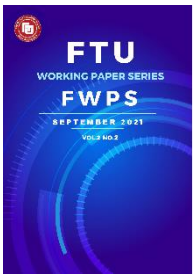


Working Paper 2021.2.4.08
- Vol 2, No 4

XÂY DỰNG MÔ HÌNH PHÂN KHÚC THỊ TRƯỜNG THEO ĐỊA LÝ DÂN SỐ TẠI HÀ NỘI

Văn Đức Mạnh[1], Nguyễn Quỳnh Chi, Bùi Thiên Bình, Trần Ngọc Diệp, Phạm Hương Giang
Sinh viên K57 CTTT Quản trị kinh doanh - Khoa Quản trị kinh doanh
Trường Đại học Ngoại thương, Hà Nội, Việt Nam

Lê Thu Hằng
Giảng viên Khoa Quản trị kinh doanh
Trường Đại học Ngoại thương, Hà Nội, Việt Nam

Tóm tắt

Vị trí địa lý là một trong những yếu tố quan trọng nhất và có sự ảnh hướng lớn đối với chiến lược của các doanh nghiệp khi quyết định gia nhập hay khai thác, mở rộng ở một thị trường mới, một khu vực mới. Bài viết này đưa ra cách xây dựng một mô hình phân tích chi tiết về sự phân bổ của thị trường tại các quận trong thành phố Hà Nội sử dụng phương pháp phân cụm K-means và phân tích thành phần chính (PCA). Mô hình thể hiện rõ đặc điểm dân cư của từng khu vực như tuổi tác, nghề nghiệp, trình độ học vấn,… qua đó đưa ra nguồn thông tin chính xác và tổng quan về mặt địa lý cho các doanh nghiệp, giúp rút ngắn chi phí và thời gian trong quá trình quyết định chiến lược gia nhập hoặc mở rộng ở một khu vực địa lý mới.

Từ khóa: Phân khúc thị trường theo địa lý dân số, Phân cụm K-means, Dữ liệu lớn, Phân tích thành phần chính (PCA), Phân tích vị trí

CREATING A GEODEMOGRAPHIC SEGMENTATION MODEL FOR HANOI

Abstract

Location is one of the most crucial factors which has a great influence on enterprises' strategy when entering, exploiting or expanding in a new market or a new area. The study illustrates how to create a detailed analytical model of the market segmentation in all districts of Hanoi using K-means clustering and principal component analysis (PCA). The model describes the population characteristics of each area such as age, occupation, education level, etc.; thereby giving enterprises precise and general sources of information about geographic location, which helps reduce the cost and time in decision-making process to enter or expand the businesses in a new area.

[1] Tác giả liên hệ, Email: vanducmanhamser@gmail.com

Keywords: Geographic – demographic segmentation, K-means clustering, Big data, Principal component analysis, Location analysis.

1. Mở đầu

Trong sự phát triển nhanh chóng của xã hội hiện nay, đặc biệt là sự bùng nổ công nghệ thông tin, nền kinh tế của chúng ta ngày càng phức tạp, thị trường ngày càng mở rộng và sự cạnh tranh ngày càng trở nên khốc liệt. Điều đó đòi hỏi các doanh nghiệp trong bất kể lĩnh vực ngành nghề nào cũng đều phải tận dụng hết mọi nguồn lực và cơ hội để chiếm được lợi thế cạnh tranh trên thị trường. Một trong những điều quan trọng và tiên quyết nhất đảm bảo sự thành công của một doanh nghiệp là xác định và tiếp cận đúng đối tượng khách hàng tiềm năng. Và một phương pháp phổ biến nhất để tiếp cận khách hàng chính là tìm một vị trí địa lý phù hợp với nhu cầu của doanh nghiệp.

Bởi vì mô hình phân khúc thị trường theo vị trí địa lý dân số có tính ứng dụng vô cùng cao đối với các hoạt động của doanh nghiệp, đặc biệt là trong việc xác định đúng đối tượng khách hàng, nên có rất nhiều doanh nghiệp đã thực hiện nghiên cứu và đưa ra những mô hình địa lý phù hợp nhất với chiến lược của họ. Tuy nhiên, những nghiên cứu này không được công bố rộng rãi và cũng không thể được ứng dụng vào hoạt động của những doanh nghiệp khác. Do đó, với mong muốn đưa ra một mô hình phân tích khái quát, chính xác, và chi tiết về sự phân bố của thị trường tại các quận trong Hà Nội, nhóm nghiên cứu chọn đề tài ***Xây dựng mô hình phân khúc thị trường theo địa lý dân số tại Hà Nội***. Dựa vào mô hình này, các doanh nghiệp tại Việt Nam, đặc biệt là Hà Nội, có thể tìm kiếm những thị trường tiềm năng mới, những khu vực phù hợp giúp mở rộng doanh nghiệp về mặt địa lý.

2. Cơ sở lý thuyết

Cho đến nay, trên thế giới đã có rất nhiều các nghiên cứu về Phân khúc thị trường theo địa lý (Geo-segmentation). Trong đó, những nghiên cứu này đã làm nổi bật được nhiều ứng dụng thực tế của Geo-segmentation và phát triển tính mới của nó theo nhiều phương thức khác nhau. Có thể chia các nghiên cứu đó thành 03 phần chính sau: (1) Ứng dụng của Phân khúc thị trường theo địa lý trong Tiếp thị; (2) Phương pháp tiếp cận và phương pháp luận được sử dụng trong các nghiên cứu trước; và (3) Kết quả tổng hợp của Phân khúc thị trường theo địa lý trong các nghiên cứu đó.

Đầu tiên, nhiều nghiên cứu đã tập trung vào **Tiếp thị theo phương pháp địa lý** (Geo-Marketing, hay GM) với những cách tiếp cận khác nhau. Vào năm 2016, Guy Lansley điều tra về sự phân bố tuổi và giới tính của những người mang tên riêng tiếng Anh và xác định các xu hướng chính trong quy ước đặt tên của nước Anh. Đặc điểm tuổi và giới được biết là có ảnh hưởng lớn đến hành vi của người tiêu dùng, vì vậy việc trích xuất và sử dụng tên để tìm ra những đặc điểm này từ bộ dữ liệu người tiêu dùng có giá trị lớn đối với ngành bán lẻ và tiếp thị. Kết quả từ việc trích xuất có thể được sử dụng để suy ra cấu trúc tuổi và giới tính dự kiến của nhiều bộ dữ liệu người tiêu dùng, cũng như dự đoán các đặc điểm chính của người tiêu dùng ở cấp độ cá nhân (Lansley, 2016). Nghiên cứu vào năm 2019 đã tóm tắt một số phương pháp tiếp thị khi có những dữ liệu về vị trí của người tiêu dùng. Cụ thể, nghiên cứu đã thảo luận về vai trò của bối cảnh thực tế xã hội và thời gian đối với hiệu quả quảng cáo, tiện ích của các công cụ xác định vị trí trong việc làm rõ tính minh bạch của quảng cáo, phân khúc người tiêu dùng và các mối quan tâm về quyền riêng tư về vị trí cá nhân (Banerjee, 2019).

Thứ hai, các nghiên cứu trước đây đã sử dụng đa dạng các **phương pháp tiếp cận và phương pháp luận liên quan đến Phân khúc địa lý theo thị trường**. Vào năm 2011, nghiên cứu của Henna đã sử dụng các tiêu chí lựa chọn điểm đến trượt tuyết để phân khúc khách hàng của khu nghỉ dưỡng trượt tuyết ở Phần Lan (Konu, 2011). Các nghiên cứu của Allo đã chứng minh tính khả thi của tiếp thị địa lý và phân đoạn địa lý ở các nước đang phát triển trong trường hợp của vùng Shomolu của Nigeria và thu được các bản đồ kinh tế xã hội của khu vực bằng cách sử dụng phương pháp lập bản đồ Dasymetric, đây là một giải pháp tiềm năng để lập bản đồ mật độ dân số liên quan đến sử dụng đất thổ cư. Lập bản đồ Dasymetric mô tả dữ liệu vùng định lượng bằng cách sử dụng các ranh giới phân chia khu vực thành các khu vực tương đối đồng nhất với mục đích mô tả rõ hơn sự phân bố dân số (Allo, 2012). Vào năm 2012, Jinsoo Hwang đã xác định các yếu tố ảnh hưởng đến năm nhóm yếu tố quyết định (thực đơn ăn uống, bầu không khí, giá cả, sức khỏe và danh tiếng thương hiệu) mà khách hàng cân nhắc khi lựa chọn một nhà hàng dịch vụ trọn gói (Hwang, 2012).

Cuối cùng, **phương pháp clustering (phân cụm)** khá là phổ biến trong các nghiên cứu về Geo-marketing. Fisher và Tate (2015) so sánh các thuật toán phân cụm được sử dụng trong các nghiên cứu về phân loại nhân khẩu học dựa trên dữ liệu về dân số vào năm 2001 của UK Office for National Statistic (ONS). Họ cho thấy cả c-means và fuzzy c-means đều khiến cho kết quả của phân đoạn thị trường dựa trên khu vực địa lý trở nên thành công và đáng kể hơn. Shaffer (2015) khảo sát các nhà máy bia thủ công ở khu vực Đại đô thị Phoenix để xác định xu hướng nhân khẩu học, hành vi người tiêu dùng và mối quan hệ không gian trong thị trường bia thủ công. Vào năm 2016, Suhaibah đề xuất một sự kết hợp của phân khúc thị trường dựa trên các tiêu chí địa lý và thuật toán phân cụm cho hoạt động quản lý dữ liệu tiếp thị địa lý 3D. Từ đó giúp tinh chỉnh hoạt động tìm kiếm trong quá trình phân tích. Ông đã sử dụng phương pháp tiếp cận được đề xuất, nhờ vậy dữ liệu tiếp thị địa lý được phân loại trong cơ sở dữ liệu không gian địa lý để quản lý dữ liệu hiệu quả. Nghiên cứu vào năm 2017 của Leung, Yen và Lohmann sử dụng dữ liệu khảo sát hành khách từ Sân bay Gold Coast ở bang Queensland, Australia, để thực hiện phân tích phân loại địa lý nhân khẩu học kết hợp với dữ liệu điều tra dân số. Với dữ liệu sở thích của hành khách được mã hóa địa lý, các đặc điểm chuyến đi và sở thích về quyết định của sân bay được so sánh chéo với dữ liệu nhân khẩu học và các biến kinh tế xã hội. Họ lập bản đồ các khu vực mà khách hàng sinh sống dựa trên các điểm đến mà họ bay đến. Kết quả cho thấy sự trái ngược, đặc biệt về vị trí xuất phát của hành khách đối với các chuyến đi nội địa chặng ngắn và các chuyến đi quốc tế đường dài, trong đó hành khách từ xa sẵn sàng đi đường dài để đến sân bay hạng hai để tận dụng giá vé máy bay rẻ hơn (Leung et al., 2017).

3. Phương pháp nghiên cứu

Thứ tự các phương pháp được sử dụng trong bài nghiên cứu này như sau:

a. Thu thập dữ liệu dân số (tập tin CSV) và cơ sở dữ liệu địa lý (tập tin JSON)

b. Thực hiện xử lý thô và làm sạch dữ liệu

c. Thực hiện phân tích dữ liệu dân số bằng phương pháp Phân tích thành phần chính (PCA)

d. Các thành phần chính thu được thông qua phân tích PCA được sử dụng để xác định số lượng các cụm

e. Thực hiện phân cụm K-means từ dữ liệu dân số

f. Tối ưu số cụm n từ phân cụm K-means bằng phương pháp Elbow

g. Tìm chính xác số cụm n từ phân cụm K-means bằng phương pháp Silhouette

h. Thực hiện phân nhóm các Phường, Xã, Thị trấn vào các cụm

i. Liên kết bảng dữ liệu dân số với các điểm không gian (các đa giác) trong cơ sở dữ liệu địa lý

j. Thực hiện vẽ bản đồ phân khúc địa lý

3.1. Phương pháp Phân tích thành phần chính (PCA)

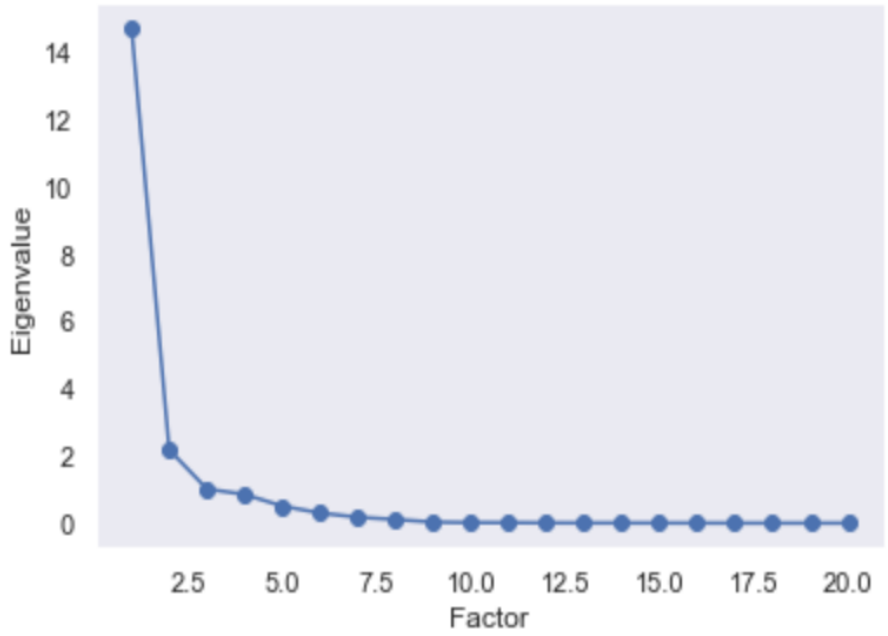
Bước 1: Sử dụng phân tích nhân tố (Factor Analysis) để xác định tải trọng (loadings) và giá trị riêng (eigenvalue)

$$Loadings = Eigenvectors.\sqrt{Eigenvalues}$$

Trong đó:

- $Loadings$: Tải trọng, là các hiệp phương sai/ tương quan giữa các biến ban đầu và các thành phần tỷ lệ đơn vị, giúp giải thích các thành phần chính, yếu tố vì chúng là các trọng số kết hợp tuyến tính (hệ số) theo đó các thành phần hoặc các yếu tố được chia tỷ lệ đơn vị xác định hoặc "tải" một biến
- $Eigenvectors$: Vectơ riêng, là một vectơ khác vectơ không mà được nhân với một hệ số vô hướng khi biến đổi tuyến tính đó được áp dụng lên nó
- $Eigenvalues$: Hệ số vô hướng vô hướng áp dụng lên vectơ riêng

Bước 2: Sử dụng lược đồ Scree Plot để xác định số thành phần chính của tập dữ liệu.



Hình 1. Lược đồ Scree Plot

Nguồn: Nhóm nghiên cứu tổng hợp qua Python 3

Kết quả từ biểu đồ Scree cho thấy, chúng ta nên giữ lại từ 3 đến 4 thành phần chính để giá trị riêng **(**Eigenvalue) gần 1. Trong bài nghiên cứu này, chúng tôi sử dụng **03 thành phần chính.**

Lưu ý: Tùy thuộc vào mục đích mà ta lựa chọn số thành phần chính cho phù hợp, giải pháp trên chỉ là một phương pháp đưa ra số thành phần chính gợi ý. (Số thành phần chính càng nhiều thì càng giải thích đầy đủ hơn cho tập hợp các biến ban đầu).

Bước 3: Thực hiện phân tích PCA với 03 thành phần chính

```
pca.explained_variance_ratio_
```

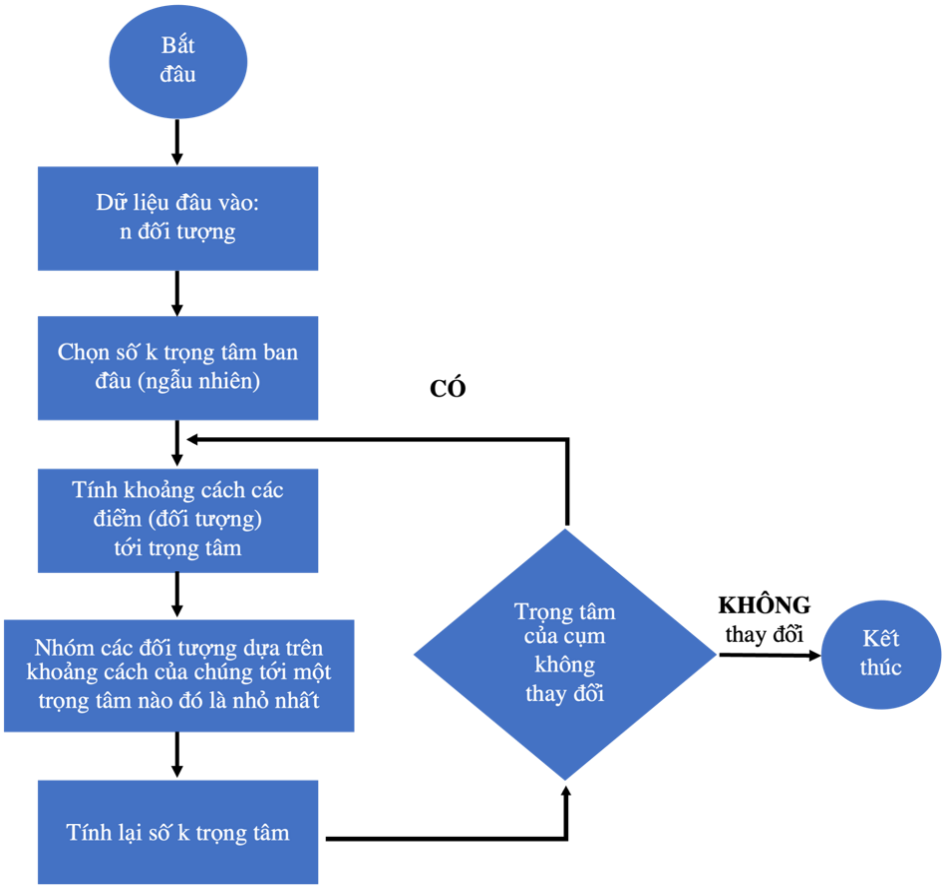
```
array([0.73861095, 0.10929314, 0.05082396])
```

Hình 2. Tỷ lệ phương sai với 03 thành phần chính

Nguồn: Nhóm nghiên cứu tổng hợp qua Python 3

Ta giải thích tỷ lệ phương sai với 03 thành phần chính. Ta có thể thấy rằng thành phần chính đầu tiên giải thích 74,86% và độ biến thiên tổng thể. Thành phần chính thứ hai và thứ ba lần lượt giải thích 10,93% và 5,08% độ biến thiên tổng thể. Cùng với nhau, hai thành phần giải thích **90,87%** tổng biến.

3.2. Phương pháp phân tích cụm – thuật toán K-means



Hình 3. Sơ đồ thuật toán K-means

Nguồn: Nhóm nghiên cứu tổng hợp

Input: Số cụm **k** và các trọng tâm cụms $\{m_j\}$; $k_j = 1$

Output: Các cụm $C[i]$ $(1 \leq i \leq k)$ và hàm tiêu chuẩn E đạt giá trị tối thiểu.

Begin

Bước 1: Khởi tạo

Chọn **k** trọng tâm $\{m_j\}$ $(1 \leq j \leq k)$, ban đầu trong không gian R_d (d là số chiều của dữ liệu). Việc lựa chọn này có thể là ngẫu nhiên hoặc theo kinh nghiệm.

Bước 2: Tính khoảng cách

Đối với mỗi điểm X_i $(1 \leq i \leq n)$, tính khoảng cách của nó tới mỗi trọng tâm $\{m_j\}$ $(1 \leq j \leq k)$. Sau đó tìm trọng tâm gần nhất đối với mỗi điểm.

Bước 3: Cập nhật lại trọng tâm

Đối với mỗi $1 \leq j \leq k$, cập nhật trọng tâm cụm m_j bằng cách xác định trung bình cộng các vectơ đối tượng dữ liệu.

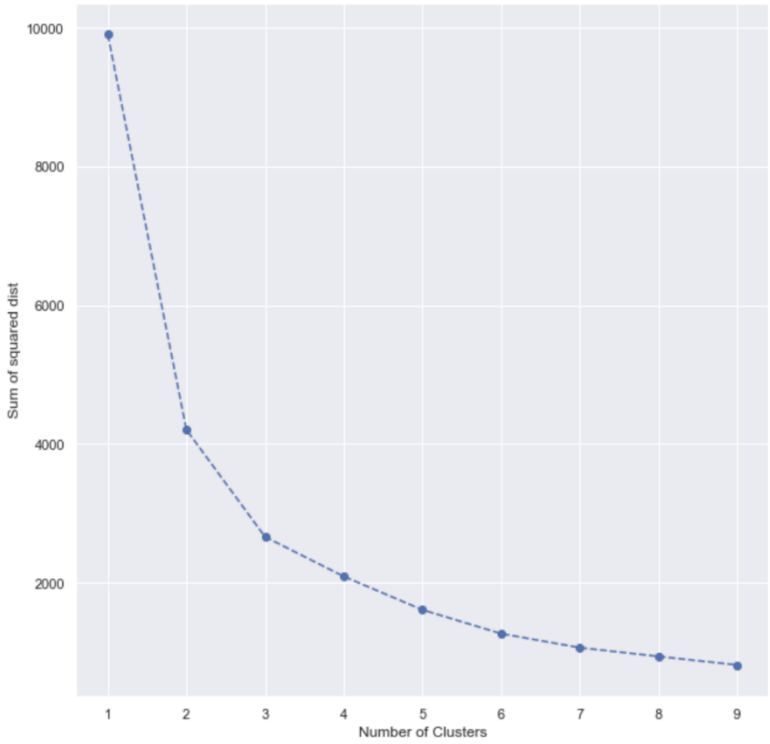
Điều kiện dừng: Lặp lại các bước 2 và 3 cho đến khi các trọng tâm của cụm không thay đổi.

Tuy nhiên, trong bài nghiên cứu này, giá trị của **k** được tìm tự động trên Python qua các lệnh, và số cụm tối ưu được tìm dựa trên phương pháp Elbow (Elbow method).

Phương pháp Elbow

Dựa vào đường cong Elbow, số **k** thích hợp là vị trí ở khúc cua (bend/knee) của đường. Tại điểm này, giá trị của khoảng cách trung bình không có sự thay đổi đáng kể khi số cụm **k** tăng.

Trong biểu đồ sau, có thể nhìn thấy rõ giá trị của **k** hợp lý nằm trong khoảng 3 hoặc 4. Để có thể tìm được chính xác số **k** tối ưu, ta sẽ tiếp tục sử dụng phương pháp Silhouette để kiểm tra các trường hợp k = 3, 4.



Hình 4. Đường cong Elbow

Nguồn: Nhóm nghiên cứu tổng hợp

Phương pháp Silhouette

Ở trên, với việc sử dụng phương pháp Elbow, ta thấy số lượng cụm thích hợp nhất giao động trong khoảng 3 hoặc 4. Ta tiếp thực hiện phương pháp Silhouette để kiểm tra tính nhất quán trong từng trường hợp.

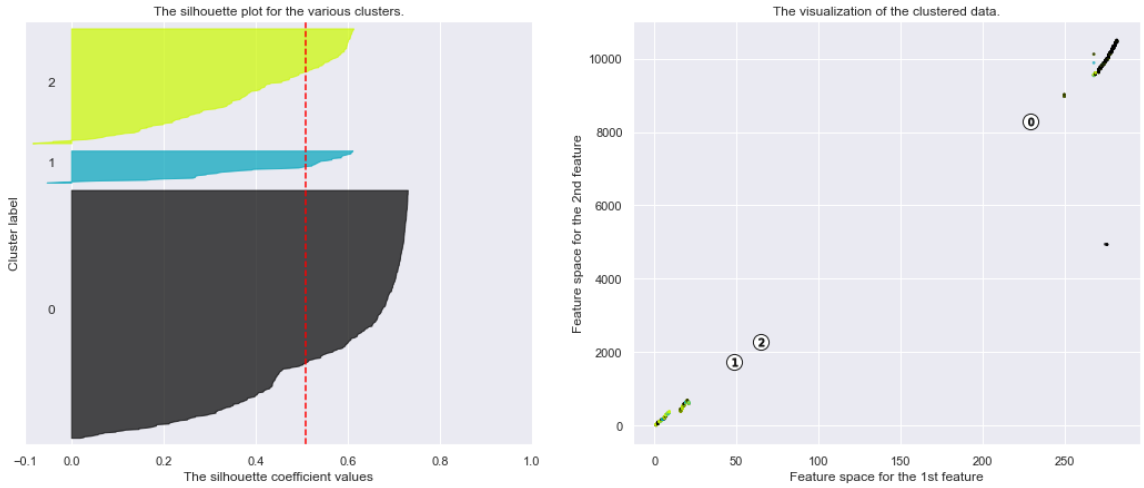
```
For n_clusters = 3 The average silhouette_score is : 0.5100543174016334
For n_clusters = 4 The average silhouette_score is : 0.47884852105752135
```

Hình 5. Silhouette score với số cụm = 3, 4

Nguồn: Nhóm nghiên cứu tổng hợp

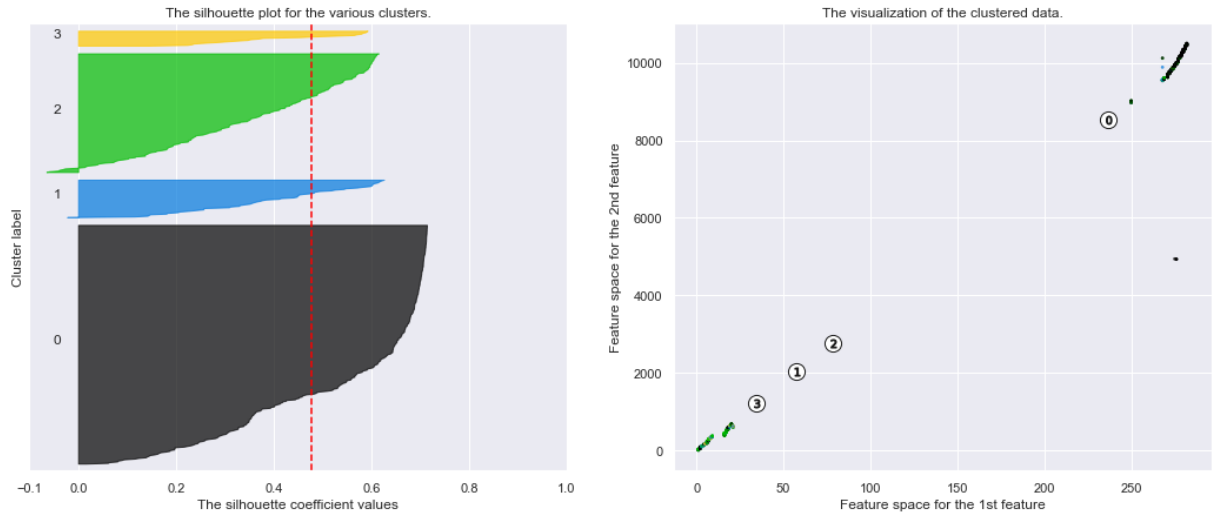
Ta thấy giá trị Silhouette score trong trường hợp 3 cụm và 4 cụm là tương đương với giá trị lần lượt là 0.51 và 0.48.

Quan sát trường hợp số cụm bằng 3 và 4 ở ***Hình 6, Hình 7***, ta thấy độ dày của biểu đồ Silhouette cho cụm có cluster label = 0 có kích thước như nhau. Có thể thấy việc sử dụng 3 cụm hay 4 cụm đều có thể đem lại kết quả mong muốn. Trong bài nghiên cứu này, ta sẽ sử số cụm k = 4 để phân tích được chi tiết hơn.



Hình 6. Biểu đồ hệ số Silhouette với số lượng cụm = 3

Nguồn: Nhóm nghiên cứu tổng hợp



Hình 7. Biểu đồ hệ số Silhouette với số lượng cụm = 4

Nguồn: Nhóm nghiên cứu tổng hợp

4. Kết quả nghiên cứu

4.1. Tổng quát

Trong bài nghiên cứu này, mô hình phân khúc địa lý được áp dụng cho 582 đơn vị hành chính cấp xã thuộc 30 đơn vị hành cấp huyện trên địa bàn Thành phố Hà Nội.

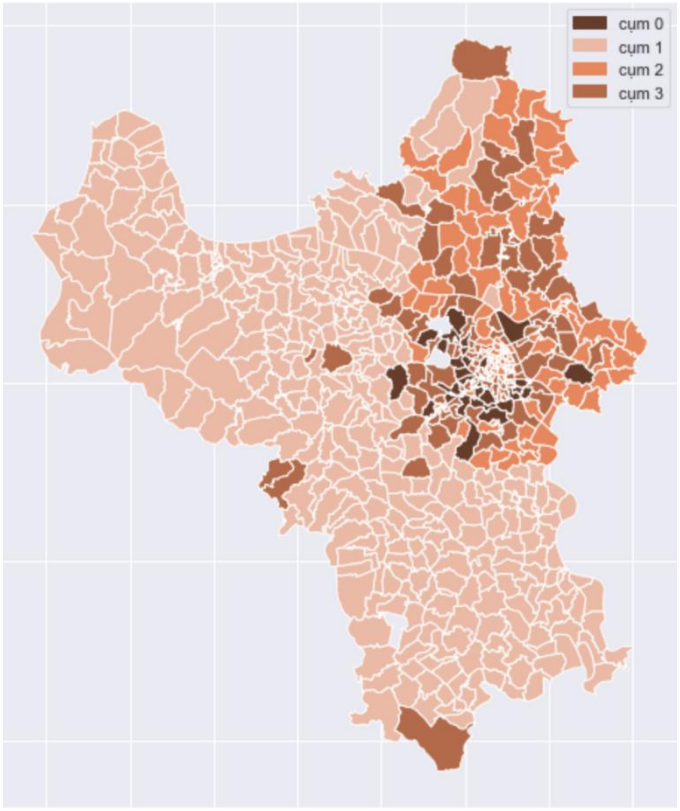
Nghiên cứu đã cố gắng xác định các đặc điểm chung của từng nhóm dựa trên kết quả của ma trận thành phần được tạo ra từ các đơn vị hành chính cấp xã. Do không có mối tương quan giữa các thành phần (cụm) được tạo thành, các thuộc tính của mỗi thành phần có thể được giải thích và xác định một cách độc lập với các thành phần (cụm) khác. Bốn thành phần chính (04 cụm) thu được là các biến phụ thuộc và các dữ liệu mô tả đã được liệt kê ở phần phương pháp nghiên cứu được sử dụng để giải thích các cụm này là các biến độc lập.

Bài nghiên cứu tập trung mô tả một số đặc điểm cơ bản về địa lý, số dân, độ tuổi, số lượng học sinh sinh viên để phân biệt các cụm trong bài nghiên cứu.

Đặc điểm cơ bản của bốn cụm thu được có thể được mô tả khái quát như sau:

Cụm 0 *(30 phường, xã)***:** Đây là khu vực có diện tích bé nhưng tập trung rất đông dân cư dẫn đến mật độ dân số rất cao. Các khu vực thuộc cụm này tập trung ở trung tâm phía Tây Thành phố Hà Nội. Dân cư ở đây chủ yếu thuộc độ tuổi lao động, trình độ cao, có nhiều sinh viên và người đang theo học cấp bậc sau Đại học. Các hộ gia đình là các gia đình trẻ có con nhỏ dưới 10 tuổi.

Cụm 1 *(325 phường, xã)***:** Đây là khu vực có diện tích lớn nhưng tập trung ít dân cư dẫn đến mật độ dân số thấp. Các khu vực thuộc cụm này chủ yếu thuộc các Huyện, chiếm khoảng 70% diện tích, tập trung nhiều ở phía Đông Bắc và phía Nam Thành phố Hà Nội. Dân cư ở đây chủ yếu thuộc độ tuổi lao động nhưng chủ yếu là làm nông và làm nghề tiểu thủ công nghiệp, trình độ dân trí không cao.



Hình 8. Bản đồ phân khúc địa lý thành phố Hà Nội

Nguồn: Nhóm nghiên cứu tổng hợp

Cụm 2 *(128 phường xã)***:** Cụm 2 có mật độ dân số cao nhưng phân bố không đồng đều. Các khu vực thuộc cụm này tập trung ở phía Tây Bắc Thành phố Hà Nội. Dân cư ở đây chủ yếu thuộc độ tuổi lao động, trình độ cao. Đây là cụm có nhiều hộ gia đình sinh sống và kinh doanh tại gia với nhiều hình thức, tuy nhiên tỷ lệ trẻ em dưới 10 tuổi ở đây khá thấp.

Cụm 3 *(101 phường, xã)***:** Đây là khu vực phân bố đồng đều ở trung tâm phía Tây Thành phố Hà Nội, mật độ dân số trung bình, nguồn lao động trình độ cao dồi dào, có nhiều dự án đang được triển khai.

Bảng 1. Bảng thống kê các quận, huyện, thị xã và số lượng phân bố cụm

STT	Tên quận, huyện, thị xã	Chi tiết cụm
Quận, huyện, thị xã có đặc điểm của 1 cụm		
1	Huyện Ba Vì	Cụm 1
2	Huyện Phú Xuyên	
3	Huyện Phúc Thọ	
4	Huyện Thường Tín	
5	Huyện Ứng Hoà	
6	Thị xã Sơn Tây	
Quận, huyện, thị xã có đặc điểm của 2 cụm		
7	Huyện Hoài Đức	Cụm 1 – 0
8	Huyện Chương Mỹ	Cụm 1 – 3
9	Huyện Đan Phượng	

STT	Tên quận, huyện, thị xã	Chi tiết cụm
10	Huyện Mê Linh	
11	Huyện Mỹ Đức	
12	Huyện Quốc Oai	
13	Huyện Thạch Thất	
14	Huyện Thanh Oai	
15	Quận Cầu Giấy	Cụm 0 – 3
16	Quận Ba Đình	
17	Quận Hoàn Kiếm	Cụm 2 – 3
18	Quận Tây Hồ	
Quận, huyện, thị xã có đặc điểm của 3 cụm		
19	Huyện Đông Anh	Cụm 1 – 2 – 3
20	Huyện Sóc Sơn	
21	Huyện Gia Lâm	
22	Huyện Thanh Trì	
23	Quận Bắc Từ liêm	
24	Quận Đống Đa	
25	Quận Hai Bà Trưng	Cụm 0 – 2 – 3
26	Quận Hoàng Mai	
27	Quận Long Biên	
28	Quận Nam Từ Liêm	
29	Quận Thanh Xuân	
Quận, huyện, thị xã có đặc điểm của 4 cụm		
30	Quận Hà Đông	Cụm 0 – 1 – 2 – 3

Nguồn: Nhóm nghiên cứu tổng hợp

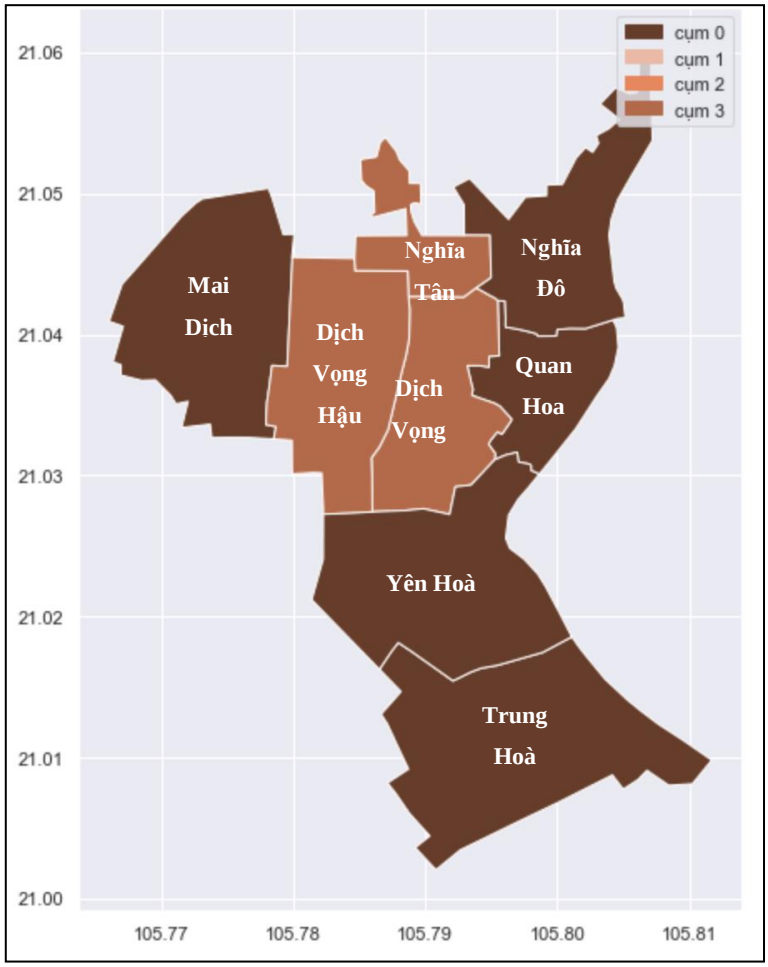
4.2. Phân tích mẫu – Quận Cầu Giấy (Quận có đặc điểm của 2 cụm)

Quận Cầu Giấy gồm có 8 phường thuộc hai cụm:

• Cụm 3: Phường Dịch Vọng, Mai Dịch, Nghĩa Tân

• Cụm 0: Phường Nghĩa Đô, Quan Hoa, Trung Hòa, Yên Hòa, Dịch Vọng Hậu.

Các phường thuộc cụm 3 là nơi tập trung các cơ sở kinh doanh, các đơn vị hành chính, giáo dục quan trọng của Quận, tập trung ở các khu vực có bán kinh từ 4 đến 5km quanh ngã tư Xuân Thủy. Vì nằm ở vị trí trung tâm của quận, nơi có lưu lượng người đi lại trong ngày cao nên các khu vực này cần được đầu tư hơn về phát triển cơ sở hạ tầng giao thông và các khu trung tâm thương mại phục vụ một số lượng học sinh, sinh viên khá đông bên cạnh các cơ sở kinh doanh nhỏ lẻ (hiện nay mới chỉ có khu Discovery Complex và khu Indochina Plaza Hà Nội).



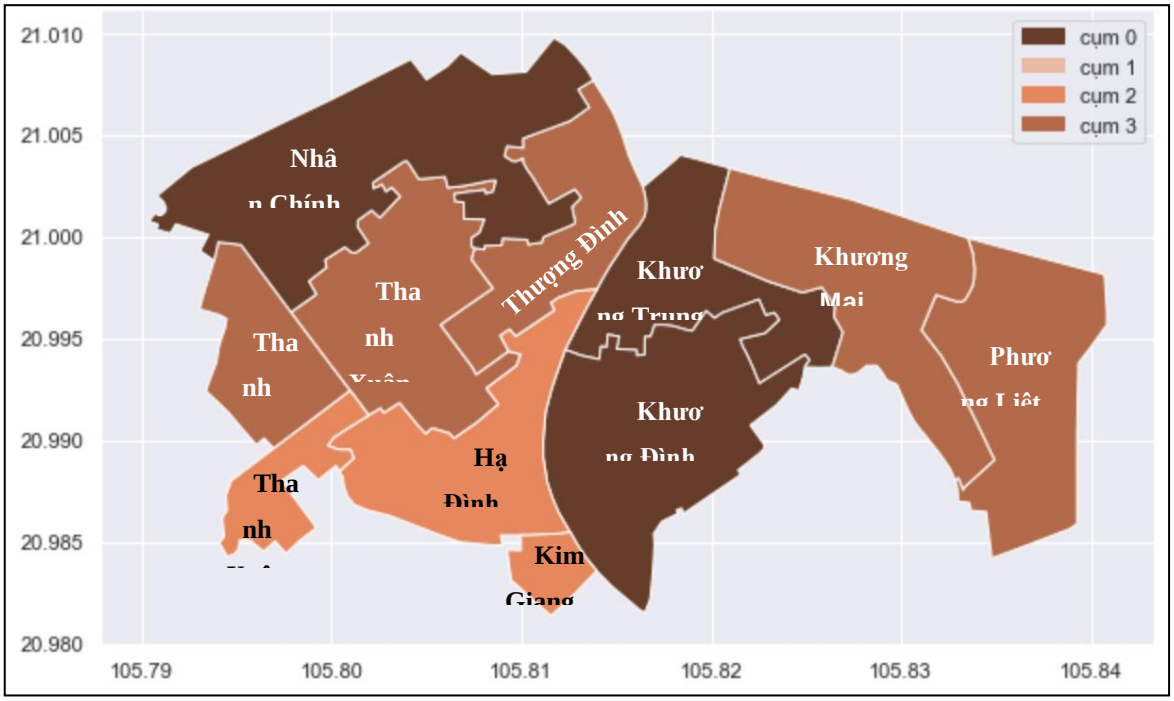
Hình 9. Bản đồ phân khúc địa lý quận Cầu Giấy

Nguồn: Nhóm nghiên cứu tổng hợp

Khu vực các phường thuộc **cụm 0** là nơi có mật độ dân cư sinh sống cao, có nhiều khu chung cư tập trung chủ yếu là người dân học tập và làm việc ở quận Cầu Giấy. Khu vực xung quanh đoạn tiếp giáp phường Yên Hòa và phường Dịch Vọng, Dịch Vọng Hậu đang trong quá trình đầu tư phát triển cơ sở hạ tầng cho đặc khu Công Nghệ Thông Tin với các tập đoàn lớn như Viettel hay FPT, dần trở thành khu vực kinh tế mũi nhọn của quận, thu hút thêm các doanh nghiệp công nghệ cao tạo ra sản phẩm giá trị gia tăng có sức cạnh tranh, hạn chế thấp nhất những tác động xấu đến môi trường khu vực. Tuy nhiên, cần phải có giải pháp xử lý một số công trình dân cư bị trì trệ, ảnh hưởng đến cảnh quan đô thị và đời sống người dân khu vực lân cận.

Với những đặc điểm kể trên Quận Cầu Giấy trở thành khu đô thị mới tập trung các hoạt động kinh tế của Thủ đô Hà Nội. Bên cạnh đó, vẫn cần duy trì công tác bảo tồn các di sản văn hóa, các làng nghề cổ truyền như Chùa Hà, Làng Vòng, …

4.3. Phân tích mẫu – Quận Thanh Xuân (Quận có đặc điểm của 3 cụm)



Hình 10. Bản đồ phân khúc địa lý quận Thanh Xuân

Nguồn: Nhóm nghiên cứu tổng hợp

Quận Thanh Xuân gồm 11 phường:

- Cụm 0: Phường Nhân Chính, Khương Đình, Khương Trung
- Cụm 2: Phường Hạ Đình, Kim Giang, Thanh Xuân Nam
- Cụm 3: Phường Khương Mai, Phương Liệt, Thanh Xuân Bắc, Thanh Xuân Trung, Thượng Đình

Đời sống ngày càng phát triển, quận tập trung được nhiều các dự án lớn phát triển thị trường địa ốc trên toàn khu vực Hà Nội và nhiều tuyến đường huyết mạch như Lê Văn Lương, Khuất Duy Tiến,… **Cụm 0** của quận là khu vực đất chật người đông, mật độ dân số cao, có nhiều hộ gia đình trên 3 người, tập trung nhiều khu đô thị lớn: Khu đô thị Trung Hòa Nhân Chính, Khu đô thị Mandarin Garden,... Khu vực này là nơi có giao thông thuận tiện, giá thuê nhà hợp lý, đây nơi tập trung nhiều công ty vừa và nhỏ, cửa hàng tiện ích cũng như dân văn phòng, dân nhập cư về thủ đô sinh sống và học tập.

Cụm 2 có diện tích bé, dân số ít so với các vùng khác trên địa bàn quận. Khu vực này có nhiều ngõ ngách nhỏ, đường xá chật hẹp, có nhiều khu đất trống, nhà cấp 4 do đất đó là đất nông nghiệp chưa được chuyển đổi, người dân chỉ xây nhà ở tạm, hoặc dãy nhà trọ cho thuê.

Các phường ở **cụm 3** có trình độ giáo dục cao, tập trung nhiều trường đại học như Đại học Khoa học Tự nhiên, Đại học Khoa học Xã hội Nhân văn, ... Bên cạnh đó, đây là nơi tập trung những bệnh viện khám và chữa bệnh hàng đầu tại Việt Nam, nên dân cư các tỉnh về đây khám chữa bệnh rất đông.

4.4. Phân tích mẫu – Quận Hà Đông (Quận có đặc điểm của 4 cụm)

Quận Hà Đông gồm 17 phường:

- Cụm 0: Phường La Khê, Mộ Lao

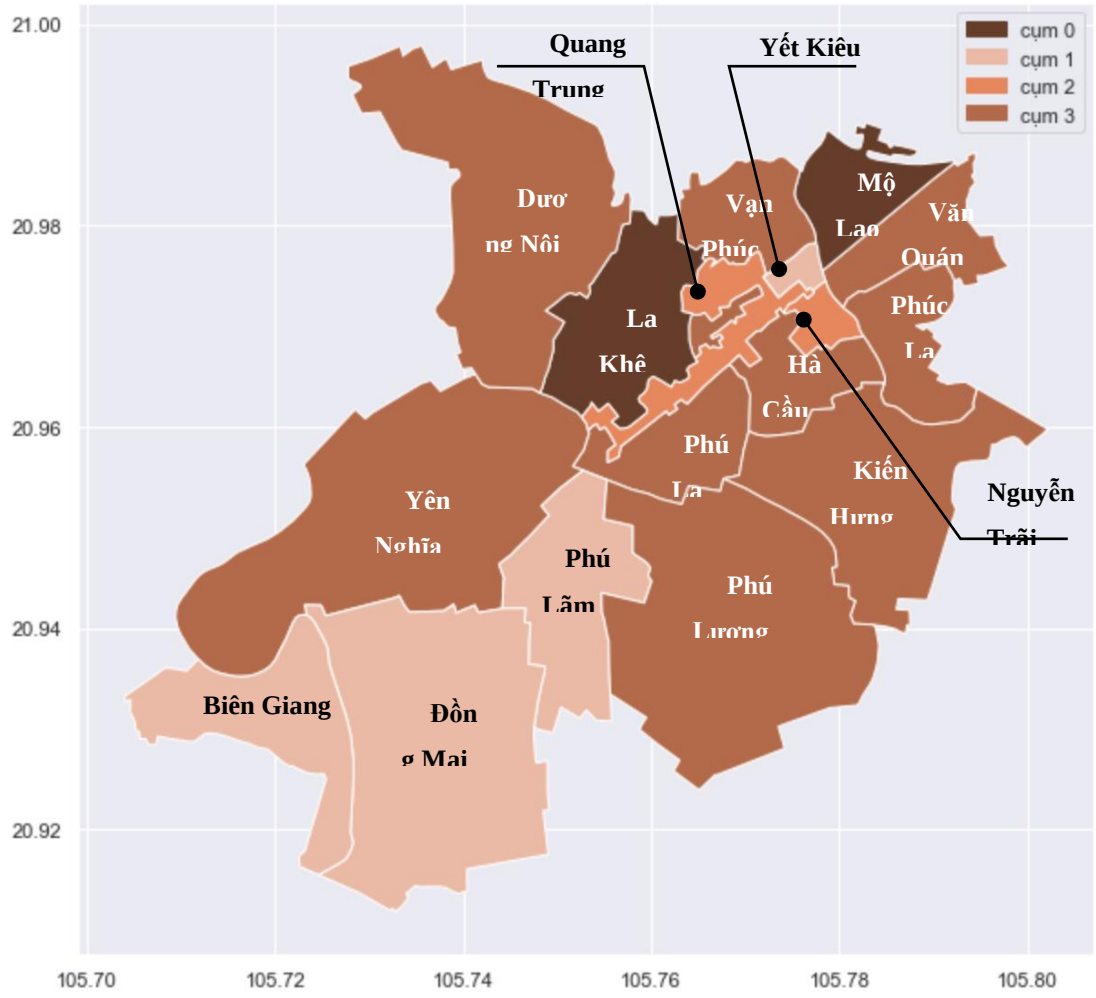
- Cụm 1: Phường Biên Giang, Đồng Mai, Phú Lãm, Yết Kiêu
- Cụm 2: Phường Nguyễn Trãi, Quang Trung
- Cụm 3: Các phường còn lại (xem trên bản đồ ***Hình 9***)

Các phường thuộc **cụm 0** đều là những địa bàn có nhiều cơ quan, đơn vị, trường học thuộc thành phố đóng trụ sở và có một số tuyến đường lớn chạy qua như trục quốc lộ 6, trục đường Quang Trung, Lê Trọng Tấn hay Lê Văn Lương.

Hai phường La Khê và Mộ Lao qua quá trình quy hoạch xây dựng mới đã nhanh chóng phát triển thêm theo hướng thương mại dịch vụ. Cơ sở hạ tầng ở những địa phương này cũng được cải thiện đáng kể, tạo điều kiện thuận lợi về mặt di chuyển. Bởi vậy, doanh nghiệp cần lượng khách hàng lớn như vận tải (giống GHTK), trung tâm thương mại, trung tâm giáo dục hay kinh doanh nhỏ lẻ có nhiều cơ hội phát triển hơn ở **cụm 0** này.

Những phường được xếp vào **cụm 1** hoặc là nơi cách quá xa trung tâm hoặc là nơi nằm giữa những phường khác, không có đường lớn đi qua như phường Yết Kiêu. Hầu hết các doanh nghiệp có cơ hội phát triển ở khu vực này thường nghiêng về mặt kinh doanh, buôn bán nhỏ. Trong đó, phường Yết Kiêu tập trung dân cư của những khu tập thể nhà máy như nhà cơ khí nông cụ. Vì vậy, nơi đây tạo nhiều cơ hội hơn cho các doanh nghiệp nhỏ hoặc các doanh nghiệp cần mở thêm chi nhánh ở những nơi đông dân như phòng khám tư, trung tâm giáo dục, chi nhánh ngân hàng.

Khu vực thuộc **cụm 2** là phường Quang Trung và phường Nguyễn Trãi. Hai phường này đều nằm trên trục đường to Quang Trung – Trần Phú – Nguyễn Trãi, nối liền Hà Đông với trung tâm Hà Nội. Dân cư sống ở đây nhìn chung đều có trình độ học vấn và thu nhập cao vì đa phần đều có cửa hàng kinh doanh nhỏ lẻ, đa dạng thuộc mọi ngành nghề. Khu vực quanh trục đường lớn vẫn là những nơi đất vàng, tập trung nhiều sự chú ý của các doanh nghiệp nhờ vào vị trí địa lý thuận lợi đối với việc tiếp cận khách hàng. Bởi vậy, khu vực này phù hợp với hầu hết các lĩnh vực thuộc ngành thương mại, dịch vụ.



Hình 11. Bản đồ phân khúc địa lý quận Hà Đông

Nguồn: Nhóm nghiên cứu tổng hợp

Mật độ dân cư ở khu vực thuộc **cụm 3** ở mức độ trung bình, vừa phải. Những phường này đều có nhiều khu đô thị mới, nhà cao tầng với nhiều dự án triển vọng cao trong tương lai khi mà dân số ở quận Hà Đông đang tăng lên một cách nhanh chóng và khiến cho việc cung cấp nhà ở khu thuộc **cụm 2** trở nên khó khăn. Những nơi này đã xuất hiện một số doanh nghiệp nhỏ, chủ yếu về bất động sản hay kinh doanh tạp hoá, quần áo nhỏ lẻ. Bởi vậy, vẫn còn rất nhiều không gian và cơ hội cho những doanh nghiệp thuộc các lĩnh vực khác. Dựa vào đối tượng khách hàng tiềm năng của những dự án bất động sản hiện có, khu vực này sẽ là lý tưởng đối với những doanh nghiệp có đối tượng khách hàng có thu thập tầm trung và cao.

5. Kết luận

Với sự phát triển nhanh chóng của công nghệ thông tin cùng nền kinh tế càng ngày càng phức tạp và được mở rộng, sự tồn tại của các doanh nghiệp trong bất kể lĩnh vực nào đều tỉ lệ thuận với sự cạnh tranh của nó trên thị trường. Để nâng cao lợi thế cạnh tranh của mình so với đối thủ, các doanh nghiệp cần nhận định tầm quan trọng của việc xác định mọi yếu tố liên quan tới đối tượng khách hàng tiềm năng. Kết hợp với công nghệ quản lý thông tin địa lý, các nhà quản trị có thể đưa ra quyết về địa điểm tiếp cận và quảng bá thương hiệu tới khách hàng một cách hiệu quả rút ra từ dữ liệu dân cư của mỗi khu vực cụ thể.

Dựa vào cơ sở lý thuyết phân cụm K-means, phép phân tích thành phần chính (PCA) với sự hỗ trợ của ứng dụng lập trình Python, bài nghiên cứu này đã hoàn thành phân tích dữ liệu dân số Thành phố Hà Nội năm 2020 quy lại thành bốn cụm có những đặc trưng riêng và chỉ ra cụm phổ biến ở các quận trong cơ sở dữ liệu địa lý Thành phố Hà Nội. Từ đó, tạo ra mô hình phân khúc thị trường theo địa lý cho những doanh nghiệp có mong muốn tìm hiểu và hoạt động ở khu vực Hà Nội trong tương lai.

Mặc dù đã đạt được kết quả như trên, công trình nghiên cứu vẫn còn hạn chế nhiều về số liệu thống kê thỏa mãn tiêu chí phân tích cũng như hạn chế về thời gian và tiền bạc. Do đó, sẽ không tránh được những nhận định, bình luận mang tính suy đoán.

Sở dĩ tính liên ngành của dữ liệu địa lý và sự đa dạng trong ứng dụng của nó, thông tin, số liệu được thu thập hiện tại có thể qua sự thay đổi để phù hợp phân tích cho các ngành khác nhau, đặc biệt là ngành marketing. Nên đề xuất của nghiên cứu có thể chưa hoàn toàn thuyết phục khi khái quát cho khu vực Hà Nội hay các khu vực khác ở Việt Nam. Vì vậy, những nghiên cứu trong tương lai sẽ cần phải được đầu tư vào thời gian, công sức và tiền bạc cũng như phân tích chi tiết kỹ hơn.

Tài liệu tham khảo

Abdi, H. & Williams, L.J. (2010), "Principal component analysis", *Wiley interdisciplinary reviews Computational Statistics*, Vol. 2 No. 4, pp. 433 – 459.

Allo, N.B. (2012), "The potential and prospects for enabling small area geodemographics and Geo-marketing in developing countries: a case study on Nigeria", Unpublished PhD Thesis, Kingston University.

AlloHenn, K., Tommi, K. & Raija, K. (2011), "Using ski destination choice criteria to segment Finnish ski resort customers", *Tourism Management*, Vol. 32, pp. 1096 – 1105.

Banerjee, S. (2019), "Geo-marketing and situated consumers: opportunities and challenges. In Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-based Recommendations", *Geosocial Networks and Geoadvertising*, pp. 13.

Fisher, P. & Tate, N.J. (2015), "Modelling class uncertainty in the geodemographic Output Area Classification", *Environment and Planning B: Planning and Design*, Vol. 42 No. 3, pp. 541 – 563.

Hadong.hanoi.gov.vn. (n.d.), "Tổng quan về Hà Nội", https://hadong.hanoi.gov.vn/english/en/overview.html, truy cập ngày 17/05/2021.

Hartigan, J.A. & Wong, M.A. (1979), "Algorithm AS 136: A *k*-Means Clustering Algorithm", *Journal of the Royal Statistical Society, Series C.*, Vol. 28 No. 1, pp. 100 – 108.

Jinsoo, H., Young, G.C., Junghoon, J.L. & Jongseung, P. (2012), "Customer Segmentation Based on Dining Preferences in Full-Service Restaurants", *Journal of Foodservice Business Research*, Vol. 15, pp. 26 – 246.

Kassambara, A. (2017), *Practical guide to cluster analysis in R: unsupervised machine learning*, Vol. 1, STHDA, France.

Kaufman, L. & Rousseeuw, P.J. (2009), *Finding groups in data: an introduction to cluster analysis*, Vol. 344, John Wiley & Sons.

Kodinariya, T.M., & Makwana, P.R. (2013), "Review on determining number of Cluster in K-Means Clustering", *International Journal*, Vol. 1 No. 6, pp. 90 – 95.

Lansley, G. & Longley, P. (2016), "Deriving age and gender from forenames for consumer analytics", *Journal of Retailing and Consumer Services*, Vol. 30, pp. 271 – 278.

Leung, A., Yen, B.T. & Lohmann, G. (2017), "Why passengers' geo-demographic characteristics matter to airport marketing", *Journal of Travel and Tourism Marketing*, Vol. 34 No. 6, pp. 833 – 850.

Shaffer, A.C. (2015), "The geodemographics in location intelligence: A study in craft brewery placement", PhD Thesis, Northern Arizona University.

Shlens, J. (2014), "A tutorial on principal component analysis".

Suhaibah, A., Uznir, U., Rahman, A.A., Anton, F. & Mioc, D. (2016), "3D Geo-marketing Segmentation: A Higher Spatial Dimension Planning Perspective", *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp. 42.

Thành, H. (2019), "Cầu Giấy: 21 năm xây dựng và phát triển", Tiền Phong, https://tienphong.vn/
cau-giay-21-nam-xay-dung-va-phat-trien-post1085406.tpo, truy cập ngày 06/05/2021.

Thinsungnoena, T., Kaoungkub, N., Durongdumronchaib, P., Kerdprasopb, K., & Kerdprasopb, N. (2015), "The clustering validity with silhouette and sum of squared errors", *International Conference on Industrial Application Engineering 2015*, Vol. *3* No. 7.

Tổng cục thống kê. (2019), Dữ liệu dân số Thành phố Hà Nội năm 2019.

VNGIS. (2019), Cơ sở dữ liệu địa lý Thành phố Hà Nội năm 2019.