TẠP CHÍ
KHÍ TƯỢNG THỦY VĂN

Check for updates

Bài báo khoa học

Ứng dụng thuật toán trên nền tảng ngôn ngữ R để nghiên cứu vi nhựa trong nước mặt lục địa, sông Sài Gòn và sông Đồng Nai

Huỳnh Phú[1], Huỳnh Thị Ngọc Hân[2,3]*, Nguyễn Thị Huệ[3], Võ Hoàng Khang[4]

[1] Viện khoa học ứng dụng HUTECH, Trường Đại học Công nghệ TP. Hồ Chí Minh; h.phu@hutech.edu.vn;

[2] Trường Đại học Tài nguyên và Môi trường Thành phố Hồ Chí Minh; htnhan_ctn@hcmunre.edu.vn

[3] Viện Công nghệ môi trường, Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học Công nghệ Việt Nam; nthue2003@gmail.com; ngochanosh@gmail.com

[4] Khoa Công nghệ thông tin, Trường Đại học Công nghệ TP Hồ Chí Minh, vh.khang@hutech.edu.vn

*Tác giả liên hệ: htnhan_ctn@hcmunre.edu.vn; Tel.: +84–975397953

Ban Biên tập nhận bài: 5/11/2023; Ngày phản biện xong: 6/12/2023; Ngày đăng bài: 25/3/2024

Tóm tắt: Nghiên cứu dự báo diễn biến phát tán vi nhựa trong nước mặt lục địa bằng ngôn ngữ lập trình R, thuật toán có chức năng mô hình hóa các đặc điểm hình dạng của vi nhựa theo mùa trong năm và mực nước thủy triều trong ngày. Sử dụng nền tảng R để hình thành hàm phân tích tương quan, thành phần chính dữ liệu, xử lý đa cộng tuyến dữ liệu và phân tích cụm nhằm mục đích dự đoán xu hướng hình dạng của vi nhựa trong nước sông trong thời gian gần. Kết quả ứng dụng nghiên cứu cho thấy tất cả các vị trí lấy mẫu trên lưu vực các sông Sài Gòn - Đồng Nai đều đang có mức vi nhựa dạng sợi cao và tăng tịnh tiến theo các năm. Kết quả nghiên cứu là công cụ hiệu quả trong dự đoán diễn biến thay đổi về đặc điểm, hình dạng của vi nhựa dưới tác động của môi trường, theo mùa trong năm, theo thủy triều lên và xuống. Hơn nữa, kết quả nghiên cứu này có thể được sử dụng để điều chỉnh các mô hình phù hợp với dữ liệu được thu thập trong điều kiện liên tục, lượng dữ liệu lưu trữ lớn, độ chính xác cao. Cần có sự chú trọng trong nguồn gốc làm phát sinh vi nhựa để có sự kiểm soát và quản lý kịp thời. Đây là công cụ đóng góp quan trọng trong nghiên cứu vi nhựa trong nước sông Sài gòn, sông Đồng Nai và sẽ là ứng dụng rộng rãi cho nghiên cứu vi nhựa nước mặt lục địa.

Từ khóa: Phân tích cụm; Phân tích tương quan đa biến; Phân tích thành phần chính; R; Vi nhựa.

1. Giới thiệu

Vi nhựa có mặt khắp nơi và gây ảnh hưởng đến nhiều môi trường, như biển [1], sông, hồ [2], Bắc Cực [3], đất [4] và không khí [5]. Các nghiên cứu đã cho thấy tác động tiêu cực của vi nhựa đối với môi trường và hệ sinh thái [6], đưa ra nhiều luận cứ về nguồn gốc và tác động đến sức khỏe người [7]. Vi nhựa có thể chứa các hóa chất độc hại, chẳng hạn như Phthalate hoặc Ete Diphenyl Polybrominated, đồng thời có khả năng hấp phụ, hấp thụ và giải phóng, phân tán các chất hữu cơ khó phân hủy [8]. Hơn nữa, tốc độ phân hủy nhựa chậm, cần nhiều thời gian và tùy theo điều kiện môi trường vật lý, chúng có thể tồn tại trong môi trường trong nhiều năm, nhiều thế kỷ [9–10].

Phần lớn các kết quả nghiên cứu trước đây đều tập trung vào vi nhựa trong nước biển, trong khi đó môi trường nước ngọt, nước mặt lục địa ít được chú ý [11–12]. Nhưng gần đây, việc nghiên cứu vi nhựa trong các sông đã được tăng cường [2]. Các phát hiện sự tồn tại của vi nhựa trong các mẫu nước sông đã cho thấy sự phong phú về mật độ của vi nhựa ở sông có mối tương quan với nhau với các yếu tố nhân tạo và mật độ cao hơn thường được phát hiện ở các khu vực có mật độ dân số và tập trung đông dân cư sinh sống. Điều này đã được quan sát và công bố kết quả trong các nghiên cứu khác nhau ở nhiều nơi trên khắp toàn cầu như: Châu Âu [13–16], Châu Á [17–21], Úc [22–23] hoặc Bắc Mỹ [2, 24–25]. Một số nghiên cứu không thể chỉ ra mối tương quan tích cực giữa sự phong phú về mật độ của vi hạt nhựa và mật độ dân số [26–28]. Mặc dù cũng có một số nghiên cứu có thể chỉ ra mối quan hệ rõ ràng giữa mật độ của vi nhựa ngày càng tăng là do mức độ đô thị hóa ngày càng tăng hay sự phát triển của công nghiệp hóa ở các khu vực lưu vực sông [29–30]. Trong khi ở một vài nghiên cứu khác không có mối quan hệ rõ ràng giữa sự thay đổi môi trường nước sông và lượng vi nhựa dồi dào đã được tìm thấy [21, 31]. Một số công bố kết quả điều tra về mật độ của vi nhựa dọc theo sông từ khu vực thưa dân đến nơi tập trung đô thị [18, 21]. Quy mô nghiên cứu các yếu tố khu vực, điều kiện môi trường theo mùa hay thủy triều và địa phương nghiên cứu có thể là nguyên nhân gây ra sự hạn chế như vậy giữa các công bố [2, 27], nhưng nội dung đi sâu vào diễn biến, đặc điểm của vi nhựa trong các yếu tố điều kiện là rất cần thiết để hiểu đầy đủ về chúng.

Cho đến hiện tại, Sông Sài Gòn và sông Đồng Nai vẫn là nguồn cung cấp nước sinh hoạt chính cho người dân 2 bên lưu vực sông. Hệ thống cửa sông Sài Gòn, sau khi chảy qua đô thị lớn là Thành phố Hồ Chí Minh, kết quả phân tích mẫu nước trên kênh Nhiêu Lộc, Thị Nghè cho thấy tổng khối lượng vi nhựa đại diện chiếm 11-43% [32]. Tác giả [33] đã lần đầu tiên công bố kết quả nghiên cứu về mật độ vi nhựa trong nước và trầm tích 2 con sông này đã xuất hiện vi nhựa dạng mảnh, dạng sợi và dạng hạt có kích thước 0,1-5 mm. Trong đó, phần lớn là chủng nhựa PE 51,2%, PP 27,1%, PVC 13,4% và 8,3% là các loại nhựa khác [33–34]. Emilie Strady và cộng sự (2020) cũng đã công bố trong nước sông Sài Gòn có sự xuất hiện của vi nhựa dạng sợi từ 22-251 sợi trong 1 lít nước, không tính đến yếu tố ảnh hưởng như lượng mưa, lưu lượng nước hoặc các yếu tố phi sinh học khác. Ước lượng hàng năm có từ 115 × 1012 đến 164 × 1012 vi nhựa dạng sợi được thải ra từ con sông Sài Gòn [35–36].

Ngày nay, khoa học phát triển, nhiều vấn đề môi trường cần được khám phá đã tạo ra một lượng dữ liệu rất lớn từ các số liệu phân tích, hệ thống quan trắc điện tử, hệ thống lưu trữ dữ liệu đa phương tiện và các ứng dụng trong mạng không gian của vạn vật (*Internet of things*). Sự tiến bộ về công nghệ đã giúp các nhà nghiên cứu có thể chuyển từ việc thu, nhận dữ liệu ở mức thấp sang nghiên cứu tích hợp mức cao với khả năng phân tích, nhận dạng, dự đoán các vấn đề [37]. Mô hình học máy là một trong những phương pháp giải quyết vấn đề được đặt ra, nó là một thuật toán quét qua một lượng vô cùng lớn dữ liệu để tìm ra các đặc trưng mẫu hoặc đưa ra dự báo, dự đoán. Có khá nhiều mô hình học máy (*ML - Machine learning*) như: học không giám sát, học có giám sát, học bán giám sát, học sâu,... Với sự bùng nổ của của trí thông minh nhân tạo (*AI - Artifical Intelligence*), nhiều nhà phân tích đã cho thấy học sâu (*DL - Deep learning*) là một trong những kỹ thuật thành công nhất để tìm ra các đặc trưng mẫu và đưa ra dự báo. Hiện nay trong ngành môi trường đã có nhiều phương pháp để dự báo lan truyền môi trường như mô hình Gauss tính toán lan truyền trong không khí, mô hình Berliand kỹ thuật, mô hình Suttan hay mô hình Streeter-phelps. Tuy nhiên, hầu hết các mô hình này được xây dựng có bản quyền, hoặc tính toán bởi các công cụ đánh giá và phân tích dữ liệu là phần mềm Excel của Microsoft Office với các hàm số đơn giản, hoặc ứng dụng các phần mềm được lập trình sẵn có như SPSS, Stata, EViews... Hạn chế của các phần mềm này là: i) Không được miễn phí, phiên bản miễn phí lại bị hạn chế sử dụng; ii) Người đánh giá bắt buộc phải phụ thuộc vào điều kiện hay độ tin cậy trong xử lý dữ liệu đã được áp đặt sẵn vào phần mềm từ người xây dựng, những người mà bị hạn chế hiểu biết về lĩnh vực ngành môi trường, đặc biệt là nghiên cứu vấn đề chuyên sâu như vi nhựa; iii) Các phần mềm được xây dựng sẵn chủ yếu phát huy hiệu quả mạnh trong xử lý số liệu đơn thuần và cho ra các kết quả cụ thể, mang tính rập khuôn, phụ thuộc sự ngẫu nhiên của dữ liệu đầu vào; điều này khá phù hợp đối với lĩnh vực tài chính, tiền tệ nhưng trong nghiên cứu về môi trường cần phải có phương pháp luận của các nhà chuyên môn; iv) Nếu dữ liệu cần xử lý quá nhiều, vượt hạn mức số dòng xử lý cho phép thì các phần mềm trên không thể xử lý được hoặc mức độ tin cậy về kết quả dễ dàng bị tác động.

Trên cơ sở những nội dung được trình bày, mục đích nghiên cứu thực hiện xây dựng thuật toán biểu diễn và học máy trên nền tảng ngôn ngữ R để phục vụ cho vấn đề nghiên cứu diễn biến hình dạng của vi nhựa của sông Sài Gòn và sông Đồng Nai dưới ảnh hưởng của mực nước thủy triều lên, thủy triều xuống và các mùa mưa, mùa khô dọc theo thượng nguồn đến hạ nguồn của các sông và để quan sát sự thay đổi của mức độ mật độ của vi nhựa trong quá trình chuyển đổi từ các vị trí dân cư thưa thớt đến nơi tập trung khu dân cư, đô thị, các thành phố lớn. Từ các kết quả phân tích dữ liệu và ứng dụng đại số hình học bảo giác (*Conformal Geometric Algebra - CGA*) với học máy (*Machine Learning*), nghiên cứu sẽ đưa ra các dự báo về khuynh hướng đặc điểm của vi nhựa theo sự ảnh hưởng bởi các biến yếu tố tác động với độ tin cậy cao trong xử lý số liệu. Kết quả nghiên cứu sẽ giúp các chuyên gia, các nhà quản lý nắm bắt hiện trạng và xu hướng, từ đó quản lý và giảm thiểu vi nhựa trong môi trường nước sông hiệu quả.

2. Phương pháp nghiên cứu

2.1. Vị trí và phương pháp lấy mẫu

Công tác lấy mẫu nước và các yếu tố về không gian, thời gian được thực hiện tại 18 vị trí trên lưu vực sông Sài Gòn và sông Đồng Nai. Trong đó, 13 vị trí từ hồ Dầu Tiếng đến ngã ba Rạch Kỳ Hà được ký hiệu SGL.1, SGL.2, SGL.3, SGL.4, SGL.5, SGL.6, SGL.7, SGL.8, SGL.9, SGL.10, SGL.11, SGL.12, SGL.13 và 5 vị trí từ hồ Trị An đến cửa sông Đồng Nai-Soài Rạp được ký hiệu DNL.1, DNL.2, DNL.3, DNL.4, DNL.5 (Hình 1).

Hình 1. Sơ đồ vị trí lấy mẫu.

Mẫu được thu bằng lưới Manta có kích thước mắt lưới 0,3 mm (tỷ lệ diện tích mở của lưới là 39,1%). Diện tích của khung là 0,30 × 0,15 m và cho phép lấy mẫu lớp nước bề mặt cao 30 cm. Tốc độ dòng chảy của sông tại mỗi vị trí lấy mẫu được đo bằng thiết bị đo dòng chảy ADCP cố định (*Workhorse Waves Array*), đồng hồ đo lưu lượng cơ học định hướng. Các phép đo của đồng hồ đo lưu lượng được chuyển đổi thành lượng nước đi qua lưới Manta. Tại mỗi vị trí lưới chắn nước và đồng hồ đo lưu lượng được đặt trong 5 phút xuống sông, nơi chúng được đặt đứng yên. Tại mỗi vị trí, 2 mẫu được lấy lặp lại cách nhau 30 phút.

Công tác lấy mẫu được thực hiện vào lúc thủy triều lên và thủy triều xuống trong ngày, vào mùa mưa và mùa khô trong các năm 2021 và 2022. Các mẫu sau khi lấy qua lưới Manta sẽ được loại bỏ thủ công các tạp chất lớn hơn 5mm như cành cây, lá, vật liệu xây dựng, những mảnh nhựa lớn... Tổng cộng 144 mẫu nước được bảo quản trong các chai thủy tinh tối màu nhằm tránh bị ảnh hưởng nhiệt độ hay ánh sáng môi trường tự nhiên và được vận chuyển đến phòng thí nghiệm Nation Lab và phòng phân tích Viện Môi trường và Kinh tế tuần hoàn Miền Nam (IECES) để phân tích xác định vi nhựa.

2.4. Phương pháp học có giám sát (Supervised Learning)

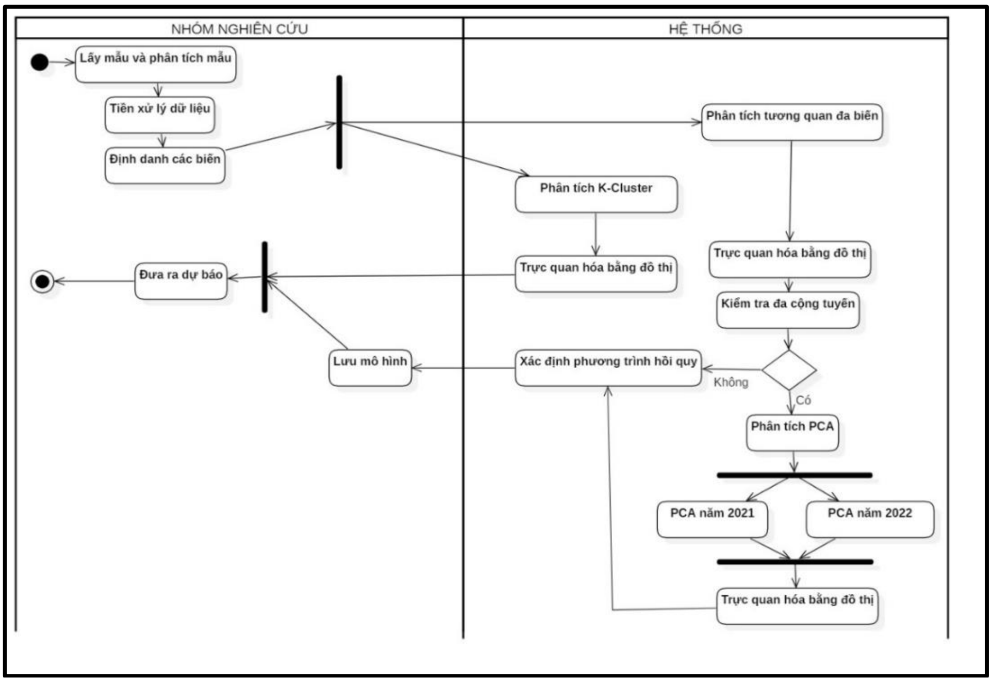
Phương pháp học giám sát là một mô hình hỗ trợ quyết định sử dụng biểu đồ hoặc mô hình của các kết quả có thể xảy ra, bao gồm kết quả sự hiện diện ngẫu nhiên, dữ liệu vi nhựa và môi trường. Học giám sát (*Supervised Learning*) là mô hình học máy dự đoán đầu ra của một dữ liệu mới dựa trên các cặp dữ liệu đầu vào cho trước dựa trên 2 thuật toán:

- Cây quyết định (*Decision Trees*): là số câu hỏi Có/Không tối thiểu được đặt ra để đánh giá xác suất đưa ra quyết định về kết quả là đúng. Cây quyết định cho phép nghiên cứu tiếp cận vấn đề một cách có cấu trúc và có hệ thống để đạt được một kết luận hợp lý.

- Sự phân lớp (*Naïve Bayes*): là một nhóm các sự phân loại xác suất đơn giản dựa trên việc áp dụng định lý Bayes với các giả định độc lập giữa các tính năng.

Trong Machine Learning (Học Máy), một "data set" (Bộ dữ liệu) là một tập hợp các dữ liệu thực tế hoặc ảo được sử dụng để đào tạo và kiểm tra các mô hình máy học. Data set này thường bao gồm hai phần chính: i) Tập dữ liệu đào tạo (*Training Data*): Đây là một phần của data set được sử dụng để đào tạo mô hình máy học. Tập dữ liệu này thường bao gồm các cặp (đầu vào, đầu ra) hoặc các mẫu dữ liệu được gán nhãn tương ứng. Trong quá trình huấn luyện, mô hình học từ các mẫu này để có thể thực hiện dự đoán hoặc phân loại dữ liệu mới sau này; ii) Tập dữ liệu kiểm tra (*Test Data*): Đây là một phần khác của data set được sử dụng để kiểm tra hiệu suất của mô hình học máy sau khi nó đã được đào tạo. Dữ liệu kiểm tra thường không được sử dụng trong quá trình đào tạo, và mô hình không biết trước kết quả của các mẫu này. Dữ liệu kiểm tra giúp đánh giá khả năng tổng quát hóa của mô hình, tức là khả năng của nó trong việc dự đoán dữ liệu mới mà nó chưa từng thấy.

Mục tiêu của mô hình học có giám sát trong ML là tìm ra một hàm số dự báo mà giá trị của chúng khác với "ground truth" là nhỏ nhất. Ground truth ở đây chính là giá trị của biến mục tiêu. Sự sai khác này có được thông qua các hàm mất mát (*Loss function*). Huấn luyện mô hình học máy thực chất là quy về tìm cực trị của hàm mất mát. Tùy thuộc vào bài toán mà có các dạng hàm mất mát khác nhau. Trong bài toán dự báo nghiên cứu này sử dụng hàm MSE (*Mean Square Error*) làm hàm mất mát. Hàm số này có giá trị bằng trung bình của tổng bình phương sai số giữa giá trị dự báo và giá trị thực. Sơ đồ mô hình xử lý dữ liệu của nghiên cứu được thể hiện trong Hình 2.



Hình 2. Sơ đồ kiến trúc mô hình xử lý.

Hình 2 đã thể hiện đầy đủ các mô hình Machine learning đã được sử dụng trong nghiên cứu này bao gồm: i) Mô hình hồi quy tuyến tính đa biến: Dùng cho các mẫu trong bảng dữ liệu số, trong nghiên cứu này là bảng tính Excel các dữ liệu về đặc điểm hình dạng vi nhựa. Các mô hình tuyến tính sử dụng đại số để dự đoán mối quan hệ giữa các biến trong các dự báo; ii) Mô hình phân lớp, phân cụm K-Means: Dự đoán cụm các biến gần; iii) Mô hình đồ họa: Dùng trực quang hóa kết quả phân tích, thể hiện dưới dạng biểu đồ một xác suất, hay xu hướng; iv) Cây quyết định/ Rừng ngẫu nhiên: Câu hỏi Có/Không được đặt ra để đánh giá xác suất đưa ra quyết định là đúng.

2.5. Phương pháp hồi quy tuyến tính đa biến

Hồi qui tuyến tính đa biến là phương pháp hồi quy tuyến tính với nhiều biến đầu vào. Trong nghiên cứu này, các biến đầu vào là đặc điểm hình dạng (sợi, mảnh, hạt và dạng khác) của vi nhựa theo mùa và diễn biến của thủy triều. Phương trình hồi quy có dạng:

$$\widehat{\mathrm{Y}}\mathrm{i} = \mathrm{F}(\mathrm{X}1, \mathrm{X}2, \ldots, \mathrm{Xk}) = \omega\mathrm{o} + \omega1\mathrm{Xi}1 + \omega2\mathrm{Xi}2 + \cdots + \omega\mathrm{pXik} = \mathrm{W}^{\mathrm{T}}\mathrm{Xi} \quad (1)$$

Ở đây ta xem X_i là một vec tơ đại diện cho quan sát thứ i. Cụ thể nó gồm các giá trị (X_1, X_2, ..., X_k). Ma trận X có kích thước n × k, trong đó cứ mỗi dòng là một quan sát, và mỗi cột là một biến số. Còn X_i là quan sát thứ i của biến thứ k. Ma trận mở rộng của X được (ký hiệu là $\bar{X}$), chính là ma trận có thêm vec tơ cột 1 được thêm vào đầu tiên. Lúc này, xét toàn bộ tập dữ liệu, ta được:

$$\widehat{\mathrm{Y}} = \mathrm{F}(\mathrm{X}) = \begin{bmatrix} 1 & \mathrm{X}_{11} & \ldots & \mathrm{X}_{1k} \\ 1 & \mathrm{X}_{21} & \ldots & \mathrm{X}_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \mathrm{X}_{n1} & \ldots & \mathrm{X}_{nk} \end{bmatrix} \quad (2)$$

Vectơ sai số giữa hiệu (Y-$\widehat{\mathrm{Y}}$) được biểu diễn thành:

$$\mathrm{E} = \mathrm{Y} - \widehat{\mathrm{Y}} = \mathrm{Y} - \overline{\mathrm{X}}_{\mathrm{z}} \quad (3)$$

Hàm MSE là trung bình của tổng bình phương của các sai số, nó được biểu diễn:

$$\mathcal{L}(\mathrm{Z}) = \frac{1}{2\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{Y_i} - \widehat{\mathrm{Y}}_{\mathrm{I}})^2 = \frac{1}{2}\mathrm{e}^{\mathrm{T}}\mathrm{e} = (\mathrm{Y} - \overline{\mathrm{X}}_{\mathrm{z}})^{\mathrm{T}}(\mathrm{Y} - \overline{\mathrm{X}}_{\mathrm{z}}) = \|\overline{\mathrm{X}}_{\mathrm{z}} - \mathrm{Y}\|_2^2 \quad (4)$$

Biểu thức $\|\bar{X}_z - Y\|_2^2$ là bình phương Norm chuẩn bậc hai (*Euclidean norm*). Bằng cách khai triển đại số tuyến tính sẽ tính được đạo hàm của hàm MSE:

$$\frac{\partial \mathrm{L}(\mathrm{z})}{\partial \mathcal{L}} = \overline{\mathrm{X}}^{\mathrm{T}}(\overline{\mathrm{X}}_{\mathrm{z}} - \mathrm{Y}) \quad (5)$$

Phương trình hồi qui có nghiệm là:

$$Z = (\bar{X}^T\bar{X})^{-1}\bar{X}^T Y = (A^{-1}b) \quad (6)$$

$$\text{Rút gọn } A = \bar{X}^T\bar{X} \text{ và } \bar{X}^T Y = b \quad (7)$$

Phương hình hồi quy đa biến có nghiệm khi là khả nghịch.

2.6. Ngôn ngữ R

Ngôn ngữ R là một ngôn ngữ lập trình và môi trường phần mềm dành cho tính toán và đồ họa thống kê. Đây là một bản hiện thực ngôn ngữ lập trình do Ross Ihaka và Robert Gentleman tạo ra [38]. Cho đến nay ngôn ngữ R là do R Development Core Team tiếp tục phát triển. Ngôn ngữ R còn được viết tắt là R. Đây là ngôn ngữ miễn phí với đặc tính nổi trội là mã nguồn mở, hiệu quả mạnh và được hỗ trợ mở rộng và là môi trường tối ưu cho nghiên cứu, thống kê và trình bày dữ liệu, đặc biệt là các tập dữ liệu lớn (Big data). Hiệu quả và sức mạnh của ngôn ngữ lập trình này thông qua các hành vi: i) Chạy các mã code không cần trình biên dịch, hay nói cách khác chính code R là một ngôn ngữ thông dịch (*Interpreted language*); ii) Thực thi mọi phép tính trên vec tơ (*Vectors language*) mà không cần đến vòng lặp mà vẫn có thể dùng bất cứ chức năng (*Function*) nào; iii) Ứng dụng trong việc thống kê dữ liệu (*Statistical language*), có thể đáp ứng hoàn thành tất cả các thuật toán vì bản chất của code R là một loại ngôn ngữ "*turing - language*". Hay nói cách khác "*turing - language*" là ngôn ngữ lập trình linh hoạt trong mọi bài toán; iv) Ngôn ngữ R là một mã nguồn mở (*Open-source*) hoàn toàn miễn phí nên sẽ tiết kiệm nhiều chi phí mà vẫn mô phỏng dữ liệu rất hiệu quả.

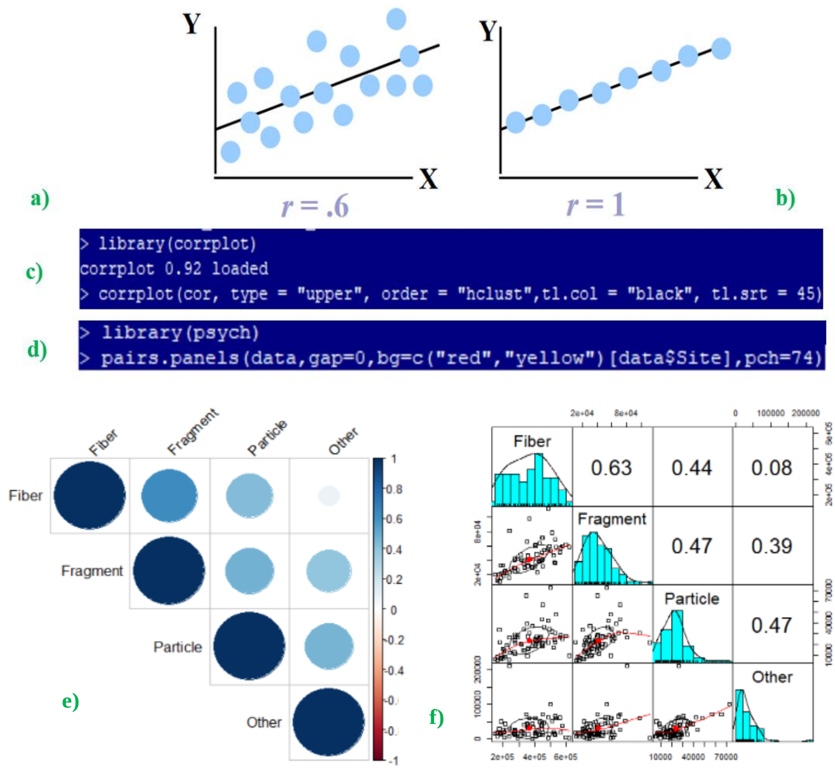
Các gói Package hỗ trợ thống kê tối ưu cùng công nghệ hiện đại, được cập nhật thường xuyên bởi cộng đồng chuyên môn khắp thế giới. Hầu hết các ý tưởng mới đều được ứng dụng trên nền tảng R trước tiên và bất kỳ nhà lập trình nào cũng có thể sử dụng, cải tiến theo các code riêng, các sáng tạo và mục đích riêng. Bên cạnh đó, người dùng R có thể hiểu rõ cách vận hành của chúng thông qua phân tích source code, từ đó tự chủ động thêm tính năng, sửa lỗi, tích hợp sử dụng thêm các ngôn ngữ khác như JAVA, SQL, C, C++… Ngôn ngữ này là một nền tảng độc lập, vì vậy, phù hợp để chạy được trên mọi hệ điều hành. Hệ sinh thái R rất tối ưu và mạnh mẽ, phù hợp nhiều loại dữ liệu nên có các dạng hệ cơ sở sữ liệu phong phú.

2.7. Huấn luyện mô hình

Các gói Package trong R là tập hợp các hàm và bộ dữ liệu do cộng đồng phát triển. Trong môi trường R, các gói này được lưu trữ trong thư mục có tên Library. Thư viện hệ thống trong R (*System library*) có chứa các gói Package quan trọng và đã được cài đặt sẵn hỗ trợ người sử dụng. Trong nghiên cứu này, các gói được sử dụng là Psych (*Corrplot*), Factoextra, Ggbiplot, Magrittr, và Dplyr.

2.8. Phương pháp phân tích tương quan đa biến

Phương pháp này được dùng để đánh giá mối tương quan giữa nhiều biến cùng một lúc. Hệ số tương quan (*Coefficient of correlation*) được ký hiệu trong nghiên cứu này là r và được sử dụng để đo lường độ lớn của mối quan hệ giữa hai biến số. Các Package và Hàm lệnh phân tích tương quan đa biến, xác định r và biểu diễn đồ thị được thể hiện trong Hình 3.

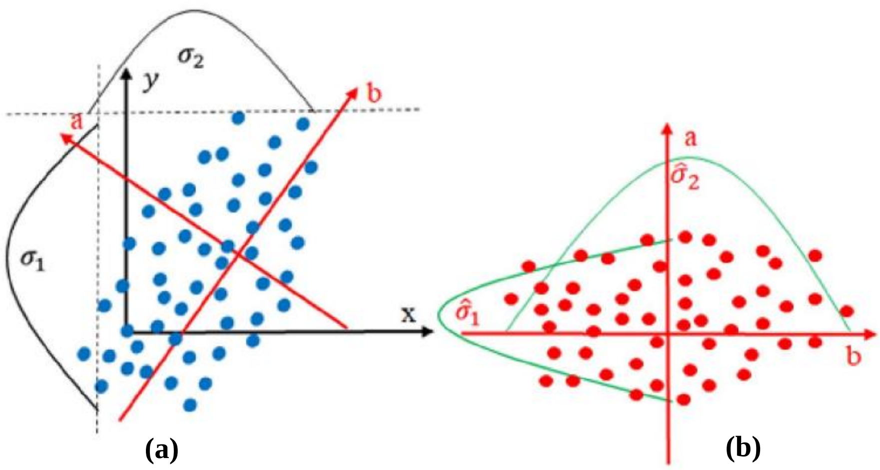


Hình 3. Phân tích tương quan đa biến bằng ngôn ngữ R: a) Mô hình biểu diễn mức tương quan giữa biến X và Y với = 0,6; b) Mô hình biểu diễn mức tương quan hoàn hảo giữa biến X và Y với r = 1; c) Hàm phân tích tương quan bằng R với gói Corrplot; d) Hàm phân tích bằng R với gói Psych; e) Đồ thị tương quan được biểu diễn bởi gói Corrplot; f) Đồ thị tương quan được biểu diễn bởi gói Psych.

Trong nghiên cứu này, phương pháp phân tích tương quan đối tượng vi nhựa trên ngôn ngữ R được sử dụng bởi hàm phân tích (*Pairs.panels*) dựa trên Package Psych.

2.9. Phương pháp phân tích thành phần chính

Phân tích thành phần chính (*Principal Components Analysis - PCA*) là hữu ích nhất khi dữ liệu nằm trên hoặc gần với một không gian con tuyến tính của tập dữ liệu. Với loại dữ liệu này, PCA tìm một cơ sở cho không gian con tuyến tính và cho phép bỏ qua các đặc trưng không liên quan. Với một tập dữ liệu được cho, nơi mỗi mẫu dữ liệu có *D* chiều (tức là *D* đặc trưng), PCA tính một tập các véc tơ đặc trưng *D* chiều được gióng với các hướng mà có phương sai cực đại của dữ liệu. Các thành phần chính có một số cách sử dụng: (1) Chiếu dữ liệu gốc lên các thành phần chính này; (2) Sử dụng các thành phần chính này để tạo ra các điểm mới. Phương pháp phân tích thành phần chính là một thuật toán thống kê sử dụng phép biến đổi trực giao để biến đổi một tập hợp dữ liệu từ một không gian nhiều chiều sang một không gian mới ít chiều hơn nhằm tối ưu hóa sự biến thiên của dữ liệu (*Maximize the variability*) [38–41]. Hình 4 phát thảo về cách mà dữ liệu được biểu diễn trong không gian mới của PCA.



Hình 4. Biểu diễn dữ liệu trong không gian mới của PCA: (a) Không gian dữ liệu ban đầu; (b) Không gian dữ liệu mới.

Trong không gian dữ liệu ban đầu, Hình 4a với các tập điểm màu xanh được quan sát trong hệ trục xOy, theo mỗi chiều đều có phương sai lớn. Trong không gian dữ liệu mới mới, Hình 4b với tập điểm màu đỏ được quan sát theo hệ trục aOb theo chiều thứ hai có phương sai ð1 nhỏ hơn so với ð2. Điều này cho thấy khi chiếu dữ liệu lên trục a sẽ có được các điểm gần nhau và gần với kỳ vọng. Vì vậy, phương pháp phân tgich1 PCA còn được coi là phương pháp giảm chiều của dữ liệu mà vẫn giữ được tổng phương sai còn lại là lớn nhất.

Giả sử cho tập dữ liệu huấn luyện: $X = \{X_i | X_i \epsilon R^D\}, i\epsilon\{1 \dots N\}$. Với X_i là vectơ thuộc không gian D chiều, N là số lượng vectơ trong tập X. Thay vì giữ lại các trục tọa độ của không gian cũ D chiều, PCA xây dựng một không gian mới K chiều với K<D, nhưng lại có khả năng biểu diễn dữ liệu tốt tương đương không gian cũ, nghĩa là đảm bảo độ biến thiên (*Variability*) của dữ liệu trên mỗi chiều mới. Trong không gian mới, PCA sẽ chiếu dữ liệu lên hệ trục mới để xác định phương sai lớn nhất của phép biến đổi. Sau đó, sắp xếp theo thứ tự giảm dần của giá trị riêng ứng với K vectơ riêng lớn nhất. Cuối cùng, chọn K vectơ riêng đầu tiên, gọi là các thành phần chính, để tạo thành một không gian con gần với phân bố của dữ liệu ban đầu. Trong bài báo này, nghiên cứu chọn K vec tơ đầu tiên tạo được tổng phương sai lớn hơn 60% để thực hiện nghiên cứu tiếp theo.

Trong nền tảng ngôn ngữ R, để phân tích PCA các đặc tính của vi nhựa theo thời gian và không gian, nghiên cứu sử dụng các hàm R tích hợp Prcomp() và Princomp(). Nghiên cứu thực hiện dự đoán khuynh hướng biến đổi của các đặc tính vi nhựa trong môi trường nước sông Sài Gòn và sông Đồng Nai trước sự thay đổi của thủy triều và thời gian vào mùa khô, mùa mưa của năm. Định dạng đơn giản của 2 hàm này là:

> Prcomp (x,scale = FALSE)

> Princomp (x,cor = FALSE,scores = TRUE)

x: là biến ma trận số hay khung dữ liệu

cor: là một giá trị hợp lý, nếu TRUE, dữ liệu căn giữa và chia tỉ lệ trước khi được phân tích

scores: là một giá trị hợp lý, nếu TRUE, tọa độ trên từng PC được tính toán

Các kết quả đầu ra được hàm Prcomp() và Princomp() trả về bao gồm: i) Độ lệch chuẩn của các thành phần chính; ii/ Ma trận tải trọng thay đổi (các cột là vectơ riêng); iii/ Các khả biến; iv/ Độ lệch chuẩn của các biến (tỷ lệ áp dụng cho từng biến); v) Tọa độ là các đại lượng tuyến tính hay góc chỉ vị trí của một điểm trong mối quan hệ với một hệ quy chiếu cho trước (quan sát) trên các thành phần chính. Gói Packages được sử dụng cho các hàm này gồm Devtools và Factoextra.

> If(!require(devtools))install.packages("devtools")

> Devtools:install_github ("kassambara/factoextra")

2.10. Phương pháp phân cụm

Phân cụm trong R đề cập đến quá trình nhóm các điểm dữ liệu tương tự lại với nhau dựa trên đặc điểm của chúng. R, ngôn ngữ lập trình cho điện toán thống kê và đồ họa, cung cấp nhiều thuật toán phân cụm khác nhau như K-Means, phân cụm theo cấp bậc. Phân cụm K-Mean là phương pháp phân vùng phổ biến nhất. Nó yêu cầu nhà phân tích chỉ định số lượng cụm cần trích xuất. Biểu đồ tổng bình phương của các nhóm bên trong theo số cụm được trích xuất có thể giúp xác định số cụm thích hợp. Nhà phân tích tìm kiếm điểm uốn cong trong biểu đồ tương tự như phép kiểm tra nền trong phân tích nhân tố.

3. Kết quả nghiên cứu

Toàn bộ tập dữ liệu sẽ được thực hiện tiền xử lý trước khi sử dụng R. Trong quy trình làm sạch tập dữ liệu về vi nhựa, thứ tự của các bước là rất quan trọng. Các bước làm sạch dữ liệu được nghiên cứu thực hiện bao gồm: Nhập dữ liệu, làm sạch hoặc đổi tên cột, loại bỏ trùng lặp, tạo và chuyển đổi cột (ví dụ: mã hóa lại hoặc chuẩn hóa các giá trị), lọc hoặc thêm hàng. Thực hiện tạo dữ liệu:

```
>ThS_HuynhThiNgocHan=data.frame(DangSoi, DangManh, DangHat, DangKhac)
## # A tibble: 5 × 144
##                              ViTri   DangSoi   DangManh   DangHat   DangKhac
##                              <chr>   <dbl>     <dbl>      <dbl>     <dbl>
## 1  SGL1-High_tide-Dry_season        163969    13613      8331      17271
## 2  SGL2-High_tide-Dry_season        202466    18451      7380      17713
## 3  SGL3-High_tide-Dry_season        212350    25563      8715      43864
## ....
....
>data (ThS_HuynhThiNgocHan)
```

Kết quả nhập dữ liệu được xuất dưới nền tảng R được biểu diễn trong Hình 5.

Trong bài báo này, nghiên cứu sẽ xây dựng và cung cấp mã R để tính toán và trực quan hóa PCA trong R bằng cách sử dụng hàm Prcomp() và gói Factoextra và thực hiện từng bước:

Bước 1: Tải Package Factoextra để trực quan hóa dữ liệu

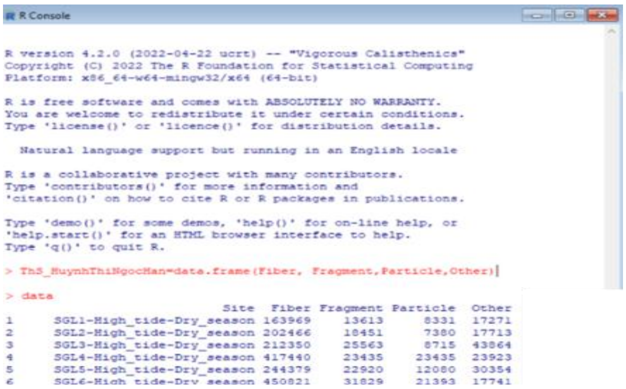
>library (factoextra)

Bước 2: Tính toán PCA

> res.pca<-prcomp(ThS_HuynhThiNgocHan,scale=TRUE)

Bước 3: Trực quan hóa các giá trị riêng (sơ đồ Scree Plot). Hiển thị tỷ lệ phần trăm phương sai được giải thích bởi từng thành phần chính.

> fviz_eig(res.pca)



Hình 5. Nhập và xuất dữ liệu trên giao diện ngôn ngữ R.

Bước 4: Đồ thị của các thành phần. Những thành phần có đặc điểm giống nhau sẽ được nhóm lại với nhau.

```
> fviz_pca_ind (res.pca,
        col.ind="cos2", # Màu sắc theo thuộc tính đại diện
        gradient.cols=c ("#00AFBB", "#E7B800", "#FC4E07"),
        repel=TRUE     # Tránh chồng chéo văn bản)
```

Bước 5: Đồ thị của các biến. Các biến tương quan dương chỉ về cùng một phía của biểu đồ. Các biến tương quan âm chỉ ra các cạnh đối diện của biểu đồ.

```
>fviz_pca_var (res.pca,
        col.var="contrib", # Màu sắc do đóng góp cho PC
        gradient.cols=c ("#00AFBB", "#E7B800", "#FC4E07"),
        repel=TRUE     # Tránh chồng chéo văn bản)
```

Bước 6: Vẽ biểu đồ Biplot

```
> fviz_pca_biplot (res.pca,repel = TRUE,
         col.var="#2E9FDF", # Màu biến
         col.ind="#696969" # Màu thành phần)
```

Bước 7: Truy cập vào kết quả PCA.

Bước 8: Dự đoán tọa độ của các biến thành phần và biến bổ sung chỉ bằng cách sử dụng thông tin do PCA đã thực hiện trước đó cung cấp. Dữ liệu mới phải chứa các cột (biến) có cùng tên và theo cùng thứ tự với dữ liệu hiện hoạt được sử dụng để tính toán PCA. Sử dụng hàm cơ sở R dự đoán():

```
> ind.sup.coord<-predict(res.pca,newdata=ind.sup)
```

Sơ đồ thành phần bao gồm cả thành phần bổ sung:

```
## Sơ đồ các đại diện chính
> p<- fviz_pca_ind (res.pca,repel=TRUE)
## Bổ sung đại diện
> fviz_add (p,ind.sup.coord,color="blue")
```

Tọa độ dự đoán của các thành phần có thể được tính toán thủ công như sau: i) Tập trung và chia tỷ lệ dữ liệu thành phần mới bằng cách sử dụng trung tâm và thang đo của PCA; ii) Tính tọa độ dự đoán bằng cách nhân các giá trị tỷ lệ với các vectơ riêng (Loadings) của các thành phần chính.

Bước 9: Các biến định tính/phân loại có thể được sử dụng để tô màu các cá nhân theo nhóm. Biến nhóm phải có cùng độ dài với số lượng thành phần đang hoạt động.

```
> groups<-as.factor(ThS_HuynhThiNgocHan$Competition[1:])
> fviz_pca_ind (res.pca,
        col.ind=groups, # Màu nhóm
        palette=c("#00AFBB",  "#FC4E07"),
        addEllipses=TRUE, # Elip tập trung
        ellipse.type="confidence",
```

```
        legend.title="Groups",
        repel=TRUE)
```

Tính toán tọa độ cho các cấp độ của các biến nhóm. Tọa độ của một nhóm nhất định được tính bằng tọa độ trung bình của các thành phần trong nhóm.

```
> library (magrittr) #cho loại %>%
> library (dplyr)   # Các loại khác
# 1. Tọa độ riêng
> res.ind<-get_pca_ind (res.pca)
# 2. Tọa độ nhóm
> coord.groups<-res.ind$coord %>%
>coord.groups
```

Tọa độ của một biến định lượng nhất định được tính bằng mối tương quan giữa các biến định lượng và các thành phần chính.

```
# Dự đoán tọa độ và tính Cos2
>quanti.coord <- cor(quanti.sup, res.pca$x)
# Đồ thị các biến bao gồm các biến bổ sung
>p <- fviz_pca_var(res.pca)
```

Kết quả phân tích tương quan đa biến và PCA các đặc tính hình dạng của vi nhựa trong môi trường nước sông Sài Gòn và sông Đồng Nai trên nền tảng R được thể hiện trên Hình 6 và Hình 7. Bên cạnh đó, nghiên cứu sử dụng phương pháp vẽ đồ thị có thể để hình dung ảnh hưởng của độ không đảm bảo tham số trong những dự đoán của mô hình nghiên cứu.

Dựa theo kết quả thu được sau khi phân tích PCA bằng ngôn ngữ R được thể hiện trong Hình 6 đã cho thấy những đóng góp đầu tiên thành phần chính về hình dáng vi nhựa trong năm 2021 thỏa mãn phần trăm giải thích phương sai >10% là PC1(56,6%), PC2(23,8%) và PC3 (12,7%) được biểu diễn trên Hình 6a và Hình 6c. Tuy nhiên, nghiên cứu này chỉ lựa chọn 2 thành phần chính đầu tiên vì tổng 2 thành phần này đã giải thích được toàn bộ 80,5% tập dữ liệu yêu cầu. Các vec tơ màu xanh cho biết mối quan hệ giữa các biến ban đầu (hình dáng MPs) và các thành phần chính, độ dài của vector cho biết độ mạnh của mối tương quan của biến ban đầu với thành phần chính. Theo kết quả phân tích ma trận xoay trong Hình 6a, mối tương quan giữa các hình dáng vi nhựa trong Hình 6d và kết quả giá trị P-Value trong Hình 6b cho thấy rằng: i) Dạng sợi - dạng khác gần như không có mối tương quan (r = 0,08) và dạng mảnh - dạng hạt có mối tương quan yếu (r = 0,47), các giá trị P-Value < 0,05 (có ý nghĩa thống kê); ii) Ngoài ra, Hình 6a cũng cho thấy dạng mảnh và dạng hạt có mối tương quan với thành phần chính PC1 cao hơn các dạng khác và dạng sợi, ngoài ra trong PC2 thì dạng sợi và dạng khác có mối tương quan cao với thành phần chính này (nhưng giá trị P-Value giữa dạng sợi-dạng khác trong Hình 6b lại cho thấy chúng không có ý nghĩa thống kê P-Value = 0,48).

a) Kết quả phân tích PCA của R; b) Kết quả trình diễn hệ số tương quan và giá trị P-Value của các hình dạng vi nhựa từ R; c) Biểu đồ Scree plot thể hiện phần trăm phương sai mà các thành phần chính giải thích bộ dữ liệu; d) Biểu đồ tương quan đa biến các hình dạng của vi nhựa được tìm thấy; e) Biểu đồ dự báo xu hướng biến đổi của các hình dạng vi nhựa theo PC1 và PC2 (Hình 6).

Từ các phân tích trên, nghiên cứu lựa chọn thực hiện phép trích các hình dáng vi nhựa cho các thành phần chính như sau:

PC1 gồm: Dạng sợi (hệ số xác định là 49%), dạng mảnh (56%) và dạng hạt (52,8%)

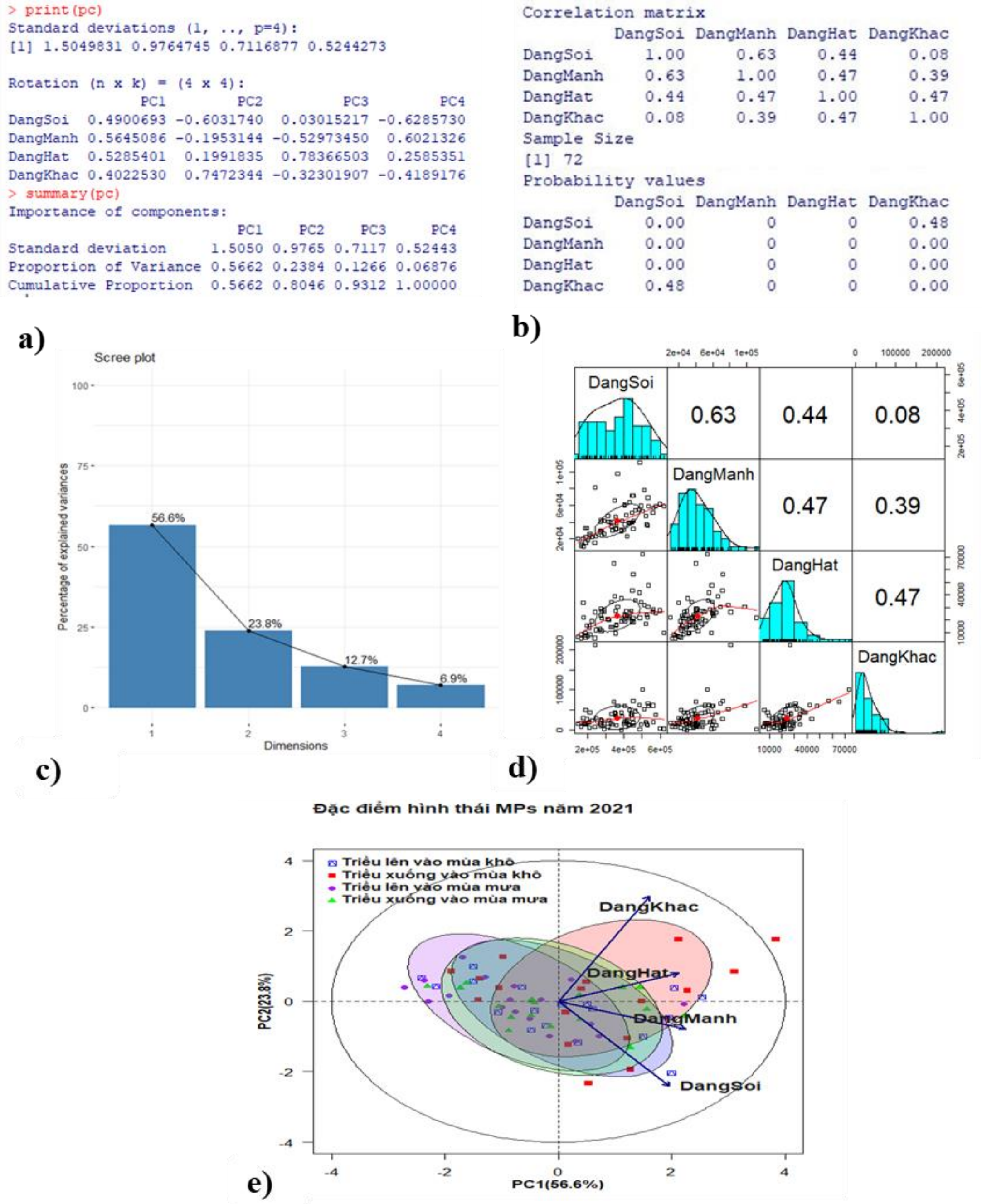
PC2 gồm: Dạng khác (74,7%)

Biểu diễn phương trình cho thành phần chính PC1:

$$\mathrm{PC1} = 0{,}49\mathrm{DangSoi} + 0{,}56\mathrm{DangManh} + 0{,}53\mathrm{DangHạt} \tag{8}$$

Biểu diễn phương trình cho thành phần chính PC2:

$$\mathrm{PC2} = 0{,}75\mathrm{DangKhac} \tag{9}$$



Hình 6. Kết quả phân tích tương quan và PCA các đặc điểm về hình dạng của vi nhựa trong môi trường nước sông Sài Gòn - Đồng Nai trong năm 2021.

Hình 6e cho thấy gần như các hình dáng vi nhựa trong các điểm khảo sát trên sông Sài Gòn và sông Đồng Nai đề có liên quan đến thành phần chính đầu tiên. Các vi nhựa bị phân rã ngoài dạng sợi, mảnh, hạt có liên quan đến thành phần chính thứ hai. Các hạt dạng sợi, hạt và bị phân mảnh là các loại vi nhựa chính được tìm thấy trong nước bề mặt lưu vực sông trong cả hai mùa, vào lúc triều lên và xuống trong ngày.

Phân tích tương tự cho năm 2022, Hình 7a và Hình 7c cho thấy những đóng góp đầu tiên thành phần chính về hình dáng vi nhựa trong năm 2022 thỏa mãn phần trăm giải thích phương sai > 10% là PC1 (65,3%), PC2 (19,1%) và PC3 (12,3%). Nghiên cứu này chỉ lựa chọn 2 thành phần chính đầu tiên vì tổng 2 thành phần này đã giải thích được toàn bộ 84,3% tập dữ liệu yêu cầu trong năm 2022 (Hình 7c). Vec tơ hình dáng vi nhựa dạng sợi có chiều dài vượt trội hơn so với năm 2021. Điều này chứng tỏ mức nhiễm vi nhựa dạng sợi trên lưu vực 2 con sông vẫn cao và tăng hơn năm 2021. Kết luận này được khẳng định bởi ý nghĩa thống kê của dữ liệu khảo sát là rất rõ ràng bởi giá trị P-Values hoàn toàn < 0,05 (Hình 7b).

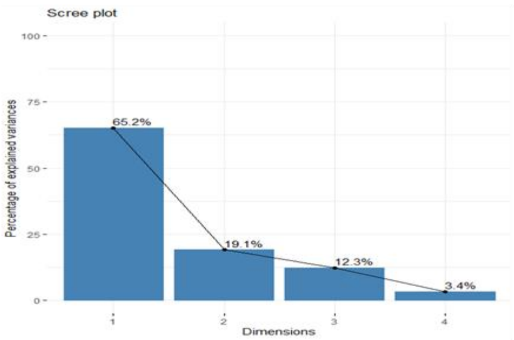
```
> print(pc)
Standard deviations (1, .., p=4):
[1] 1.6147184 0.8748295 0.7016527 0.3674798

Rotation (n x k) = (4 x 4):
                PC1         PC2         PC3        PC4
DangSoi   0.4255148  0.78690590 -0.27508579  0.3521989
DangManh  0.5112127 -0.19430465  0.73889952  0.3936175
DangHat   0.5907643  0.05233241  0.03357134 -0.8044450
DangKhac  0.4567233 -0.58334048 -0.61418912  0.2718260
> summary(pc)
Importance of components:
                          PC1    PC2    PC3     PC4
Standard deviation     1.6147 0.8748 0.7017 0.36748
Proportion of Variance 0.6518 0.1913 0.1231 0.03376
Cumulative Proportion  0.6518 0.8432 0.9662 1.00000
```

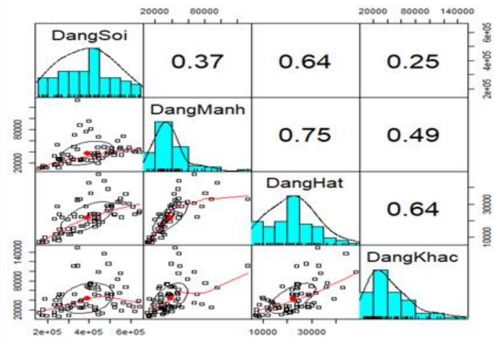
a)

```
Correlation matrix
         DangSoi DangManh DangHat DangKhac
DangSoi     1.00     0.37    0.64     0.25
DangManh    0.37     1.00    0.75     0.49
DangHat     0.64     0.75    1.00     0.64
DangKhac    0.25     0.49    0.64     1.00
Sample Size
[1] 72
Probability values
         DangSoi DangManh DangHat DangKhac
DangSoi     0.00        0       0     0.03
DangManh    0.00        0       0     0.00
DangHat     0.00        0       0     0.00
DangKhac    0.03        0       0     0.00
```

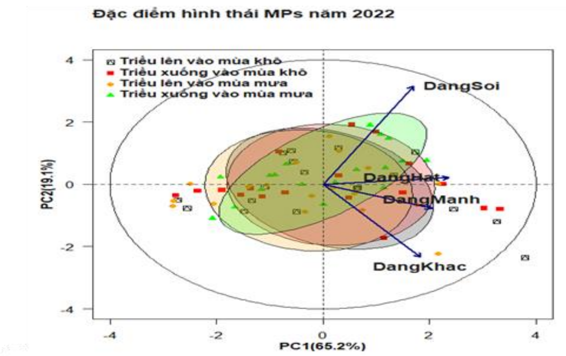
b)



c)



d)



e)

Hình 7. Kết quả phân tích tương quan và PCA các đặc điểm về hình dạng của vi nhựa trong môi trường nước sông Sài Gòn - Đồng Nai trong năm 2022.

a) Kết quả phân tích PCA của R; b) Kết quả trình diễn hệ số tương quan và giá trị P-Value của các hình dạng vi nhựa từ R; c) Biểu đồ Scree plot thể hiện phần trăm phương sai mà các thành phần chính giải thích bộ dữ liệu; d) Biểu đồ tương quan đa biến các hình dạng của vi nhựa được tìm thấy; e) Biểu đồ dự báo xu hướng biến đổi của các hình dạng vi nhựa theo PC1 và PC2.

Để xác định các nhân tố trong 2 thành phần PC1 và PC2, nghiên cứu dựa vào Hình 7a và Hình 7d và lựa chọn kết quả trích hình dạng vi nhựa như sau:

- PC1 gồm: Dạng sợi (hệ số xác định là 43%), dạng mảnh (51%) và dạng khác (46%)
- PC2 gồm: Dạng khác (52%)

Phương trình biểu diễn thành phần chính PC1:

$$PC1 = 0.43DangSoi + 0.51DangManh + 0.46DangKhac \quad (10)$$

Phương trình biểu diễn thành phần chính PC2:

$$PC2 = 0.52DangHat \quad (11)$$

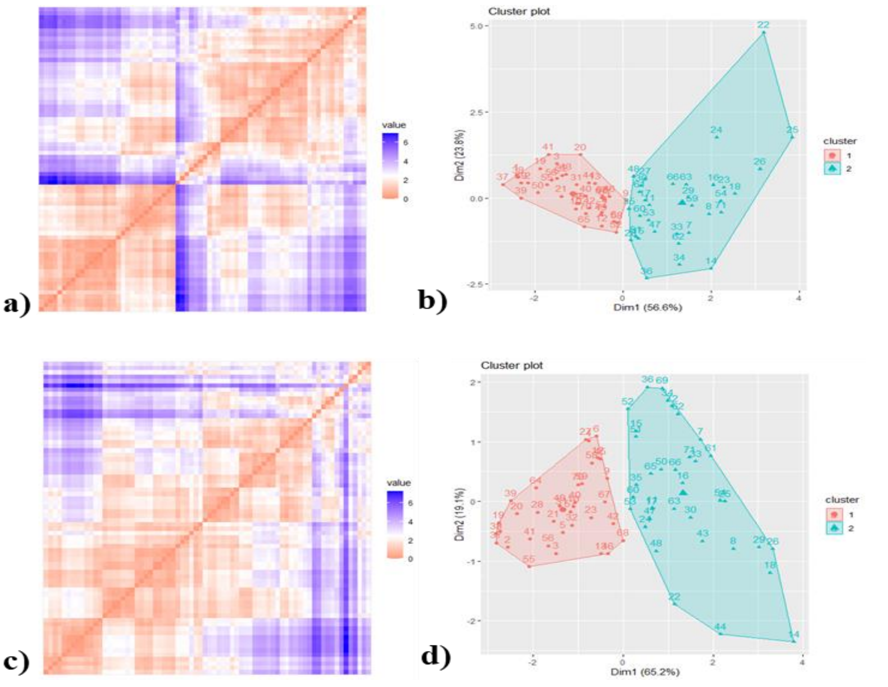
Gần như các hình dáng vi nhựa trong các điểm khảo sát trên sông Sài Gòn, sông Đồng Nai đề có liên quan đến thành phần chính đầu tiên (dạng sợi, mảnh). Các vi nhựa bị phân rã dạng hạt có liên quan đến thành phần chính thứ hai. Các vi nhựa dạng sợi, mảnh và bị phân mảnh là các loại vi nhựa chính được tìm thấy trong nước bề mặt lưu vực sông vào lúc triều lên và xuống trong ngày, trong cả hai mùa, đặc biệt là vào mùa khô. Trong đó hàm lượng vi nhựa dạng sơi khá cao. Trong Hình 7e, vi nhựa trong triều lên và triều xuống của mùa khô và vào lúc triều lên của mùa mưa đều chịu ảnh hưởng nhiều bởi PC1 (giải thích được 65,3% phương sai) trong khoản giá trị PC1 là từ -2 đến 2 (cụm elip xám, đỏ và cam) và vi nhựa các hình dáng còn lại xuất hiện trong mùa mưa lúc triều xuống chịu ảnh hưởng chính bởi PC2 (giải thích được 19,1% phương sai), có cụm elip màu xanh lá. Ngoài ra, kết quả phân tích PCA cho thấy lưu vực sông Sài Gòn và sông Đồng Nai bị ảnh hưởng bởi các dạng vi nhựa sợi, mảnh nhiều hơn trong cả hai mùa của năm 2022. Hàm lượng vi nhựa dạng sợi từ các nguồn phát sinh ảnh hưởng lưu vực các sông rất cao.

Như vậy, trong 2 năm 2021-2022, mức nhiễm vi nhựa dạng sợi, mảnh khá vượt trội hơn các dạng khác trong cả các mùa của năm, năm 2022 cao hơn năm 2021. Cần có sự chú trọng trong nguồn gốc làm phát sinh vi nhựa dạng này để có sự kiểm soát và quản lý kịp thời.

Để khám phá những địa điểm lấy mẫu có đặc điểm hình dáng vi nhựa tương tự nhau, một phân tích cụm được tiến hành dựa trên hình dạng của vi nhựa trung bình trong mùa khô và mùa mưa (Hình 8).

```
# Chuẩn bị dữ liệu
>ThS_HuynhThiNgocHan<-na.omit(ThS_HuynhThiNgocHan) #Danh sách xóa thiếu
>ThS_HuynhThiNgocHan<-scale(ThS_HuynhThiNgocHan) #Chuẩn hóa các biến
# Xác định số cụm
> get_clust_tendency (df, n=15,graph=T)
> cl$hopkins
# Chỉ số Hopkins (-1:1) cho thấy có thể phân nhóm được không, càng gần 1 thì khả năng phân nhóm tốt
> print (cl$plot)
```

Các vị trí lấy mẫu được chia thành 2 nhóm trong cả 2 mùa của năm 2021 gồm 2 cụm mùa cơ bản là màu xanh dương và màu đỏ với chỉ số Hopkins = 0,76 (dữ liệu khá thích hợp để phân tích cụm) và K-Cluster tối ưu là 2 (nghiên cứu chọn phân nhóm các vị trí lấy mẫu thành 2 cụm để quan sát). Có sự khác biệt tương đối rõ rệt trong phân loại các vị trí lấy mẫu.



Hình 8. Kết quả phân tích cụm dựa trên hình dạng của vi nhựa trong trong 2 năm 2021 - 2022: a) Hình minh họa mối liên kết theo màu của 2 cụm (xanh dương và đỏ) trong năm 2021; b) Biểu đồ cụm năm 2021của 18 vị trí khảo sát trên các sông Sài Gòn và sông Đồng Nai; c) Hình minh họa mối liên kết theo màu của 2 cụm (xanh dương và đỏ) năm 2022; b) Biểu đồ cụm (cluster plot) vào năm 2022 của 18 vị trí khảo sát trên các sông Sài Gòn và sông Đồng Nai.

Các vị trí thuộc nhóm màu xanh dương và nhóm màu đỏ có chung đặc điểm về hình dạng vi nhựa, chịu ảnh hưởng bởi PC1 giải thích 56,6% phương sai gồm dạng sợi (hệ số xác định là 49%), dạng mảnh (56%) và dạng hạt (53%). Các phân loại đặc điểm hình dáng vi nhựa tại địa điểm lấy mẫu cho các cụm khác nhau được trình diễn trong Hình 7b. Tương tự như vậy, vào năm 2022, vị trí lấy mẫu được chia thành 2 cụm màu cơ bản là màu xanh dương và màu đỏ với chỉ số Hopkins = 0,78 và K-Cluster tối ưu là 2. Có sự khác biệt tương đối rõ rệt trong phân loại các vị trí lấy mẫu. Các vị trí thuộc nhóm màu xanh dương và nhóm màu đỏ có chung đặc điểm về hình dạng vi nhựa, chịu ảnh hưởng bởi PC1 giải thích 65,2% phương sai gồm dạng sợi (hệ số xác định là 43%), dạng mảnh (51%) và dạng khác (46%). Các phân loại đặc điểm hình dáng vi nhựa tại địa điểm lấy mẫu cho các cụm khác nhau được trình diễn trong Hình 8.

Kết luận về kết quả phân tích cụm đã đáp ứng phù hợp với kết quả PCA về hình dáng MPs. Tất cả các vị trí lấy mẫu trên lưu vực các sông Sài Gòn - Đồng Nai đều đang có mức MPs dạng sợi cao và tăng tịnh tiến theo năm. Kết quả nghiên cứu đặc điểm của hình dạng vi nhựa theo mùa và thủy triều bằng ngôn ngữ R có thể đạt được môi trường xử lý lượng dữ liệu lớn, đưa ra được các kết quả đồ thị thể hiện trực quang rõ ràng về dữ liệu nghiên cứu, đồng thời với phương pháp Machine learning, các dữ liệu được xử lý phối hợp nhiều mô hình đưa ra được kết quả với độ tin cậy cao hơn mà hiện nay các phương pháp mô hình khác chưa thực hiện được. Nếu dữ liệu các năm được thêm và bổ sung vào tệp dữ liệu, số mẫu tăng tịnh tiến, hệ thống sẽ tự học và tự phân tích đưa ra kết quả về diễn biến hình dạng vi nhựa trong tương lai, nhiều kết quả nghiên cứu sẽ đạt độ chính xác về dự báo càng tối ưu. Phương pháp này rất phù hợp và cần được phát triển trong vấn đề xử lý dữ liệu lớn của môi trường, như đánh giá tệp dữ liệu tịnh tiến về các yếu tố gây tổn thương và ô nhiễm để dự báo mức độ biến đổi khí hậu. Đặc biệt, việc dùng phương pháp dự báo với tệp dữ liệu có số dòng mẫu tăng tịnh tiến ở mức cao (Big data) bằng ngôn ngữ R đối với các chất "ô nhiễm mới" như vi nhựa là vô cùng quan trọng trong công việc nghiên cứu, bởi vì sự tạo thành các chất này (thời gian phân rã chất thải nhựa thành vi nhựa) lại cần độ dài thời gian lớn, đôi khi vượt quá tuổi thọ loài người.

4. Kết luận

Ứng dụng thành công phương pháp phân tích dữ liệu thực tế để dự báo diễn biến đặc điểm hình dạng vi nhựa trong môi trường nước sông Sài Gòn và sông Đồng Nai trên nền tảng ngôn ngữ R là kết quả nghiên cứu hiệu quả trong lĩnh vực nghiên cứu về vi nhựa hiện nay. Một trong những khía cạnh thách thức nhất của việc xây dựng phương pháp dự báo trên nền tảng R là cần phải xác định package nào được sử dụng trong trường hợp cụ thể. Trong kết quả xây dựng được, nghiên cứu đã cung cấp được ít nhất các package đặc biệt được thiết kế riêng cho mục đích phân tích tương quan, phân tích thành phần chính và phân tích cụm hiệu quả, trình diễn các biểu đồ diễn giải kết quả phân tích một cách rõ ràng các đặc tính hình dạng của vi nhựa theo mùa và theo mực nước thủy triều. Phương pháp xây dựng thuật toán giảm chiều dữ liệu và phân cụm trên R dự đoán khuynh hướng đặc điểm hình dạng của vi nhựa trong nước mặt lục địa sông Sài Gòn và sông Đồng Nai đã giải quyết được nhu cầu về một công cụ đánh giá với độ tin cậy cao, xử lý được vấn đề đa cộng tuyến trong dữ liệu thực nghiệm, xử lý được thư viện dữ liệu lớn và hoàn toàn tránh được chi phí tài chính cho bản quyền từ các phần mềm được bán sẵn trên thị trường. Ngoài ra, phương pháp này có thể được lập trình các hàm đồ họa, đưa ra các hình ảnh, biểu đồ hiển thị phong phú, giải thích rõ ràng các nội dung mà các nhà chuyên môn môi trường, các nhà nghiên cứu vi nhựa có thể luận giải được trong khám phá dữ liệu nghiên cứu vi nhựa. Bên cạnh đó, hạn chế trong kết quả đạt được của nghiên cứu là dữ liệu quan sát vi nhựa trên 2 lưu vực sông còn ít, chỉ mới thu thập được trong 2 năm là 2021 và 2022 bởi vì "chất ô nhiễm" mới này hiện nay gần như chỉ mới được nghiên cứu ở mức độ phát hiện, đánh giá về mật độ. Chính lượng dữ liệu còn hạn chế làm giảm độ tin cậy khi phân tích bằng các mô hình ML của ngôn ngữ R mặc dù độ chính xác của phương pháp rất cao. Các mã code, gói Package luôn cần phải được điều chỉnh thích hợp khi cập nhật hệ điều hành máy tính và các phiên bản của R.

Đóng góp của tác giả: Xây dựng ý tưởng nghiên cứu: H.P., H.T.N.H.; Lựa chọn phương pháp nghiên cứu: H.P., N.T.H.; Xử lý số liệu: H.T.N.H., V.H.K.; Lấy mẫu: H.P., H.T.N.H.; Viết bản thảo bài báo: H.P., H.T.N.H.; Chỉnh sửa bài báo: H.P., H.T.N.H.

Lời cảm ơn: Nghiên cứu này được thực hiện dưới sự tài trợ của dự án nghiên cứu Viện Môi trường và Kinh tế tuần hoàn Miền Nam (IECES).

Lời cam đoan: Các tác giả đảm bảo rằng bài viết này là công trình của các tác giả, chưa được xuất bản ở nơi khác, không được sao chép từ nghiên cứu trước đó; không có xung đột lợi ích trong nhóm tác giả.

Tài liệu tham khảo

1. Avio, C.G.; Gorbi, S.; Regoli, F. Plastics and microplastics in the oceans: from emerging pollutants to emerged threat. *Mar. Environ. Res*. **2017**, *128*, 2–11. https://doi.org/10.1016/j.marenvres.2016.05.012.
2. Talbot, R.; Chang, H. Microplastics in freshwater: a global review of factors affecting spatial and temporal variations. *Environ. Pollut*. **2022**, *292*, 118393. https://doi.org/10.1016/j.envpol.2021.118393.
3. Peeken, I.; Primpke, S.; Beyer, B.; Gutermann, J.; Katlein, C.; Krumpen, T.; Bergmann, M.; Hehemann, L.; Gerdts, G. Arctic sea ice is an important temporal sink and means of transport for microplastic. *Nat. Commun*. **2018**, *9*, 1505. https://doi.org/10.1038/s41467-018-03825-5.
4. Nizzetto, L.; Bussi, G.; Futter, M.N.; Butterfield, D.; Whitehead, P.G. A theoretical assessment of microplastic transport in river catchments and their retention by soils and river sediments. *Environ. Sci. Process Impacts*. **2016**, *18*, 1050–1059. https://doi.org/10.1039/c6em00206d.
5. Bergmann, M.; Mützel, S.; Primpke, S.; Tekman, M.B.; Trachsel, J.; Gerdts, G. White and wonderful? Microplastics prevail in snow from the alps to the arctic. *Sci. Adv*. **2019**, *5*, eaax1157. https://doi.org/10.1126/sciadv.aax1157.
6. Windsor, F.M.; Tilley, R.M.; Tyler, C.R.; Ormerod, S.J. Microplastic ingestion by riverine macroinvertebrates. *Sci. Total Environ*. **2019**, *646*, 68–74. https://doi.org/10.1016/j.scitotenv.2018.07.271.
7. Phu, H.; Han, H.T.N.; Thao, N.L. Plastic waste, microplastics in the Saigon – Dong Nai river basin, the risk of impacts on the health of people. *J. Hydrometeorol.* **2022**, *736(1)*, 14–27.
8. Crawford, C.B.; Quinn, B. Microplastic Pollutants, first ed. Elsevier. 2017. https://doi.org/10.1016/C2015-0-04315-5.
9. Born, M.P.; Brüll, C. From model to nature - A review on the transferability of marine (micro-) plastic fragmentation studies. *Sci. Total Environ*. **2022**, *811*, 151389. https://doi.org/10.1016/j.scitotenv.2021.151389.
10. Chamas, A.; Moon, H.; Zheng, J.; Qiu, Y.; Tabassum, T.; Jang, J.H.; Abu-Omar, M.; Scott, S.L.; Suh, S. Degradation rates of plastics in the environment. *ACS Sustainable Chem. Eng*. **2020**, *8*, 3494–3511. https://doi.org/10.1021/acssuschemeng.9b06635.
11. Blettler, M.C.M.; Abrial, E.; Khan, F.R.; Sivri, N.; Espinola, L.A. Freshwater plastic pollution: recognizing research biases and identifying knowledge gaps. *Water Res*. **2018**, *143*, 416–424. https://doi.org/10.1016/j.watres.2018.06.015.
12. Eerkes-Medrano, D.; Thompson, R. Occurrence, fate, and effect of microplastics in freshwater systems. In: Zeng, E.Y. (Ed.). Microplastic Contamination in Aquatic Environments. 2018, 95–132. https://doi.org/10.1016/B978-0-12-813747-5.00004-7.
13. Dris, R.; Gasperi, J.; Rocher, V.; Saad, M.; Renault, N.; Tassin, B. Microplastic contamination in an urban area: a case study in Greater Paris. *Environ. Chem.* **2015**, *12(5)*, 592–599. https://doi.org/10.1071/EN14167.

14. Liu, F.; Olesen, K.B., Borregaard, A.R.; Vollertsen, J. Microplastics in urban and highway stormwater retention ponds. *Sci. Total Environ*. **2019**, *671*, 992–1000. https://doi.org/10.1016/j.scitotenv.2019.03.416.
15. Lechner, A.; Keckeis, H.; Lumesberger-Loisl, F.; Zens, B.; Krusch, R.; Tritthart, M.; Glas, M.; Schludermann, E. The Danube so colourful: a potpourri of plastic litter outnumbers fish larvae in Europe's second largest river. *Environ. Pollut*. **2014**, *188*, 177–181. https://doi.org/10.1016/j.envpol.2014.02.006.
16. Mani, T.; Hauk, A.; Walter, U.; Burkhardt-Holm, P. Microplastics profile along the rhine river. *Sci. Rep*. **2015**, *5*, 17988. https://doi.org/10.1038/srep17988.
17. Irfan, M.; Qadir, A.; Mumtaz, M.; Ahmad, S.R. An unintended challenge of microplastic pollution in the urban surface water system of Lahore, Pakistan. *Environ. Sci. Pollut. Res. Int*. **2020**, *27*, 16718–16730. https://doi.org/10.1007/s11356-020-08114-7.
18. Chen, H.L.; Gibbins, C.N.; Selvam, S.B.; Ting, K.N. Spatio-temporal variation of microplastic along a rural to urban transition in a tropical river. *Environ. Pollut*. **2021**, *289*, 117895. https://doi.org/10.1016/j.envpol.2021.117895.
19. Lin, C.T.; Chiu, M.C.; Kuo, M.H. Effects of anthropogenic activities on microplastics in deposit-feeders (Diptera: chironomidae) in an urban river of Taiwan. *Sci. Rep*. **2021**, *11*, 400. https://doi.org/10.1038/s41598-020-79881-z.
20. Kameda, Y.; Yamada, N.; Fujita, E. Source- and polymer-specific size distributions of fine microplastics in surface water in an urban river. *Environ. Pollut*. **2021**, *284*, 117516. https://doi.org/10.1016/j.envpol.2021.117516.
21. Lahens, L.; Strady, E.; Kieu-Le, T.C.; Dris, R.; Boukerma, K.; Rinnert, E.; Gasperi, J.; Tassin, B. Macroplastic and microplastic contamination assessment of a tropical river (Saigon River, Vietnam) transversed by a developing megacity. *Environ. Pollut*. **2018**, *236*, 661–671. https://doi.org/10.1016/j.envpol.2018.02.005.
22. Su, L.; Sharp, S.M.; Pettigrove, V.J.; Craig, N.J.; Nan, B.; Du, F.; Shi, H. Superimposed microplastic pollution in a coastal metropolis. *Water Res.* **2020**, *168*, 115140. https://doi.org/10.1016/j.watres.2019.115140.
23. Leterme, S.C.; Tuuri, E.M.; Drummond, W.J.; Jones, R.; Gascooke, J.R. Microplastics in urban freshwater streams in Adelaide, Australia: a source of plastic pollution in the Gulf St Vincent. *Sci. Total Environ*. **2023**, *856*, 158672. https://doi.org/10.1016/j.scitotenv.2022.158672.
24. Baldwin, A.K.; Corsi, S.R.; Mason, S.A. Plastic debris in 29 great lakes tributaries: relations to watershed attributes and hydrology. *Environ. Sci. Technol*. **2016**, *50*, 10377–10385. https://doi.org/10.1021/acs.est.6b02917.
25. Yonkos, L.T.; Friedel, E.A.; Perez-Reyes, A.C.; Ghosal, S.; Arthur, C.D. Microplastics in four estuarine rivers in the Chesapeake Bay. U.S.A. *Environ. Sci. Technol*. **2014**, *48*, 14195–14202. https://doi.org/10.1021/es5036317.
26. Corcoran, P.L.; Belontz, S.L.; Ryan, K.; Walzak, M.J. Factors controlling the distribution of microplastic particles in benthic sediment of the thames river, Canada. *Environ. Sci. Technol*. **2020**, *54*, 818–825. https://doi.org/10.1021/acs.est.9b04896.
27. Dikareva, N.; Simon, K.S. Microplastic pollution in streams spanning an urbanisation gradient. *Environ. Pollut*. **2019**, *250*, 292–299. https://doi.org/10.1016/j.envpol.2019.03.105.
28. Wong, G.; L¨owemark, L.; Kunz, A. Microplastic pollution of the Tamsui River and its tributaries in northern Taiwan: spatial heterogeneity and correlation with precipitation. *Environ. Pollut*. **2020**, *260*, 113935.
29. Ma, C.Y.; Sheu, Y.T.; Hsia, K.F.; Dong, C.D.; Chen, C.W.; Huang, Y.C.; Kao, C.M. Development of water and sediment quality management strategies for an urban river basin: a case study in Taiwan. *J. Water Supply Res. Technol. Aqua.* **2018**, *67*, 810–823. https://doi.org/10.2166/aqua.2018.084.
30. Schell, T.; Hurley, R.; Nizzetto, L.; Rico, A.; Vighi, M. Spatio-temporal distribution of microplastics in a Mediterranean river catchment: the importance of wastewater as an

environmental pathway. *J. Hazard Mater*. **2021**, *420*, 126481. https://doi.org/10.1016/j.jhazmat.2021.126481.
31. Wang, T.; Wang, J.; Lei, Q.; Zhao, Y.; Wang, L.; Wang, X.; Zhang, W. Microplastic pollution in sophisticated urban river systems: combined influence of land-use types and physicochemical characteristics. *Environ. Pollut*. **2021**, *287*, 117604. https://doi.org/10.1016/j.envpol.2021.117604.
32. Lahens, L.; Strady, E.; Kiều, L.; Dris, R. Macroplastic and microplastic contamination assessment of a tropical river (Saigon River, Vietnam) transversed by a developing megacity. *Environmental Pollution*. **2018**, *236,* 661–671. Doi:10.1016/j.envpol.2018.02.005.
33. Phú, H.; Hân, H.T.N.; Thảo, N.L.; Đông, Đ.V.; Hân T.G. Nghiên cứu mức độ ô nhiễm vi nhựa trong nước và trầm tích sông Sài Gòn – Đồng Nai. *Tạp chí Khí tượng thủy văn* **2021**, *731*, 69–81. Doi:10.36335/VNJHM.2021(731).69-81.
34. Phu, H.; Han, H.T.N.; Hue, N.T. Developing a circular economy from plastic waste and identifying microplastics in drinking and domestic water supplies in Ho Chi Minh City and Southeast provinces. The 3rd International Conference on Environment, Resources and Earth Sciences (ICERES 2023). Green and Nature-based Solutions for Environmental Sustainability and Resources Management. Ho Chi Minh City University of Technology, Vietnam National University-Ho Chi Minh City, October 21st, 2023.
35. Strady, E.; Dang, T.H.; Dao, T.D.; Dinh, H.N.; Do, T.T.D.; Duong, T.N.; Duong, T.T.; Hoang, D.A.; Kieu-Le, T.C.; Le, T.P.Q.; Mai, H.; Trinh, D.M.; Nguyen, Q.H.; Tran-Nguyen, Q.A.; Tran, Q.V.; Truong, T.N.S.; Chu, V.H.; Vo, V.C. Baseline assessment of microplastic concentrations in marine and freshwater environments of a developing southeast Asian country, Viet Nam. Mar. *Pollut. Bull*. 2021, *162*, 111870.
36. Liu, D.H.; Wang, J.J. A PCA-LSTM model for stock index prediction. Proceeding of the 2018 International Conference on Electrical, Control, Automation and Robotics (ECAR 2018), 2018.
37. Smith, L.I. A tutorial on Principal Components Analysis, February 26, 2002.
38. Michael, E.; Christopher, T.; Bishop, M. Probabilistic principal component analysis. *J. Royal Stat. Soc. Ser. B*, **1999**, *61(3)*, pp. 611–622.
39. Jolliffe, I.T. Principal component analysis, New York: 2nd Edn. Springer Verlag, 2002.
40. Trực tuyến: http://gentleman.fhcrc.org/ (Truy cập ngày 30 tháng 11 năm 2023).

Applying algorithm based on the R language platform to study microplastics in continental surface water, Saigon River and Dong Nai River

Huynh Phu[1], Huynh Thi Ngoc Han[2,3]*, Nguyen Thi Hue[3], Vo Hoang Khang[4]

[1] HUTECH Institute of Applied Sciences; HUTECH University; h.phu@hutech.edu.vn
[2] Hochiminh City University of Nature Resources & Environment; htnhan_ctn@hcmunre.edu.vn
[3] Institute of Environmental Technology, Vietnam Academy of Science and Technology; nthue2003@gmail.com; ngochanosh@gmail.com
[4] Faculty of Information Technology; HUTECH University; vh.khang@hutech.edu.vn

Abstract: The study has developed an algorithm to predict the spread of microplastics in continental surface water using the R programming language. The algorithm has the functions of modeling, correlation analysis, principal component analysis, data multicollinearity processing and cluster analysis of shape characteristics of microplastics according to the season of the year and the tidal water level of the day. Research results have shown that sampling locations in the Saigon River and Dong Nai River all have high density of fibrous microplastics and increase year by year. The research results are an effective tool in predicting changes in the shape characteristics of microplastics according to the seasons of the year and the high and low tides of the day. Furthermore, this result can be used to adapt models to data collected under continuous conditions, large amounts of stored data, and high accuracy. From there, it is the basis to pay more attention to the source of microplastics and have timely control and management. This is an important contributing tool in the study of microplastics in water of the Saigon River and Dong Nai River, and will be widely applied to the study of microplastics in continental surface water.

Keywords: Cluster analysis; Multivariate correlation analysis; Principal component analysis; R; Microplastics.