Advances in Economics, Business and Management Research, volume 196

Proceedings of the International Conference on Emerging Challenges: Business Transformation and Circular Economy (ICECH 2021)

Building a Geo-Demographic Segmentation Model: the Case of Hanoi City, Vietnam

Le Thu HANG[1*], Bui Nguyen Anh TUAN[2], Van Duc MANH[3], Nguyen Quynh CHI[4], Bui Thien BINH[5] and Tran Ngoc DIEP[6]

[1] PhD Candidate, Faculty of Business Administration, Foreign Trade University, Hanoi, Vietnam

[2] PhD Candidate, Vietnam Competition Council, Ministry of Industry and Trade, Vietnam

[3,4,5,6] English 1 Advanced Program – K57, Faculty of Business Administration, Foreign Trade University, Hanoi, Vietnam

[*]Corresponding author: hanglt@ftu.edu.vn

Abstract

Location is one of the most crucial factors determining enterprises' strategies when entering, exploiting or expanding into 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 demographic characteristics of each area such as age, occupation, education level, etc.; thereby giving enterprises precise sources of information about geographic location, which helps reducing the cost and time in decision-making process to enter or expand the businesses in a new area.

Research purpose:

In the rapid development of today's society, especially the boom of information technology, our economy is becoming more complex, the market is expanding, and the competition is becoming more and more fierce. This trend requires businesses in any industry to make full use of all resources and opportunities to gain a competitive advantage in the market. One of the most important and prerequisite things to ensure the success of a business is identifying and reaching the right potential customers. More sepcfically , one of the most popular methods to reach customers is to find a geographical location that fits the needs of the business.

Because the model of market segmentation by geographic location and population is highly applicable to business activities, especially in identifying the right customers, many businesses have conducted research and came up with the geographic models that best fit their strategies. However, these studies are not widely published and cannot be applied to the activities of other businesses. Therefore, with the desire to provide an accurate and customized analytical model of the market distribution in different districts in Hanoi, our research team decided to choose the topic ***Building a geo-demographic segmentation model: The case of Hanoi****. Based on this model, businesses in Vietnam, especially Hanoi, can seek for new potential markets and suitable areas to help expand their businesses geographically.*

Research motivation:

Because the model of market segmentation by geographic location and population is highly applicable to business activities, especially in identifying the right customers, many businesses have conducted research and came up with the geographic models that best fit their strategies. However, these studies are not widely published and cannot be applied to the activities of other businesses. Therefore, with the desire to provide a general, accurate, and detailed analytical model of the market distribution in districts in Hanoi, our research team decided to choose the topic ***Building a geo-demographic segmentation model: The case of Hanoi****. Based on this model, businesses in Vietnam, especially Hanoi, can seek for new potential markets and suitable areas to help expand their businesses geographically.*

Research design, approach and method:

The order of methods used in this study is as follows:

1. Population data collection (CSV file) and geodatabase (JSON file)

2. Perform raw processing and data cleaning

3. Perform population data analysis using Principal Component Analysis (PCA) method

4. Principal components obtained through PCA analysis were used to determine the number of clusters

5. Perform K-means clustering from population data

6. Optimizing the number of clusters n from K-means clustering by Elbow method

7. Find the exact number of clusters n from K-means clustering using Silhouette method

8. Grouping wards, communes and townships into clusters

9. Link the population data table to the spatial points (polygons) in the geodatabase

10. Perform geographic segmentation mapping

Main findings:

In this study, the geographic segmentation model is applied to 582 commune-level administrative units of 30 district-level administrative units in Hanoi city.

The study tries to identify the common characteristics of each group based on the results of the component matrix generated from the commune-level administrative units. Since there is no correlation between the components (clusters) formed, the properties of each component can be interpreted and determined independently of the other components (clusters). The four main components (04 clusters) obtained are dependent variables and the descriptive data listed in the research methods section used to explain these clusters are independent variables.

The study focuses on describing some basic characteristics of geography, population, age, number of students to distinguish clusters in the research paper.

Practical/managerial implications:

With the rapid development of information technology and the complex growth and expansion of economy, the existence of enterprises in any fields is directly proportional to their competition in the market. To enhance their competitive advantages over competitors, businesses need to realize the importance of identifying all factors related to their potential customers. Combined with geographic information management technology, administrators can effectively make decisions about where to reach and promote brands to customers, which is drawn from population data of each specific area.

Based on K-means clustering theory, Principal Component Analysis (PCA) with the assistance of Python programming application, this study has completed the analysis of population data of Hanoi in 2020 dicvided into four clusters with their own characteristics and shows the common clusters in the districts based on geodatabase of Hanoi City. Thence, creating a geographic market segmentation model for businesses wishing to learn and operate in the Hanoi area in the future.

Keywords: Geo-demographic segmentation, K-means clustering, Big data, Principal Component Analysis, Location analysis

1. INTRODUCTION

In today's rapidly developing society, our economy is increasingly complex, the market is expanding, and the competition is becoming more and more fierce, especially with the boom of information technology. This requires businesses in any industry to make full use of all resources and opportunities to gain a competitive advantage in the market. Thus, identifying and reaching the right potential customers are two of the most important steps to ensure the success of a business. A popular method to accomplish this is to find a geographical location that fits the needs of the business.

Because the model of market segmentation by geographic location and population is highly applicable to business activities especially in identifying the right customers, many businesses have conducted research and produced geographic models that best fit their strategies. However, these studies are not widely published and cannot be applied to the activities of other businesses. Therefore, with the desire to provide a general, accurate, and detailed analytical model of the market distribution in districts in Hanoi, our research team decided to build a geodemographic segmentation model.

Based on the latest available population and housing 2019 Census data in Hanoi, we came up with 4 clusters for the 582 basic administrative units by using k-means clustering technique. From the results of this study (possibly first of the kind in Vietnam), businesses in Vietnam, especially in the city of Hanoi, can seek suitable potential markets and areas in which to expand.

2. LITERATURE REVIEW

In international academia there have been many studies on geo-segmentation. These studies have highlighted many practical applications of geo-segmentation and developed them in different ways. These studies can be divided into 3 main parts: (1) application of geographic segmentation in marketing; (2) approach and methodology used in previous studies; (3) aggregate results of geographic segmentation in those studies. First, many studies have focused on **Geographic Marketing** (Geo-Marketing, or GM) with different approaches. Miller et al. (2014) describes the implementation and evaluation of Geographic Information Systems (GIS) at a regional university as well as their potential for deployment in other marketing departments. In 2016, Lansley and Longley investigated the age and sex distribution of English proper names and identified key trends in British naming conventions. Age and gender characteristics are known to have a great influence on consumer behavior, so extracting and using names to find these characteristics from consumer data sets is of immense value to the retail and marketing industries. The results from the extraction can be used to infer the expected age and sex structures of many consumer data sets, as well as to predict key consumer characteristics at the individual level. A 2019 study by Banerjee summarized a number of marketing methods for when there is data on the locations of consumers. Specifically, the study discussed the role of time and social real-time context in advertising effectiveness, utility of positioning tools in clarifying advertising transparency, consumer segmentation, and concerns over location privacy (Banerjee, 2019).

Second, previous studies have used a variety of approaches and methodologies related to **Geographic Segmentation by Market**. In 2011, Konu et al. used selection criteria for a ski destination to segment customers of ski resorts in Finland. Allo (2012) demonstrates the feasability of geo-marketing and geo-segmentation in developing countries through the use of dasymetric mapping to obtain socio-economic maps of the Shomolu region of Nigeria. This is a potential solution for population density mapping in relation to residential land use. Dasymetric mapping describes quantitative zonal data using boundaries that divide an area into rela-tively homogeneous regions for the purpose of better depicting population distribution. A study conducted in the same year evaluated geolocation segmentation and its applications in general terms and emphasized that GIS can be sued more specifically and in a region-specific manner (Longley, 2012). Around the same time, Hwang et al. (2012) identified factors that influence five groups of determinants – dining menu, atmosphere, price, health, and brand reputation – that customers consider when choosing a full-service restaurant.

Finally, the method of **clustering** is quite popular in geo-marketing studies. Fisher and Tate (2015) compared the clustering algorithms used in the 2001 UK Office for National Statistics (ONS) population-based demographic classification studies. They show that both c-means and fuzzy c-means make the results of market segmentation based on geographic region more successful and significant. Shaffer (2015) surveyed craft breweries in the Greater Phoenix area to identify demographic trends, consumer behavior, and spatial relationships in the craft beer market. In 2016, Suhaibah et al. proposed a combination of geographic-based market segmentation and clustering algorithms for 3D geo-marketing data management. This helps refine the search during the analysis. They used the recommended approach whereby geo-marketing data is classified in a geospatial database for efficient data management. A 2017 study by Leung, Yen, and Lohmann used passenger survey data from Gold Coast Airport in Queensland, Australia, to perform a geo-demographic taxonomy analysis combined with census data. With geo-coded passenger preference data, trip characteristics and airport decision preferences are compared with demographic data and socio-economic variables. They map the areas where customers live based on the destinations they fly to. The results show contrasts, especially in terms of passenger origin for short-distance domestic trips and long-distance international trips, where remote passengers are willing to travel long distances to reach second-class airports to take advantage of cheaper airfares.

3. RESEARCH METHODS

A note on the data source:

In the Census 2019, the households over the country were surveyed by the General Statistics Office for a long form questionnaire to get more detailed information. From the ‘huge’ dataset, which consists of about more than two million households in Hanoi, we obtained a secondary data set for 582 administrative units (communes/wards/towns). Variables for each household are various (such as administrative district, family size, residence, housing, age, sex, education, occupation etc.) From these variables, we selected 20 basic variables reflecting socioeconomic characteristics and living types for cluster analysis.

The order of methods used in this study are as follows:

1. Process population data (CSV type file) and geodatabase (JSON type file)

2. Perform raw processing and data cleaning

3. Perform population data analysis using (PCA) method

4. Use principal components obtained through PCA to determine the number of clusters

5. Perform K-means clustering from population data

6. Optimize the number of clusters (n) from K-means clustering with the Elbow method

7. Find the exact number of clusters (n) from K-means clustering using Silhouette method

8. Group wards, communes, and townships into clusters

9. Link the population data table to the spatial points (polygons) in the geo-database

10. Perform geographic segmentation mapping

3.1 Principal Components Analysis (PCA)

Step 1: Use factor analysis to determine the loads and eigenvalues.

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

In which:

- Loadings: Loads, which are the covariances/correlations between the original variables and the unit-scaled components which help to explain the principal components. Factors, since they are linearly combined weights (coefficients) whereby components or elements are scaled to a defined unit or load variable.
- Eigenvectors: a non-zero vector that is multiplied by a scalar factor when that linear transformaiton is applied to it.
- Eigenvalues: scalar coefficients applied to eigenvectors.

Step 2: Use the Scree Plot schema to determine the number of principal components of the data set.

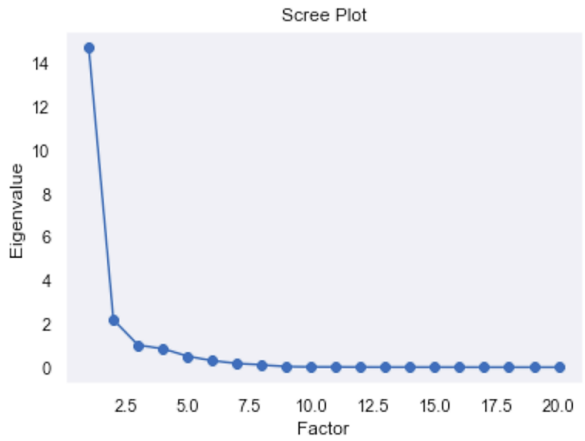

Fig. 1: The Scree Plot map

(Source: Author, 2021)

The results from the Scree chart show that we should keep 3 to 4 main components so that the eigenvalue is close to 1. In this study, we use 3 main components.

Note: The above solution is just a method to give the suggested number of principal components and depends on what our chosen purpose is. (The greater the number of principal components, the more complete the explanation for the original set of variables.)

Step 3: Perform PCA analysis with 3 main components

```
pca.explained_variance_ratio_
```

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

Fig.2. Variance ratio with 3 main components

(Source: Author, 2021)

We explain the ratio of variance with three main components. We can see that the first principal component explains 74.86% of the overall variability. The second and third principal components explain 10.93% and 5.08% of the overall variability respectively. Together, the three components explain 90.87% of the total variability.

3.2 Cluster analysis method – K-means algorithm

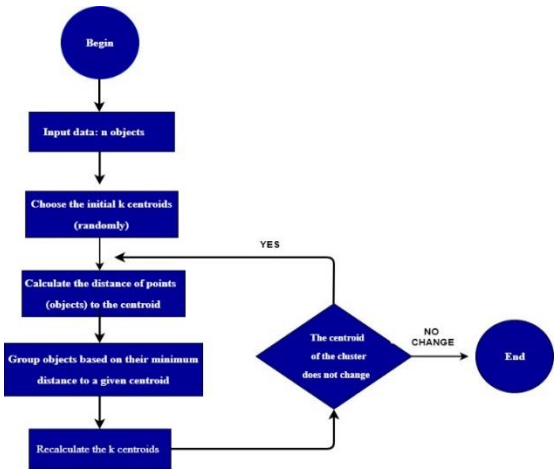

Fig. 3. K-means algorithm diagram

(Source: Author, 2021)

Input: Number of clusters k and cluster centroids $\{mj\}$; = 1**Output:** The clusters $C[i]$ $(1 \leq i \leq k)$ and the standard function E reach the minimum value.

Begin

Step 1: Initialize

Choose k centroids $\{mj\}$ $(1 \leq j \leq k)$, initially in Rd space (d is the number of dimensions of the data). This selection can be random or empirical.

Step 2: Calculate the distance

For each point Xi $(1 \leq i \leq n)$, calculate its distance to each centroid $\{mj\}$ $(1 \leq j \leq k)$. Then find the closest centroid to each point.

Step 3: Update the focus again

For each $1 \leq j \leq$, update the cluster centroid mj by determining the mean of the data feature vectors.

Stop condition: Repeat steps 2 and 3 until the centroids of the cluster do not change.

However, in this study the value of k is found automatically in Python through commands, and the optimal number of clusters is found based on the Elbow method.

Elbow method

Based on the Elbow curve, the appropriate number k is the position at the bend (knee) of the road. At this point, the value of the mean distance does not change significantly as the number of clusters k increases.

In the following diagram, it is clearly seen that a reasonable value of k is in the range 3 or 4. To be able to find the optimal k exactly, we will continue to use the Silhouette method to check the cases of k = 3 and k = 4.

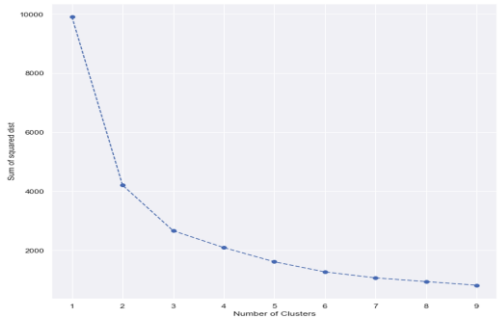

Fig. 4: Elbow curve

(Source: Author, 2021)

Silhouette method

Above, using the Elbow method, we see that the number of most suitable clusters fluctuates between 3 or 4. We continue to implement the Silhouette method to check the consistency in each case.

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

Fig. 5: Silhouette score with cluster count = 3.4

(Source: Author, 2021)

We see that the Silhouette score in the case of 3 clusters and 4 clusters is equivalent to 0.51 and 0.48 respectively.

Observing the case where the number of clusters is equal to 3 in Fig. 6 and 4 in Fig. 7, we see that the thickness of the Silhouette plot for clusters with cluster label = 0 has the same size. It can be seen that using 3 clusters or 4 clusters can bring the desired results. In this study, we will use the number of clusters k = 4 to analyze in more detail.

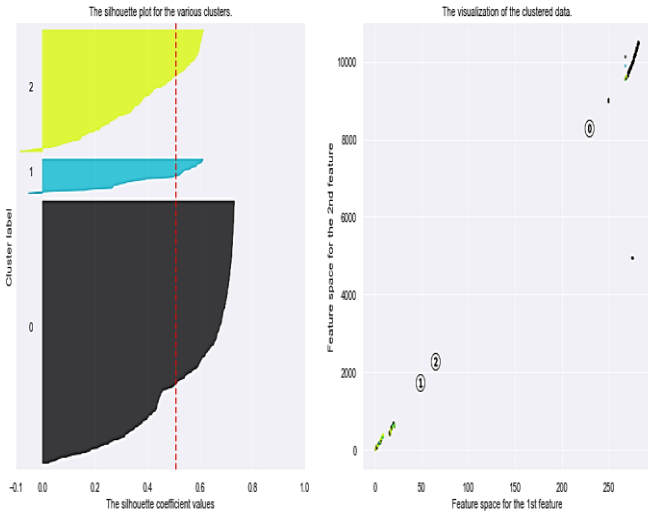

Fig. 6. Silhouette coefficient plot with number of clusters = 3

(Source: Simulated by research team, 2021)

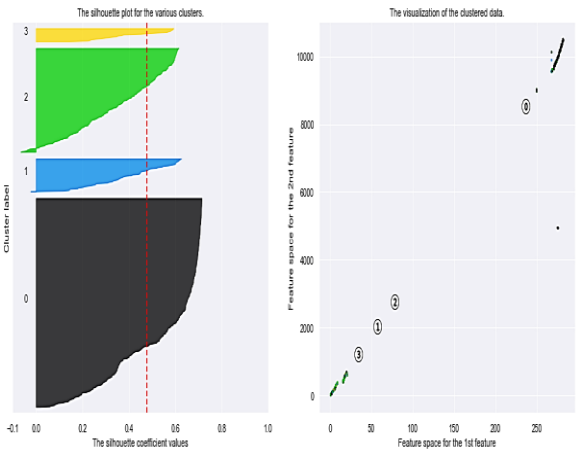

Fig. 7. Silhouette coefficient plot with number of clusters =4

(Source: Simulated by research team, 2021)

4. KEY FINDINGS

4.1. General

In this study, the geo-demographic segmentation model is applied to 582 commune-level administrative units of 30 district-level administrative units in Hanoi city.

It tried to identify the common characteristics of each group based on the results of the component matrix generated from the commune-level administrative units. Since there is no correlation between the components (clusters) formed, the properties of each component can be interpreted and determined independently of the other components (clusters). The four main components (04 clusters) obtained are dependent variables and the descriptive data listed in the research methods section used to explain these clusters are independent variables.

The study focuses on describing a number of basic characteristics of geography, population, age, and number of students to distinguish clusters in the research paper.

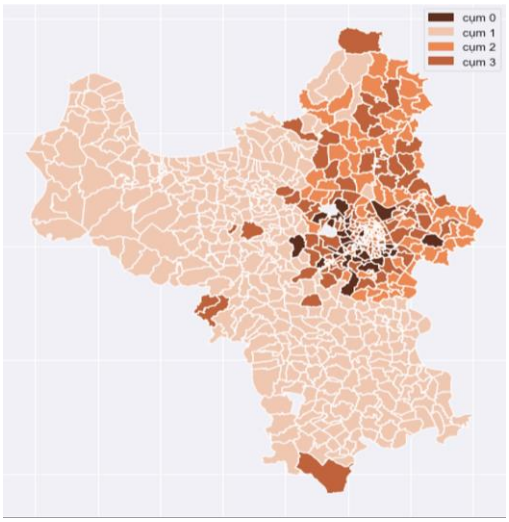

Fig. 7. Basic characteristics of clusters

(Source: Simulated by research team, 2021)

The basic characteristics of the four clusters obtained can be broadly described as follows:

Cluster 0 (30 wards and communes): This is a small area with a very high population density. Areas belonging to this cluster are concentrated in the western center of Hanoi City. The population here is mainly of working age, highly educated, with many students and people studying at the postgraduate level. Households are young families with children under 10 years old.

Cluster 1 (325 wards and communes): This is a large area with with a low concentration of population, which results in a low population density. The areas in this cluster are mainly located in the districts, accounting for about 70% of the total area, concentrated in the northeast and the south of Hanoi City. The population here is mainly of working age, but mainly engaged in farming and handicrafts, with low educational attainment.

Cluster 2 (128 wards and communes): This cluster has a high population density, but with an uneven distribution. Areas belonging to this cluster are concentrated in the northwest of Hanoi City. The population here is mainly of working age and has a high level of education. There are many households and types of businesses, but the percentage of children under 10 years old is low.

Cluster 3 (101 wards and communes): This is an area evenly distributed in the west center of Hanoi City, with an average population density, an abundant source of highly qualified labor, and many projects under development.

Table 1. Statistical table of districts and number of clusters distribution

No.	Name of District	Cluster
Districts have the characteristics of one cluster		
1	Ba Vi	Cluster 1
2	Phu Xuyen	
3	Phuc Tho	
4	Thuong Tin	
5	Ung Hoa	
6	Son Tay	
Districts have the characteristics of two clusters		
7	Hoai Duc	Cluster 1-0
8	Chuong My	Cluster 1-3
9	Dan Phuong	
10	Me Linh	
11	My Duc	
12	Quoc Oai	
13	Thach That	
14	Thanh Oai	
15	Cau Giay	Cluster 0-3
16	Ba Dinh	Cluster 2-3
17	Hoan Kiem	
18	Tay Ho	
Districts have the characteristics of three clusters		
19	Dong Anh	Cluster 1-2-3
20	Soc Son	
21	Gia Lam	Cluster 0-2-3
22	Thanh Tri	
23	Bac Tu Liem	
24	Dong Da	
25	Hai Ba Trung	
26	Hoang Mai	
27	Long Bien	
28	Nam Tu Liem	
29	Thanh Xuan	
Districts have the characteristics of four clusters		
30	Ha Dong	Cluster 0-1-2-3

(Source: Author, 2021)

4.2. Sample analysis – Ha Dong District (District has characteristics of four clusters)

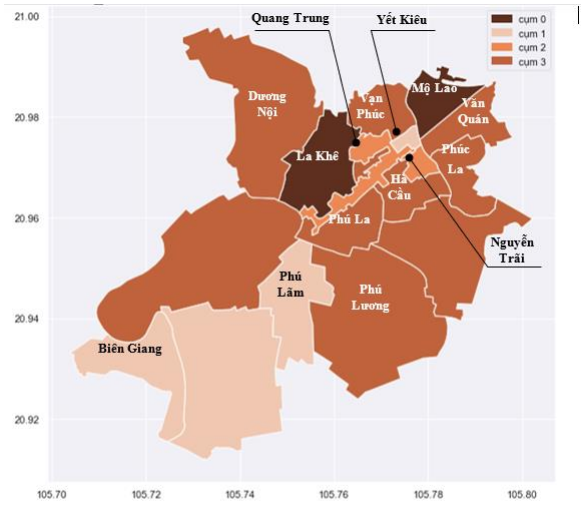

Fig. 8. Geographic segmentation map of Ha Dong district

(Source: Author, 2021)

Ha Dong District was established on May 8, 2009. The district is located between Nhue and Day rivers, located in the geometric center of Hanoi city (10km from the center of Hanoi) and is the southwest gateway of the capital (hadong.hanoi.gov.vn, nd). It borders Thanh Tri district to the east, Thanh Xuan district to the northeast, Nam Tu Liem district to the north, Hoai Duc and Quoc Oai districts to the west, Chuong My district to the southwest, and Thanh Oai district to the south. The area

of the district is 49.64km2 with a population of 397,854 people (General Statistics Office, 2019). Ha Dong district is the only district in Hanoi city where the characteristics of all four geographical clusters converge. The district consists of 17 wards:

- Cluster 0: La Khe, Mo Lao Ward
- Cluster 1: Bien Giang, Dong Mai, Phu Lam, Yet Kieu ward
- Cluster 2: Nguyen Trai, Quang Trung Ward
- Cluster 3: The other wards (**Fig. 8)**

Wards in **Cluster 0** include many agencies, units and schools are located in the city, and there are a number of major roads running through it, such as National Highway 6, Quang Trung, Le Trong Tan, and Le Van Luong.

Besides maintaining their legacy as craft villages, wards like La Khe and Mo Lao have rapidly developed with an emphasis on trade and services. Fields are superseded by high-rise buildings, adjacent villas, companies, and businesses with strong brands in the market such as Tran Anh supermarket, Saigon. These two areas also have a lot of schools of all levels from kindergarten to higher education, ensuring the educational level of employees. The infrastructure in these wards has also been significantly improved, creating favorable conditions for movement. Therefore, businesses that need a large number of customers – like transportation and logistics outlets (i.e. GHTK), shopping centers, educational centers, or small businesses – have more development opportunities in this cluster (**Cluster 0**).

Wards that are divided into **Cluster 1** are either too far from the center, and are near the suburbs such as Bien Giang, Dong Mai, and Phu Lam or are located in the middle of other wards with no major roads passing through like Yet Kieu. In addition to Yet Kieu ward, the other three wards mainly focus on agricultural production due to their large natural land area and running along the Day River. These areas are more suitable for development in terms of commerce - services such as tours of relics with spirituality because there are many people following religious beliefs (Buddhism, Christianity), or they can also develop in terms of cottage industries. Most of the businesses that have the opportunity to develop in this area are often small business. In addition, Yet Kieu ward is densely populated with a quarter of factory collective housing units such as agricultural machinery factory. Therefore, this place creates more opportunities for small businesses or businesses that need to open more branches in densely populated places such as private clinics, educational centers, bank branches.

The areas in **Cluster 2** are Quang Trung and Nguyen Trai ward. These two wards are located on the main road Quang Trung - Tran Phu - Nguyen Trai. This main road connecting Ha Dong with the center of Hanoi next to Le Van Luong - To Hieu street includes many schools of all levels, especially universities with a large number of students such as Hanoi Architectural University, Hanoi University, or some political schools such as the Academy of Politics. People living here in general have a high level of education and income because most of them run small businesses. These areas have a low land price, are not too far from the center, and are densely populated; therefore, the companies and businesses here are extremely diverse. They include trading in garments, stationery, furniture, equipment, education, and tourism. The areas around the main road are still golden lands, attracting a lot of attention from businesses thanks to its favorable geographical position for approaching customers. Therefore, these areas are suitable for most fields in trade and services.

The population density of **Cluster 3** is moderate. These wards all have many new urban areas, high-rise buildings with many promising projects, such as Duong Noi Nam Cuong urban area or Ha Cau Residence apartment complex. These are all areas with great potential, and are predicted to grow immensely in the future as the population in Ha Dong district is increasing rapidly and making housing supply unavailable. In these places have appeared a number of small businesses, mainly in real estate, small groceries or small clothing business. Therefore, there are still a lot of spaces and opportunities for other businesses. Based on the potential customers of real estate projects in these wards, they will be ideal for businesses with medium and high income customers.

4.3 Districts have the characteristics of cluster 0 and cluster 3

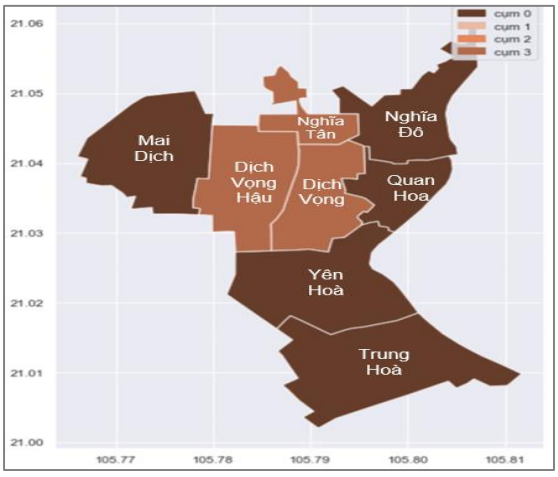

Fig. 9. Map of geographical segment-ation of Cau Giay District

(Source: Author, 2021)

Cau Giay District was established under Decree No. 74-CP dated 22/11/1996 of the Government and officially came into operation on 01/09/1997. It borders Dong Da and Ba Dinh districts to the east, Nam Tu Liem and Bac Tu Liem to the west, Thanh Xuan district to the south, and Tay Ho district to the north. The natural land area of the district is 12.32km^2 with 292,536 citizens (General

Statistics Office, 2019). The district consists of 8 wards belonging to two clusters:

- Cluster 3: Dich Vong, Mai Dich, Nghia Tan Ward.
- Cluster 0: Nghia Do, Quan Hoa, Trung Hoa, Yen Hoa, Dich Vong Hau Ward.

Dich Vong, Dich Vong Hau, and Nghia Tan Ward (**Cluster 3**) make up the central area of business establishments, administrative units, important educational institutions with a radius of 4 to 5 km around Xuan Thuy intersection. However, this area does not have too many people living there, it is mainly a place for business rental, offices, and schools. So the population density here is not too high. Because they are located in the center of the district, where there is a high daily traffic volume, these areas need more investment in the development of transport infrastructure and commercial centers serving a number of students besides small businesses (currently only Discovery Complex and Indochina Plaza Hanoi).

The area of wards in **Cluster 0** is where the population density is high, many apartment buildings are concentrated there, which are occupied mainly by people studying and working in Cau Giay district. In Trung Hoa ward, there is a hot spot on Tran Duy Hung street with many corporate offices and business establishments as well as large commercial centers. Adjacent to Yen Hoa and Dich Vong wards, Dich Vong Hau is in the process of investing in the development of infrastructure for the Information Technology district with major corporations such as Viettel and FPT, becoming key economic areas of the district, attracting more high-tech enterprises to create competitive value-added products, and minimizing adverse impacts on the regional environment. However, it is necessary to have solutions to handle a number of stagnant residential works, affecting the urban landscape and people's lives in the vicinity.

With the features mentioned above, Cau Giay has become the focus of new economic activity in Hanoi. In addition, it is still necessary to maintain the conservation of cultural heritage and traditional craft villages such as Ha Pagoda and Vong Village.

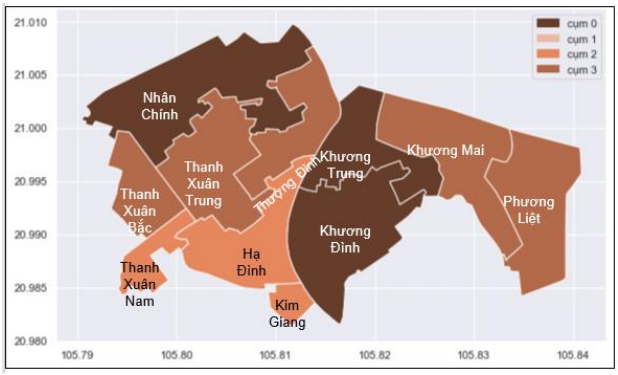

Fig. 10. Map of geographical segmentation of Thanh Xuan district

(Source: Author, 2021)

Thanh Xuan district is located in the southwest of Hanoi inner city. It borders Hai Ba Trung district to the east, Nam Tu Liem district to the west, Ha Dong district to the southwest, Hoang Mai and Thanh Tri districts to the south, and Dong Da and Cau Giay districts to the north. With an area of 9.11km² and a population of about 293,520 people (General Statistics Office, 2019), the district consists of 11 wards:

- Cluster 0: Nhan Chinh, Khuong Dinh, and Khuong Trung ward
- Cluster 2: Ha Dinh, Kim Giang, and Thanh Xuan Nam ward
- Cluster 3: Khuong Mai, Phuong Liet, Thanh Xuan Bac, Thanh Xuan Trung, and Thuong Dinh ward

With the continuous development of life, the district has attracted many large projects to develop the real estate market throughout Hanoi. The district has arterial roads: Le Van Luong, Quan Nhan, Nhan Hoa, Cu Loc, Khuat Duy Tien, Nguy Nhu Kon Tum, and Nguyen Tuan. Cluster 0 of the district is a densely populated area with many households of more than 3 people and the concentration of many large urban areas: Trung Hoa Nhan Chinh urban area, Mandarin Garden urban area, Tower SJC house - Le Van Luong, Song Da Nhan Chinh Building, Nguyen Tuan street, and The Golden Palm- Le Van Luong street. There are many working-age population here, so there are still many residential areas. To serve the people, besides the long-standing market like Khuong Dinh market, there are also toad markets, temporary markets interspersed in the residential areas, and old collective areas. The market provides relatively cheap food suitable for most working people. The district is a place with convenient transportation, reasonable rent, many small and medium-sized companies, convenience stores as well as office workers and immigrants coming to the capital city to work and study.

Cluster 2 has a small area and a small population compared to other areas in the district. For example, Ha Dinh ward has many small alleys and narrow roads. In Ha Dinh ward, it can be seen that there are many vacant land areas and four-level houses because they are agricultural land areas that have not been converted. People only build temporary houses, or a row of boarding houses for rent.

Wards in **Cluster 3** have a high level of education, with many universities such as VNU University of Science, VNU University of Social Sciences and Humanities, and Hanoi University, etc. Thuong Dinh Ward has a luxury urban area called Royal City, located at 74 Nguyen Trai Street. Khuong Mai ward has 6 major hospitals: Bach Mai Hospital, National Hospital of Tropical Diseases, National Otorhinorarynology Hospital of Vietnam, Viet Phap Hospital, National Hospital of Dermatology and Venereology, and National Geriatric Hospital. This ward is the concentration of major hospitals in the field of medical examination and treatment in Vietnam. Therefore, there are a great number of people from other provinces coming here for medical examination and treatment.

5. CONCLUSION

With the rapid development of information technology and the complex growth and expansion of economy, the existence of enterprises in any field is directly proportional to their competition in the market. To enhance their competitive advantage, businesses need to realize the importance of identifying all factors related to potential customers. Combined with geographic information management technology, administrators can effectively make decisions about where to reach and promote brands to customers, drawn from population data of each specific area.

Based on K-means clustering theory, Principal Component Analysis (PCA) with the assistance of Python programming application, this study has completed the analysis of population data of Hanoi from the Census 2019 divided into four clusters with their own characteristics and shows the common clusters in the districts based on geo-database of Hanoi City. This paper aims to develop an empirical analysis of the relevance of geo-demographic segmentation as an analytical tool for optimising business location strategy. Specifically, businesses rely on location such as retails, F&B, services… could apply this tool to explains to some extent their site selection criteria.

Although this study has achieved results, it is still limited in terms of statistics as well as time and finance.

Because of the interdisciplinary nature of geographic data and the variety in its applications, the current collected information and data can be changed to be suitable for analysis for different fields, especially marketing. Therefore, the research can be improved significantly both broader in scope (to cover all basic administrative units in the country) and deeper in the model (to analyse and simulate by using other variables that are available in the Census).

REFERENCES

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

[2]. 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.

[3]. Konu, H., Laukkanen, T. & Komppula, R. (2011), "Using ski destination choice criteria to segment Finnish ski resort customers", Tourism Management, Vol. 32, pp. 1096 – 1105.

[4]. 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. ACM.

[5]. 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, DOI: https://doi.org/10.1068/ b130176p.

[6]. Fred L., Miller, W., Glynn, M., Joy, R., Gary, B., Timothy, J., et al. (2014), "RacerGISOnline: Enhancing Learning in MarketingClasses with Web-based Business GIS", Marketing Education Review, Vol. 24, pp. 31 – 36.

[7]. General Statistics Office. (2019), Hanoi City Population Data (Census 2019).

[8]. hadong.hanoi.gov.vn. (n.d.), "Hanoi Overview", https://hadong.hanoi.gov.vn/

[9]. 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 (1), pp. 100 – 108. JSTOR 2346830.

[10]. 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.

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

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

[13]. 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.

[14]. 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.

[15]. Leung, A., Yen, B.T., and 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.

[16]. Longley, P.A. (2012), "Geodemo-graphics and the practices of geographic information science", International Journal of Geographical Information Science, Vol. 26 No. 12, pp. 2227 – 2237

[17]. Shaffer, A.C. (2015), "The geodemo-graphics in location intelligence: A study in craft brewery placement", PhD Thesis, Northern Arizona University.

[18]. Shlens, J. (2014), "A tutorial on principal component analysis", arXiv:1404.1100.

[19]. 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.

[20]. 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.