Original Article

To Discriminate General Election system in Thailand by using K-Means Clustering

Siriya Phoonokniam[1], Dr. Kanchana Kanchanasuntorn[2*], Dr. Varin Vongmanee[3]

[1,3]School of Engineering, University of the Thai Chamber of Commerce, Bangkok, Thailand

[2]Faculty of Industrial Technology and Management, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand

Email: kanchana_k@fitm.kmutnb.ac.th

DOI: 10.47750/pnr.2022.13.S06.106

Abstract

Thailand uses the ballot paper in the general election since 1933. In the present day, technology has been involved in daily basics. This study aims to explore the hypothesis of the election system in Thailand which can use the technology in the election process name as electronic voting (e-voting) or still need to use the traditional method. Before implementation, it should study in terms of the area that is ready to implement the new method and where still need to use the current one. This study takes the relevant factors to analyze with the data of each area collected from various sources. The clustering method used in this study is k-means. Then to find the acceptable k cluster the silhouette method is used. The result is 2 cluster is a perfect fit with the 11 factors that used in this study. The first cluster is Bangkok which is the capital city to be matched with the e-voting method and the second cluster is the remainder province in Thailand (76 provinces). This can be used for the next study in terms of supply chain design for e-voting and developing the ballot paper logistics.

Keywords: Election, Voting, Thailand, K-means

I. INTRODUCTION

Nowadays, technology and communication play an increasingly important role in our daily life. At the same time, the spread of COVID is a catalyst for us to adopt the technology. Thailand has had paper elections since 1933 to research whether Thailand can adopt technology in elections across the country. This study aims to find out which provinces in Thailand have still used the traditional paper-based election system or can be switched to elections that bring technology to help as electronic voting. However, each region in Thailand has differences in terms of the number of voters, number of polling stations, the distance of travel, transportation facilities, etc. This study will take all 11 relevant factors namely, the number of eligible voters, number of polling stations, the average number of academic years for a population aged 15 years and over, skills of using computers, integrity and transparency of government agencies, number of households have a computer, number of households can connect internet, number of households have a mobile phone, number of households have a landline, score of participations, and quality of road and take the data obtained in each province. All 77 provinces are grouped according to the appropriateness of the available data using an analysis method called k-means grouping. In this study, from 2 to 5 groups were tested to determine the most suitable group. To find the optimal group, another statistical tool, the silhouette method, is used to examine the similarities of the data in the same group and to find the differences in the different groups.

In this segment, the study overview has been introduced, and the next segment will represent the methodology procedure in depth. Then the result of the study together with the comparison of each cluster, the final one is the conclusion and future work of this study.

II. METHODOLOGY

A. Data sources

The data used for cluster analysis was based on data from 15 election-related factors in Thailand those have been explored using factor analysis from our previous study which was accepted for published in the Journal of Positive School Psychology then data from the respective province was collected to be able to classify how much each province has an index and how it can be classified in which electoral system. According to table 1.1, the final factor for cluster analysis was 11 cluster analysis factors. By conducting the clustering, we explore the hypothesis that provinces still use traditional election methods: ballot paper or electronic voting machines (EVM). It will lead to the development of a new election logistic model in the next step. The first factor used in this study was the average of academic years for a population aged 15 years and over [1]. The skills of using computers were computed from the average number of years of academic years for a population aged 15 years and over [2] to compare using a computer and academic level. The next factor is a report on the results of the assessment of integrity and transparency in the operation of government agencies [3] is an encouraging tool that aims to develop the Thai bureaucracy creatively. The objective is to make government agencies across the country aware of the status and problems of the organization's integrity and transparency. The results of the assessment will help government agencies can be used to improve the organization's efficiency in operating. The service can be facilitated and better respond to the people in the delivery of the overall image. The quality of the road was measured in villages where main roads are usable year-round [1], based on the percentage of villages whose roads are within village boundaries used by the majority of village residents as regular transport routes the most (only one main route) works well for the whole village with roads, meaning the roads are not potholes, easy to travel from one place to another. As for technology and communication in this study, the focus is on four factors: the number of households that have a computer, the number of households that can connect internet, the number of households that have a mobile phone, and the number of households has a landline[4]. The participation factor [1] is a fundamental factor reflecting the development of social and political participation in a democratic form of participation. There are four sub-indexes on the participant aspect: the percentage of the population who exercised the right to vote to the total eligible population of each province, the number of community organizations refers to a group of people with a management system established by members of a community to work together to benefit occupation, career development, income enhancement, housing, and environment development, or livelihood development of group members. Next, are households belonging to local groups/organizations, meaning households in villages that are members of agricultural cooperatives and households participating in village public activities by rating the score from 0-1. The next factor is the number of voters who are eligible to vote in a country aged 18 and above [5], and the final factor is the number of polling stations in each province in the 2019 House of Representatives election. [6]. Table 1.2 is the summary of the factor used in this study and table 1.3 is the source of data.

Table 1.1 Final factor for clustering analysis

From Factor analysis	For cluster analysis
Voter's age	Number of eligible voters
Number of Authority	Number of polling stations
Number of voters per polling station	
Voter's education	The average of academic years for a population aged 15 years and over
Voter's ability to use technology	Skills of using computers
Authority's ability to use technology	
The reliability of voters in the election system	Integrity and transparency of government agencies
The reliability of political parties in the election system	
The reliability of the voting system developer	
Local telecommunication systems	Number of households have a computer, Number of households can connect internet, Number of households have a mobile phone,
Available voting system	
ICT Infrastructure	

From Factor analysis	For cluster analysis
	Number of households have a landline
The involvement of the new electoral process	Score of participations
The participation of all sectors affects the electoral process at any level	
Transportation of election equipment and voters	Quality of road

Table 1.2 11 election-related factors in Thailand

Factor	Label
Number of eligible voters	C1
Number of polling stations	C2
The average of academic years for a population aged 15 years and over	C3
Skills of using computers	C4
Integrity and transparency of government agencies	C5
The number of households have a computer	C6
The number of households that can connect internet	C7
The number of households has a mobile phone	C8
The number of households that have a landline	C9
Score of participation	C10
Quality of road	C11

Table 1.3 Source of 11 election-related factors in Thailand

Factor	Source
Number of eligible voters	National Statistical Office [5]
Number of polling stations	Office of The Election Commission of Thailand [6]
The average of academic years for a population aged 15 years and over	Office of the National Economic and Social Development Council [1]
Skills of using computers	National Statistical Office [2]
Integrity and transparency of government agencies	Office of the National Anti-corruption [3]
The number of households have a computer	National Statistical Office [4]
The number of households that can connect internet	
The number of households has a mobile phone	
The number of households that have a landline	
Score of participation	Office of the National Economic and Social Development Council [1]
Quality of road	

B. K-means cluster algorithm

Clustering is a useful technique because it finalizes the partition of unknown objects into some groups. Those groups are the set of data comprised of similar points and different points of another cluster. The k-means cluster is a comparatively straightforward way to tool clustering analysis into k cluster, the smaller the distance between data, the higher the similarity. The start of the k-means method is by indicating random preliminary clusters of centroids then the data is separated into numerous groups. The Euclidean distance is used to define the similarity of data and can extract the difference in data. The new centroids are allocated to each cluster. Calculating the Euclidean distance and passing on new centroids are recurrent until members of each cluster modified. The Euclidean distance (D) is calculated by using the following equation (1):

$$D(x_a, x_b) = \sqrt{\sum_{j=i}^{n}(x_{ja} - x_{jb})^2} \qquad (1)$$

where x_ax_band are input figures and specified centroids, correspondingly, j is the data property, i is the number of properties.

C. Silhouette method

The silhouette method is considered a technique for envisioning the fitness of cluster assignment over a set of variance [7]. Contrasting the information principle methods, the silhouette penalizes having clusters which are equivalent to each other, as divergent to simply the number of clusters. Its calculation formula is as follows (2):

$$s(i) = \frac{b(i)\text{-}a(i)}{\max\{a(i),b(i)\}} \qquad (2)$$

where a(i)is the mean distance between i and the rest of its cluster, andb(i) is the lowest mean distance between i and the member of any other cluster.

The silhouette series from −1 to +1, where a nearness +1 designates that the entity is well harmonized to its cluster and ill matched to near clusters. If furthermost entities have a high rate, then the clustering conformation is fitting [7], [8].

III. Result

The data were evaluated by using IBM SPSS version 27. After obtaining the primary data, the data were standardized by z-score. To find the best figure of k clusters and silhouette method was used. This study was evaluating the result for 2-5 clusters to identify the most appropriate number of clusters.

For the first study, k = 2 clusters are explored, and the results can be shown in table 2. From table 2, we can see that the centroid value of each variable for index C1, and C6 to C9 has a high value and it has an obvious difference in centroid where the distance from cluster 1 and cluster 2 is 19.014 (table 3).

Table 2 Final Cluster Centers 2 clusters

Cluster Centers	Cluster	
	1	2
C1	6.55294	-0.08622
C2	5.27214	-0.06937
C3	3.31478	-0.04362
C4	1.35196	-0.01779
C5	1.93984	-0.02552
C6	8.01368	-0.10544
C7	7.80034	-0.10264

Cluster Centers	Cluster	
	1	2
C8	7.7393	-0.10183
C9	8.32077	-0.10948
C10	-2.26424	0.02979
C11	2.35516	-0.03099

Table 3 Distances between Final Cluster Centers 2 clusters

Cluster	1	2
1		19.014
2	19.014	

In fig. 1 show the member of the cluster which 1 has 1 province and cluster 2 has 76 provinces.

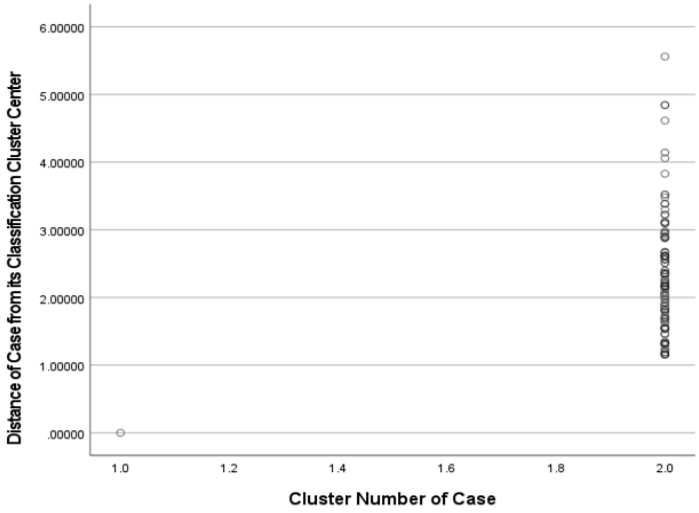

Figure 1 The member of 2 clusters

The silhouette method is calculated and the obtained results are shown in Fig 2. The value of cohesion and separation of 0.7 which is near +1 shows the high similarity of the variance in the same cluster (cohesion) and the difference between other clusters (separation).

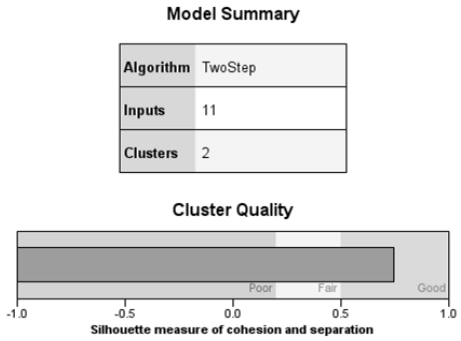

Figure 2 The silhouette value of 2 clusters

For the K= 3 clusters (table 4), there is a similar final cluster centers value in C4 between cluster number 1 (1.35196) and cluster number 3 (1.25732) compare to the distance between clusters. The distance between cluster number 1 to cluster number 2 (19.191) and number 3 (18.723) seems to be not much different (table 5).

Table 4 Final Cluster Centers 3 clusters

Cluster Centers	Cluster		
	1	2	3
C1	6.55294	-0.0192	-0.25073
C2	5.27214	0.08375	-0.44521
C3	3.31478	-0.42984	0.90439
C4	1.35196	-0.53728	1.25732
C5	1.93984	-0.19139	0.38161
C6	8.01368	-0.16847	0.04926
C7	7.80034	-0.14294	-0.0037
C8	7.7393	-0.12908	-0.03496
C9	8.32077	-0.09326	-0.1493
C10	-2.26424	0.04819	-0.01537
C11	2.35516	-0.46476	1.03372

Table 5 Distances between Final Cluster Centers 3 clusters

Cluster	1	2	3
1		19.191	18.723
2	19.191		2.827
3	18.723	2.827	

The member in these 3 clusters (fig. 3) have been divided into 1 province, 54 provinces, and 22 provinces respectively.

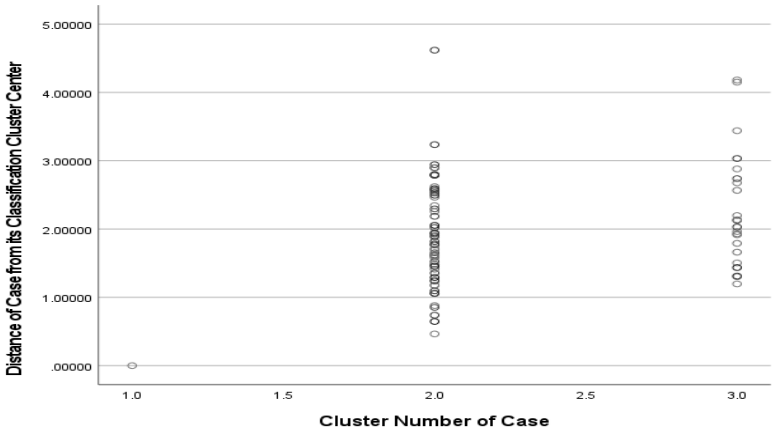

Figure 3 The member of the 3 clusters

As there is a similarity in the final cluster centers value between 3 clusters as per table 4, the silhouette value result is slightly dropped to 0.6 shown in fig 4 but it is still in good cluster quality.

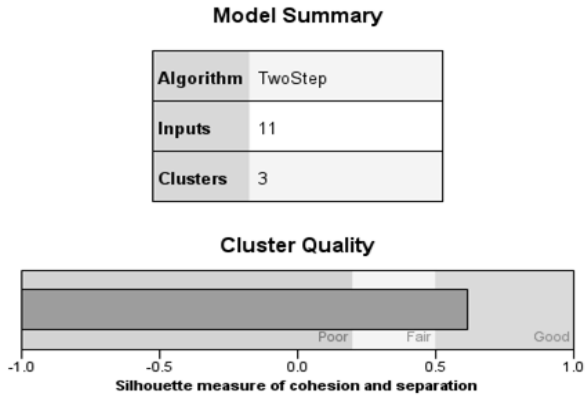

Figure 4 The silhouette value of 3 clusters

In the 4 cluster (table 6), there are 7 comparable final cluster centers in C2 between cluster number 2 (-0.45826) and cluster number 4 (-0.31957), C3 between cluster number 1 (-0.36512) and cluster number 2 (-0.25486), C4 between cluster number 3 (1.35196) and cluster number 4 (1.17846), C5 between cluster number 1 (-0.1007) and cluster number 2 (-0.29588), C7 between cluster number 1 (0.21634) and cluster number 4 (0.2724), C8 between cluster number 1 (0.27311) and cluster number 4 (0.22938), and C9 between cluster number 2 (-0.14717) and cluster number 4 (-0.15187).

Table 6 Final Cluster Centers 4 clusters

Cluster Centers	Cluster			
	1	2	3	4
C1	0.76408	-0.43882	6.55294	-0.01005
C2	1.09127	-0.45826	5.27214	-0.31957
C3	-0.36512	-0.25486	3.31478	1.24841
C4	-0.49872	-0.14166	1.35196	1.17846
C5	-0.1007	-0.29588	1.93984	1.12362
C6	0.04528	-0.26827	8.01368	0.29264
C7	0.21634	-0.32529	7.80034	0.2724
C8	0.27311	-0.33496	7.7393	0.22938
C9	0.01508	-0.14717	8.32077	
C10	-0.20875	0.2937	-2.26424	-0.62403
C11	-0.66039	-0.17117	2.35516	1.45046

The comparison to the distance between clusters. The distance of cluster number 3 to cluster number 1 (18.198), number 2 (19.650), and cluster number 4 (18.091) seem to be not much variance (table 7).

Table 7 Distances between Final Cluster Centers 4 clusters

Cluster	1	2	3	4
1		2.303	18.198	3.772
2	2.303		19.650	3.268
3	18.198			18.091
4	3.772	3.268	18.091	

In terms of members in each cluster (fig 5) is extended to 18 provinces, 46 provinces, 1 province, and 12 provinces consecutively.

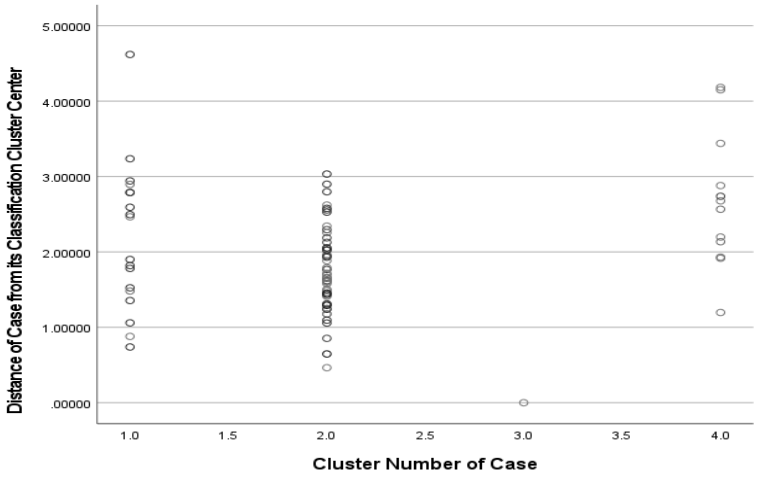

Figure 5 The member of 4 clusters

There are 7 comparable final cluster centers that affect both distances between clusters as well as the silhouette measurement directly. The clustering quality is fair, 0.4 (fig 6), which can confirm these 4 clusters don't fit this data model.

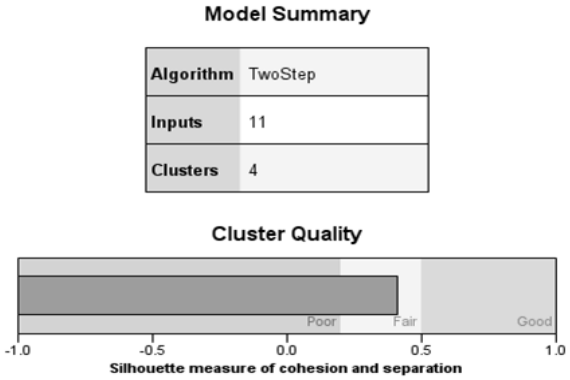

Figure 6 The silhouette value of 4 clusters

In the 5 clusters (table 8.1 and table 8.2), there are 6 similar final cluster centers in C1 between cluster number 3

(0.52409) and cluster number 5 (-0.2876), C4 between cluster number 1 (1.35196) and cluster number 4 (1.35196), C6 between cluster number 3 (-0.21967) and cluster number 5 (-0.25383), C7 between cluster number 3 (-0.28371) and cluster number 5 (-0.28093), C8 between cluster number 3

(-0.30679) and cluster number 5 (-0.28126), and C9 between cluster number 3 (-0.14888) and cluster number 5 (-0.14224).

Table 8.1 Final Cluster Centers 5 clusters

Cluster Centers	Cluster		
	1	2	3
C1	6.55294	1.16376	-0.52409
C2	5.27214	1.63197	-0.6161
C3	3.31478	-0.35788	0.59122
C4	1.35196	-0.49872	1.23629
C5	1.93984	-0.23674	0.30741
C6	8.01368	0.16663	-0.21967
C7	7.80034	0.37919	-0.28371
C8	7.7393	0.46392	-0.30679
C9	8.32077	0.16152	-0.14888
C10	-2.26424	-0.2924	0.40073
C11	2.35516	-0.67146	0.78005

Table 8.2 Final Cluster Centers 5 clusters

Cluster Centers	Cluster	
	4	5
C1	0.5248	-0.2876
C2	0.17831	-0.25113
C3	1.41625	-0.53819
C4	1.35196	-0.73006
C5	0.10335	-0.15047
C6	0.60596	-0.25383
C7	0.67646	-0.28093
C8	0.64241	-0.28126
C9	-0.15785	-0.14224
C10	-1.14836	0.16755
C11	0.57778	-0.36525

The comparison to the distance between clusters. There are 2 pairs which are the distance of cluster number 1 to cluster number 2 (17.696) and number 4 (17.365), and the distance of cluster number 1 to cluster number 3 (19.340) and number 5 (19.590) seems to be not much change (table 9).

Table 9 Distances between Final Cluster Centers 5 clusters

Cluster	1	2	3	4	5
1		17.696	19.340	17.365	19.590
2	17.696		3.997	3.452	2.704
3	19.340	3.997		2.723	2.628
4	17.365	3.452	2.723		3.767
5	19.590	2.704	2.628	3.767	

In respect of members in these 5 clusters can be seen in fig 7: 1 province, 9 provinces, 18 provinces, 8 provinces, and 41 provinces sequentially.

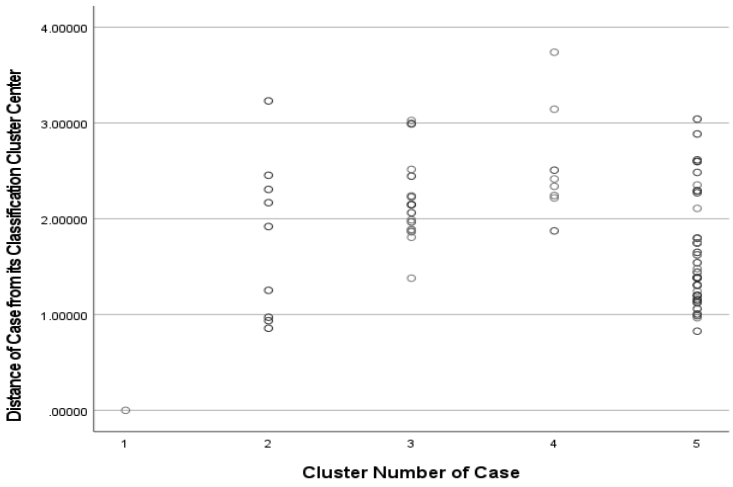

Figure 7 The member of 5 clusters

The silhouette series from −1 to +1, where a nearness +1 designates that the entity is well harmonized to its cluster and ill matched to near clusters.[7]. However, in fig. 8 the silhouette value indicates the poorest of all clustering 0.3. In the final, this clustering doesn't fit the data model either.

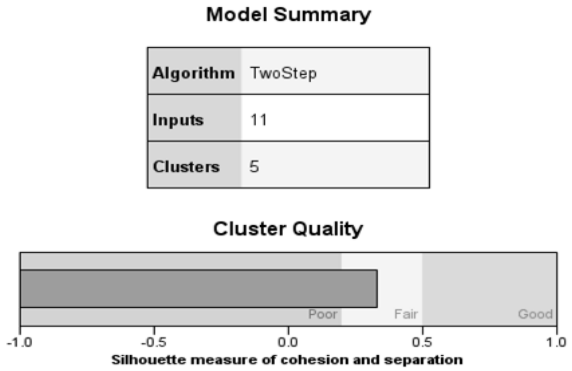

Figure 8 The silhouette value of 5 clusters

IV. CONCLUSION

From the above result, it can be concluded that k = 2 cluster is the most fitness to the data in which it has the distance between each other equal to 19.014 with a silhouette value of 0.7.

In this case, we found that the number of members in cluster 1 is only in one area which is Bangkok city. Bangkok city is the capital and most crowdy city in Thailand. The city has an estimated registered population of 5.5 million as of 2021 or equal to 8.4 percent of the country's population. Over 4.47 million people have aged over 18-years-old who have the right for voting [5], making Bangkok more significant potentially than other provinces (6.55294). In the contrast, cluster 2 contains 76 provinces with several average voters of just about 0.62 million people [1]. As a consequence of those reasons, Bangkok has the highest number of the polling station at 6,149 polling stations (5.27214) to cover the number of voters that do not overcrowd each station [6] and another cluster has an average of polling stations of only 1,134 polling stations (18% of Bangkok's polling station). On top of that road transport are the most widely used transportation mode of travel in Thailand and the index shows

an improvement in the percentage every year. The most magnificent improvement is Bangkok (2.35516), the quality of the road is high 94.70% which the main roads are available all year round to support the enormous population[1] compare to cluster 2 is 63.94%. Along with the technology and communication index; having a computer (8.01368), households in Bangkok with computers are 1.21 million households and 0.05 million in cluster 2. In the past 5 years, Thailand has increased its internet users from 52.9% in 2017 to 81.8% in 2021 (Q2) internet connection [4]. When considering internet users by region find that in the past 5 years internet users tend to increase in every region as well, and only Bangkok (7.80034) has the highest number of internet users is 92.8%, followed by the central region at 85.4%. The lowest used was the Northeastern region 75.4% [4]. In terms of households with internet connection, Bangkok has 2.8 million households that can connect. At the same time, the average of other provinces is 0.21 million households. Meanwhile, having a mobile phone (7.7393) has increased also. Bangkok is the highest number of mobile phone users, 92.2%, or an average of 2.94 million households. After that is the central region, 88.1%, and the northeastern region the lowest number of mobile phone users was 79.5% [4]. Because the average number of households is 0.24 million households. The last one is a landline (8.32077). The availability of landlines tends to decline in each province due to the replacement of mobile phones. but anyway, Bangkok was also the most active province at 52%, or 0.52 million households on average, next is the central region at 34.9%, the southern region at 10.2%, the northeastern region at 1.6%, and the lowest region is the North 1.3% [9] when considerate in term of the average households is 5.7 thousand households.

As the center of economic and modern education, the average of academic years for a population aged 15 years and over, Bangkok has made the most advances in education (3.31478) is 10.67 years [1] and 7.83 from cluster 2. Education level also affects the skills of computer users such as copying/moving files, copying/cutting/pasting text in documents, transferring files between computers and other devices such as mobile phones, etc. Bangkok (1.35196) is still a province where the population has higher abilities (51.81%) in this area than in other provinces (37%)

Plus, the integrity and transparency assessment of Bangkok's government agencies is 1 which has passed the criteria (1.93984). In the contrast, the participation factor in Bangkok (-2.26424) has a score of only 0.4 lower than cluster 2 (0.02979) has a score of 0.61. Table 10 describe in summary and compares the data in each factor.

Table 9 Compare data in each factor

	Cluster	
Factor	1*	2**
Number of eligible voters	4,479,801	626,529
Number of polling stations	6,149	1,134
The average of academic years for a population aged 15 years and over	10.67	7.83
Skills of using computers	51.81%	37.00%
Integrity and transparency of government agencies	1.00	0.20
The number of households have a computer	1,212,460	58,195
The number of households that can connect internet	2,811,840	215,453
The number of households has a mobile phone	2,949,120	242,897
The number of households that have a landline	528,994	5,754
Score of participation	0.40	0.61
Quality of road	94.60%	63.94%

Remark: * Bangkok city

** The average values per province for 76 provinces

The overall result can conclude that Bangkok has the potential to adopt the new election process which is electronic voting machines. Compare to the cluster that has 76 provinces. Most of the index has a value lower than Bangkok, especially the technology index which plays a key role to implement electronic voting. When the population lacks knowledge and is accessible to technology it might lead to difficulty to transform to new technology [10]. Therefore, the second cluster still us the traditional election method.

REFERENCES

[1] "Human Achievement Index: HAI 2020," Office of the National Economic and Social Development Council, 2021. [Online]. Available: https://www.nesdc.go.th/ewt_dl_link.php?nid=11971&filename=social. [Accessed: 29-Jan-2022].

[2] "The ITU indicators from 2020 Household Survey on the use of information and communication Technology," NATIONAL STATISTICAL OFFICE MINISTRY OF DIGITAL ECONOMY AND SOCIETY, 2021. [Online]. Available: http://www.nso.go.th/sites/2014/Pages/สำรวจ/เทคโนโลยีสารสนเทศ/รายงานตัวชี้วัดITU.aspx. [Accessed: 18-Feb-2022].

[3] "Integrity and Transparency Assessment 2021," OFFICE OF THE NATIONAL ANTI-CORRUPTION COMMISSION OF THAILAND, 2021. [Online]. Available: https://itas.nacc.go.th/home/downloaddoc/2284?fileId=216924. [Accessed: 29-Jan-2022].

[4] "The 2021 Household survey on the use of Information and Communication Technology (quarter 2)," NATIONAL STATISTICAL OFFICE MINISTRY OF DIGITAL ECONOMY AND SOCIETY, 2021. [Online]. Available: http://www.nso.go.th/sites/2014/DocLib13/Forms/AllItems.aspx?RootFolder=%2Fsites%2F2014%2FDocLib13%2Fด้านICT%2Fเทคโนโลยีในครัวเรือน%2F2564&FolderCTID=0x0120003D62909E7B39064686E0606D1231BC3A. [Accessed: 29-Jan-2022].

[5] "Number of Population from Registration by Age, Sex Region and Province: 2021," NATIONAL STATISTICAL OFFICE MINISTRY OF DIGITAL ECONOMY AND SOCIETY, 2021. [Online]. Available: http://statbbi.nso.go.th/staticreport/page/sector/th/01.aspx%0A. [Accessed: 27-Jan-2022].

[6] "Number of polling stations in the election of members of the House of Representatives," Office of The Election Commission of Thailand, 2019. [Online]. Available: https://www.ect.go.th/ect_th/more_news.php?cid=13. [Accessed: 02-Feb-2022].

[7] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," J. Comput. Appl. Math., vol. 20, pp. 53–65, Nov. 1987.

[8] O. Bocking, "Determining k in k-means clustering by exploiting attribute distributions," Untiversity of groningen, 2018.

[9] "Number of information and communication technology equipment/tools in households, classified by region and province, 2013 - 2020," National Statistical Office. [Online]. Available: http://statbbi.nso.go.th/staticreport/page/sector/th/16.aspx%0A. [Accessed: 20-Feb-2022].

[10] N. Mpekoa and D. Van Greunen, "E-voting experiences: A case of Namibia and Estonia," in 2017 IST-Africa Week Conference, IST-Africa 2017, 2017, pp. 1–8.