu lớn trên thế giới đã và đang rất phát triển với nhiều phần mềm và mô hình hiện đại như SAS, SPSS, Stata, Statistica, R... Trong bối cảnh hiện nay chưa có một nghiên cứu chuyên sâu về ứng dụng phần mềm R và các phân tích thống kê phù hợp trong phân tích, đánh giá dữ liệu và dự báo nhu cầu vận tải hành khách, hàng hóa ngành GTVT. Vì vậy, việc nghiên cứu khả năng ứng dụng phần mềm R trong thống kê và dự báo nhu cầu vận tải là rất cần thiết.

2. TỔNG QUAN VỀ PHẦN MỀM R VÀ CÁC ỨNG DỤNG TRONG LĨNH VỰC GTVT

2.1. Tổng quan phần mềm R và một số ứng dụng trên thế giới

Phần mềm R là một ngôn ngữ lập trình hàm cấp cao, vừa là một môi trường dành cho tính toán thống kê. R hỗ trợ rất nhiều công cụ cho phân tích dữ liệu. Có thể coi R là công cụ phân tích thống kê học cũng như phân tích dữ liệu nói chung.

Hiện nay, phần mềm R được ứng dụng rộng rãi trong lĩnh vực học tập và nghiên cứu về Khoa học dữ liệu (Data Science) và Quản lý dự án (Project Management). Phần mềm R cung cấp nhiều gói đồ họa mạnh mẽ, giúp người dùng hiểu rõ hơn về cách biểu đồ có thể cung cấp thông tin và trực quan hóa dữ liệu một cách hiệu quả. Phần mềm R có thể được sử dụng để phân tích và dự đoán xu hướng tài chính, xử lý dữ liệu tài chính và xây dựng mô hình dự báo.

Cho đến thời điểm hiện nay đã có một số dự án và nghiên cứu trong lĩnh vực GTVT đã sử dụng phần mềm R để phân tích dữ liệu và thực hiện các mô hình học máy. Dưới đây là một số ví dụ nổi bật về việc ứng dụng R:

- Dự án “Mô hình dự đoán giao thông” Portland, Oregon, Hoa Kỳ: Sử dụng R để xây dựng mô hình dự đoán tình trạng giao thông trên các tuyến đường cụ thể.

- Dự án “Phân tích dữ liệu vận tải” (Transportation Data Analytics) - Cơ quan quản lý GTVT công cộng TP. London: Sử dụng R để phân tích dữ liệu nhằm cải thiện hệ thống giao thông công cộng của London.

2.2. Đánh giá khả năng ứng dụng của phần mềm R trong GTVT tại Việt Nam

- *Ứng dụng phần mềm R trong phân tích dữ liệu GTVT:*

Sử dụng R trong phân tích và xử lý dữ liệu thống kê vận tải là một lựa chọn tốt nhất cho các nhà nghiên cứu và chuyên gia. R là một ngôn ngữ lập trình và môi trường phần mềm miễn phí, được sử dụng rộng rãi trong phân tích thống kê, khai phá dữ liệu và trực quan hóa dữ liệu.

Một trong những ưu điểm của R là có thể sử dụng trên nhiều hệ điều hành khác nhau, bao gồm Windows, Mac OS và Linux. Ngoài ra, R cũng có thể được sử dụng để xử lý các tập tin dữ liệu từ nhiều nguồn khác nhau, hay sử dụng để xử lý các tập tin dữ liệu lớn (Big Data), giúp cho việc phân tích dữ liệu, dự báo nhu cầu vận tải trở nên dễ dàng hơn. Điểm nổi bật nhất của R là việc trực quan hóa dữ liệu bằng cách sử dụng các gói phần mềm như ‘ggplot2’ và ‘lattice’. Các gói này cung cấp cho người dùng các công cụ để tạo ra biểu đồ chất lượng cao và trực quan. R có nhiều ưu điểm trong việc xử lý dữ liệu lớn, dưới đây là một số ưu điểm nổi bật.

Xử lý dữ liệu lớn	• R cung cấp nhiều gói phần mềm và thư viện cho phép xử lý dữ liệu lớn và phân tích dữ liệu hiệu quả
Tự động hóa	• R có khả năng tự động hóa các tác vụ phức tạp, giúp tiết kiệm thời gian và giảm thiểu sai sót R có khả năng tự động hóa các tác vụ phức tạp, giúp tiết kiệm thời gian và giảm thiểu sai sót
Tính toán nhanh	• R có khả năng tính toán nhanh chóng, đặc biệt là khi xử lý các tập tin dữ liệu lớn
Đọc mọi loại dữ liệu	• R có khả năng đọc mọi loại dữ liệu, bao gồm cả các tập tin dữ liệu không được định dạng chuẩn
Hỗ trợ tập tin dữ liệu lớn	• R hỗ trợ các tập tin dữ liệu lớn
Khả năng tái sử dụng	• Mã nguồn của R có thể được sử dụng lại với các tập tin dữ liệu khác nhau

Hình 2.1: Ưu điểm của R trong việc xử lý Big Data (Dữ liệu lớn)

R được sử dụng để thực hiện các tác vụ phân tích dữ liệu, khai thác dữ liệu và trực quan hóa dữ liệu. Ngoài ra, R cũng có thể được sử dụng để xây dựng các mô hình dự đoán và mô hình học máy. Đặc biệt, R có thể lọc các dữ liệu đầu vào dựa trên việc phát hiện giá trị ngoại lai và thực hiện các thống kê cơ bản từ các mô hình tính toán mà người dùng tự xây dựng. Với những ưu điểm trên, R có thể coi là bước đầu tiên để ứng dụng Big Data trong phân tích dữ liệu, dự báo tổng nhu cầu vận tải.

- *Ứng dụng phần mềm R trong dự báo nhu cầu vận tải:*

Để dự báo lưu lượng giao thông và phân bổ lưu lượng giao thông cho từng loại phương thức vận tải, thông thường người dùng sử dụng các mô hình mô phỏng và dự báo giao thông như mô hình 4 bước, mô hình Distance-based hoặc Activity-based cùng những biến thể và điều chỉnh của

nó. Nó rất phù hợp và cần thiết để xây dựng những báo cáo quy hoạch GTVT cần mức độ chuyên sâu và chính xác cao. Tuy nhiên, đôi khi những quyết định, báo cáo cần được thực hiện trong thời gian rất ngắn, mức độ chính xác có thể không cần quá cao. Khi đó, việc dự báo theo chuỗi thời gian (Time series) của phần mềm R trở thành một lựa chọn hữu hiệu để chỉ ra xu hướng phát triển nhất định của đối tượng.

Mặt khác, R đóng vai trò là một công cụ để tham chiếu, thẩm tra các dự báo nhu cầu vận tải khác nhau. Từ cơ sở dự báo theo chuỗi thời gian (Time series), các kết quả dự báo đều phải tuân theo một xu hướng phát triển nhất định.

Phần mềm R là một công cụ mạnh mẽ và linh hoạt, thích hợp cho việc dự báo nhu cầu quy hoạch giao thông vận tải nhờ khả năng làm việc với lượng dữ liệu lớn hơn nhiều so với các phần mềm khác. Việc sử dụng phần mềm R trong việc dự báo nhu cầu quy hoạch GTVT không chỉ giúp có được những dự báo chính xác, mà còn giúp người dùng kiểm soát được dữ liệu và xu hướng trong GTVT.

3. DỰ BÁO NHU CẦU VẬN TẢI SỬ DỤNG PHẦN MỀM R

3.1. Quy trình phân tích và dự báo nhu cầu vận tải

- *Bước 1:* Cài đặt các thư viện cần có để xử lý dữ liệu chuỗi thời gian.

- *Bước 2:* Nhập dữ liệu từ bảng theo cú pháp, kiểm tra thông số cơ bản của dữ liệu. Kết quả trả về là số lượng chuỗi thời gian, sẽ cho biết có tất cả bao nhiêu quan sát và bao nhiêu biến.

Câu lệnh 1: `Tên_tập_dữ_liệu <- read_excel` (đường link dẫn đến tập dữ liệu)

`View(Tên_tập_dữ_liệu)`

Câu lệnh 2: `dim(Tên_tập_dữ_liệu)`

- *Bước 3:* Tiến hành các thống kê cơ bản. Tiến hành kiểm tra dạng dữ liệu của các biến bằng lệnh “str”, các dữ liệu dưới dạng số học sẽ được dùng để tính toán. Kiểm tra các thông tin chung của các trường dữ liệu bằng lệnh “summary”. Kiểm tra thống kê mô tả các biến bằng lệnh “describe”. Phát hiện giá trị ngoại lai bằng lệnh “tsoutliers”.

- *Bước 4:* Kiểm tra mức độ tương quan của các biến dạng số học.

Hệ số tương quan là một thước đo thống kê về độ mạnh yếu tuyến tính của mối quan hệ chuyển động tương đối giữa hai biến.

Câu lệnh 1: `vars=cbind(Tên_biến_1,Tên_biến_2,...,Tên_biến_n)`

Câu lệnh 2: `pair.panel(vars)`.

- *Bước 5:* Dự đoán mô hình tuyến tính đa biến.

Mô hình tuyến tính đa biến là một phương trình mô tả mối quan hệ giữa biến phụ thuộc với các biến độc lập $X_1, X_2, \dots, X_n$ và số hạng sai số e . Cụ thể, mô hình này có dạng: $y = B_0 + B_1 \cdot X_1 + B_2 \cdot X_2 + \dots + B_n \cdot X_n + e$. Trong đó: y là biến phụ thuộc; $X_1, X_2, \dots, X_n$ là các biến độc lập; $B_0, B_1, B_2, \dots, B_n$ là các tham số của mô hình; e là biến ngẫu nhiên gọi là số hạng sai số.

Sử dụng mô hình tuyến tính đa biến sẽ hỗ trợ cho việc phân tích dữ liệu sơ bộ và dự đoán các nhu cầu vận tải trong tương lai. Mô hình tuyến tính đa biến có dạng công thức toán học cơ bản, dễ hiểu, dễ giải thích và dễ áp dụng vào tính toán, đưa ra các dự đoán.

Câu lệnh 1: `REGdata <- data.frame (Tên_biến_độc_lập_1, Tên_biến_độc_lập_2,..., Tên_biến_độc_lập_n)`

Câu lệnh 2: `reg <- lm(Tên_biến_phụ_thuộc~.,data=REGdata)`

Câu lệnh 3: `step(reg, direction = "both")`

Kết quả đưa ra mô hình nào có giá trị AIC thấp nhất là tốt nhất, sau đó tiếp tục kiểm định lại R2 của mô hình để xuất.

Câu lệnh 1: `check=lm(Tên_biến_phụ_thuộc ~ Tên_biến_độc_lập_để_xuất)`

Câu lệnh 2: `summary(check)`

- **Bước 6:** Kiểm định mô hình. Mô hình để xuất là phương trình tuyến tính đơn biến, nên cần kiểm định các giả thiết sau:

+ Tính tuyến tính của dữ liệu: Kiểm định tính tuyến tính trong thống kê thường liên quan đến việc kiểm tra xem mối quan hệ giữa các biến có phải là tuyến tính hay không. Câu lệnh: `plot(check,1)`.

Nếu đường màu đỏ thẳng => Dữ liệu tuyến tính và đạt yêu cầu kiểm định. Nếu đường màu đỏ không thẳng => Vi phạm giả thiết kiểm định, cần phải tiếp tục điều chỉnh mô hình.

+ Phần dư có phân phối chuẩn: Phân phối chuẩn là một phân bố xác suất quan trọng nhất trong thống kê vì nó mô tả chính xác sự phân bố các giá trị của nhiều hiện tượng tự nhiên. Câu lệnh: `shapiro.test(res)`.

Nếu p value > 0,05 là giả thiết không bị vi phạm. Nếu p value < 0,05 là giả thiết phần dư có phân phối chuẩn bị vi phạm. Khi giả thiết phần dư có phân phối chuẩn bị vi phạm cần phải tiếp tục kiểm tra, điều chỉnh lại mô hình.

+ Phần dư có trung bình bằng 0: Một trong những giả định quan trọng của hồi quy tuyến tính là phần dư có trung bình bằng 0. Điều này có nghĩa là tổng của tất cả các phần dư phải bằng 0. Câu lệnh: `t.test(res, mu=0)`.

Nếu p-value > 0,05 thì phần dư có trung bình bằng 0; Nếu p-value < 0,05 thì phần dư có trung bình không bằng 0, điều này chỉ ra rằng mô hình hồi quy có thể không phù hợp với dữ liệu.

+ Phương sai của phần dư là không đổi:

Kiểm định phương sai của phần dư trong thống kê là một quy trình quan trọng để xác định xem phương sai của các phần dư (sai số) có đồng nhất hay không. Câu lệnh: `ncvTest(check)`.

- **Bước 7:** Sử dụng mô hình ARIMA dự đoán ngắn hạn.

Mô hình ARIMA (Auto Regressive Integrate Moving Average), hay Mô hình Tự hồi quy tích hợp Trung bình Trượt, cho phép phân tích và mô phỏng các chuỗi thời gian dựa trên các quan sát trước đó.

Câu lệnh 1: `Tên.đồ.thị <- ts(Tên_biến, frequency = 1, start = (1990))`

Câu lệnh 2: `plot.ts(Tên.đồ.thị)`

Dựa vào đồ thị đưa ra nhận xét, từ đó xây dựng các mô hình ARIMA tương ứng với các trường hợp (có xét đến giá trị ngoại lai, không xét đến giá trị ngoại lai...). Dữ liệu ngoại lai (outlier) là những giá trị dữ liệu (records) được ghi nhận có sự khác biệt bất thường so với những giá trị dữ liệu khác.

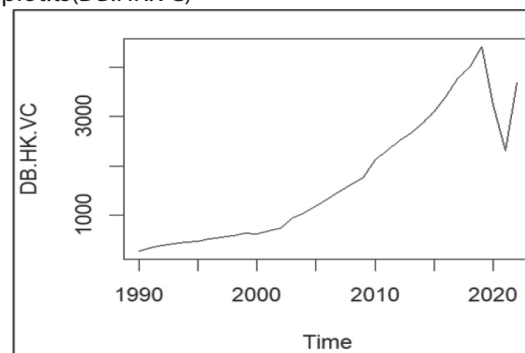
3.2. Dự báo nhu cầu vận tải hành khách bằng đường bộ

- **Xử lý dữ liệu vận tải hành khách đường bộ cơ bản:** Do đối tượng dự báo là khối lượng vận tải hành khách bằng đường bộ, cần phải gán giá trị các dữ liệu vào các thời điểm tính toán và vẽ đồ thị quan sát, cú pháp câu lệnh như sau:

`DB.HK.VC<-ts(DBo_HK_RawData$Hkvan.DB.HK, frequency = 1, start = (1990))`

`DB.HK.VC`

`plot.ts(DB.HK.VC)`



Hình 3.1: Khối lượng vận tải hành khách đường bộ có xét đến giá trị ngoại lai theo thời gian

Từ biểu đồ trên, có thể thấy dạng đồ thị có xu hướng tăng nhưng xuất hiện những điểm ngoại lai lớn trong giai đoạn 2020 - 2022. Nhằm đảm bảo tính khách quan, cần xét đến 2 kịch bản bao gồm kịch bản có xét đến yếu tố ngoại lai và kịch bản không xét đến yếu tố ngoại lai. Từ cơ sở trên, so sánh số liệu dự báo với dự báo chính thức đã được công nhận về nhu cầu vận tải hành khách tới năm 2030.

- **Xây dựng mô hình ARIMA có xét đến giá trị ngoại lai:** Tại bước này, người dùng xây dựng mô hình ARIMA (p,d,q) tự động với 'p' là phân tích tương quan liên hoàn, 'd' là bậc của sai phân, 'q' là phân tích biểu đồ tương quan liên hoàn từng phần.

Đầu tiên, cần tìm mô hình ARIMA phù hợp với cú pháp câu lệnh:

`auto.arima(DB.HK.VC)`

```
Series: DB.HK.VC
ARIMA(2,1,2) with drift

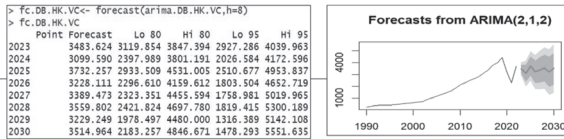
Coefficients:
      ar1      ar2      ma1      ma2      drift
    -0.0074  -0.3742  0.3751  -0.4661  108.8061
s.e.    0.3990   0.2977  0.3724   0.3585   36.5854

sigma^2 = 97693:  log likelihood = -227.61
AIC=467.23  AICC=470.59  BIC=476.02
```

Hình 3.2: Lựa chọn mô hình ARIMA có xét đến giá trị ngoại lai bằng R

Phần mềm cho biết mô hình ARIMA(2,1,2) được lựa chọn để dự báo với số bậc sai phần là 1. Tiếp theo, sử dụng mô hình ARIMA để dự báo nhu cầu vận tải đến năm 2023, cú pháp câu lệnh như sau:

```
arima.DB.HK.VC=arima(DB.HK.VC, order=c(2,1,2))
fc.DB.HK.VC<- forecast(arima.DB.HK.VC,h=8)
fc.DB.HK.VC
plot(fc.DB.HK.VC)
```



Hình 3.3: Dự báo khối lượng vận tải hành khách đường bộ có xét đến giá trị ngoại lai

- Xây dựng mô hình ARIMA không xét đến giá trị ngoại lai: Nhập lại chuỗi dữ liệu sau khi bỏ đi các giá trị ngoại lai giai đoạn 2020 - 2022, cú pháp câu lệnh như sau:

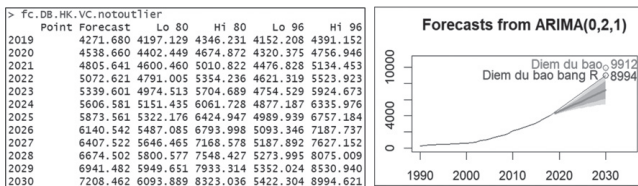
```
DB.HK.VC.notoutlier<-ts(DBo_HK_RawData$Hkvan,
DB.HK,frequency = 1, start = (1990),end=(2018))
```

```
DB.HK.VC.notoutlier
plot(DB.HK.VC.notoutlier)
```

Lựa chọn mô hình ARIMA phù hợp với câu lệnh:
auto.arima(DB.HK.VC.notoutlier)

Phần mềm cho biết mô hình ARIMA(0,2,1) được lựa chọn để dự báo với số bậc sai phần là 2. Tiếp theo, sử dụng mô hình ARIMA để dự báo nhu cầu vận tải đến năm 2023, cú pháp câu lệnh như sau:

```
arima.DB.HK.VC.notoutlier=arima(DB.HK.VC.
notoutlier, order=c(2,1,2))
fc.DB.HK.VC.notoutlier <- forecast(arima.DB.HK.
VC.notoutlier,level = c(80,96), h=12)
fc.DB.HK.VC.notoutlier
plot(fc.DB.HK.VC.notoutlier)
```



Hình 3.4: Dự báo khối lượng vận tải hành khách đường bộ có xét đến giá trị ngoại lai

Trên đồ thị, 96% giá trị của khối lượng hành khách vận chuyển bằng đường bộ sẽ đi theo đường nét đứt và đến năm 2030 sẽ đạt giá trị 8.994 triệu hành khách. Trong khi đó, giá trị dự báo hiện nay là 9.912 triệu, với mức chênh lệch so với tính toán là 9,3%. Sự chênh lệch này hoàn toàn có thể chấp nhận được do thời gian dự báo dài và phần mềm thống kê R chỉ sử dụng dữ liệu quá khứ để dự đoán, không xét đến các yếu tố tăng trưởng trong tương lai.

4. KẾT LUẬN

Dự báo GTVT là một trong những công tác quan trọng

trong xây dựng chiến lược, quy hoạch, để án phát triển nhằm xác định mục tiêu, định hướng phát triển kết cấu hạ tầng giao thông đáp ứng nhu cầu vận tải.

Trong bối cảnh việc áp dụng các thành tựu của cuộc cách mạng công nghệ 4.0 ngày càng đa dạng, việc thống kê, phân tích, xử lý dữ liệu, đặc biệt là những dữ liệu lớn (Bigdata) trong công tác dự báo GTVT là hết sức cần thiết. Nghiên cứu đã chỉ ra các ưu điểm, khả năng ứng dụng và cách vận hành cơ bản của phần mềm R (một trong những phần mềm thống kê và xử lý dữ liệu mã nguồn mở miễn phí tốt nhất hiện nay) để hỗ trợ công tác dự báo nhu cầu GTVT đảm bảo cơ sở khoa học và tính thực tiễn.

Nghiên cứu cũng đã chỉ ra quy trình dự báo cũng như tính toán khối lượng vận tải đường bộ bằng phần mềm R, chỉ ra mức độ sai số chấp nhận được của phần mềm R so với các phần mềm dự báo chuyên dùng khác.

Tài liệu tham khảo

- [1]. Bộ GTVT (2017), Thông tư số 48/2017/TT-BGTVT ngày 13/12/2017 quy định hệ thống chỉ tiêu thống kê và chế độ báo cáo thống kê ngành GTVT.
- [2]. Quốc hội (2021), Luật sửa đổi bổ sung một số điều và phụ lục danh mục chỉ tiêu thống kê quốc gia của Luật Thống kê số 01/2021/QH15.
- [3]. Geoffery J.Meaden & Đỗ Chí Thắng (1996), Hệ thống thông tin địa lý: Ứng dụng cho nghề biển, Tổ chức Lương thực và Nông nghiệp.
- [4]. Nguyễn, L. P. H. (2020), Big Data và xu hướng ứng dụng trong hoạt động thông tin - thư viện.
- [5]. Nguyễn, V. T. (2014), Phân tích dữ liệu với R, Tổng hợp TP. Hồ Chí Minh.
- [6]. Ihaka, R., & Gentleman, R. (1996), R: Ngôn ngữ phân tích dữ liệu và đồ họa, Tạp chí Thống kê tính toán và Đồ họa.
- [7]. Hayslett, H. T. (2014), Thống kê, Elsevier.
- [8]. Polit, D. F., & Hungler, B. P. (1999), Nghiên cứu điều dưỡng: Nguyên tắc và Phương pháp (Tái bản lần thứ sáu).
- [9]. Black, W. R. (2003), Giao thông vận tải: Phân tích địa lý, Tạp chí Guilford.
- [10]. Teodorović, D., & Nikolić, M. (2020), Phương pháp định tính trong GTVT, Tạp chí CRC.

Ngày nhận bài: 01/9/2023

Ngày chấp nhận đăng: 25/12/2023

Người phản biện: TS. Lương Tuấn Anh

TS. Đinh Trọng Hùng