Application of K-Means Cluster and Spatial Statistics using Python to Analyze the Indicators of Indonesia Information Technology

Bambang Suharjo
Sekolah Tinggi Teknologi Angkatan Laut
Bumimoro, Moro Krembangan, Surabaya 60187, Telp: 031-99000581, Fax: 031-99000583
e-mail: bambang_suharjo@tnial.mil.id

Abstract

The use of computers and the internet is very important for business improvement. Analysis of its use for delineation and development plans in order to provide a better role in the business field. The problem is that there is no information technology literacy map in Indonesia that can provide an overview for national policy formulation. The research was carried out to compile a map of mastery of information technology in Indonesia by data mining from the Central Bureau of Statistics and analyzed it into 4 clusters of mastery of information technology. The presentation results in the form of a spatial statistical map showing the mastery of information technology makes it easier for executive decisions to be made, which can be followed up with education, socialization and other floating plans to increase indications of mastery of information technology to increase business success.

Keywords: *k-means cluster, spatial, python*

1. Introduction

In recent years there has been a revolution in the use of computing and communication technology, and all of them indicate that technological progress and use of information technology will continue. Very rapid developments in the field of computing, information, and communication technology have changed the way people do business and in life in general. Information technology helps companies to be able to reach more customers appropriately, introduce new products and services quickly, control marketing, and collaborate with suppliers and business partners from various regions of the world. The transformation from an industrial society to an information society and an industrial economy into a knowledge economy is the result of the impact of the use of ICTs and the Internet [1]. Thus, it becomes very important to ensure the use of ICTs and the internet so that it can be well planned for improvement so that more can be used for business improvement. In addition, during the crisis caused by the co-19 pandemic, the use of information and communication technology (ICT) increased rapidly. ICT's are very important in keeping the economy going, allowing large groups of people to work and learn from home, improving social communication online, providing entertainment that is uniquely diverse and needed. [2], [3], [4]. In addition, the world of education has also experienced fundamental changes in its learning by applying online learning to most of its learning models [5]. Thus, a massive and fast policy is needed for the development of ICT in Indonesia. For this reason, support for mapping the level of ICT literacy is needed. So, these policies can be more effective and efficient than before.

On the other hand, research on the mapping of ICT technology literacy is still limited to local research [6], as well as national scale research but does not describe the mapping of ICT mastery [7], [8]. So that this mapping can complement the shortcomings of this research and will improve decision making in the ICT field. Furthermore, research on mapping is carried out with a grouping technique as a technique in which a group of objects is put into a group called a cluster.

Furthermore, in data management, the grouping technique as a technique in which a set of objects is inserted into a group called clusters. This grouping is very suitable for obtaining data

grouping and can be used to find out quickly the data position compared to others, evaluation, and follow-up planning. Thus, clustering as a very important part of data mining is needed. K-means clustering is a very well-known clustering technique and algorithm. It is also known as the nearest neighbor search [9], [10], [11]. Furthermore, the grouping/mastery of computer and internet usage indicators can be made grouping to facilitate analysis. Next, the spatial representation of data will help facilitate analysis. Some researchers visualized their data and analysis in a spatial figure [12], [13]. This is consistent with the results of research from [14] who explained that: effective communication media that present and convey data and information to help readers understand the context of reading well and effectively present complex information is in the form of a combination of text, tables, and graphics. [15]. Specifically, presenting the results of calculations in spatial data will get the same benefits as the results of research namely getting effective communication and interest [16]. Based on the explanation above, the state of the art of this research has been analyzing data, grouping them in some clusters and visualize on the map.

The problem that should be researched is ICT indicators data not yet to be analyzed. It makes difficulties to give treatment to increase ICT indicators in Indonesia. So, it becomes important, interesting, and effective if Indonesian ICT indicator data that can be extracted from the Central Statistics Agency data needs to be analyzed by clustering and presented in a spatial map in accordance with the position of each province in Indonesia. Based on the explanation above, the aim of this research is: to analyze the cluster analysis and spatial map of ICT indicators for each Indonesian province.

2. Research method

Based on the theory of clustering [10], [11], spatial statistics [15] and python programming, this research is carried out with the following steps:

Mining data from the Central Statistics Agency in the form of data on the use/mastery of computers and provincial internet use in Indonesia in 2018, in the 2019 information technology data report. Presentation of data that has been mined in a diagram of data points so that it can know the position and spread of data.

The elbow method is used to get the best number of clusters in the k-mean cluster method.

Data cluster calculations are performed using the k-mean cluster method, using the k calculation results in step c. The presentation of the results of clustering is carried out on maps of provinces throughout Indonesia using the theory of spatial statistics.

Furthermore, the calculation results in the steps above can be described in accordance with the flow chart as follows:

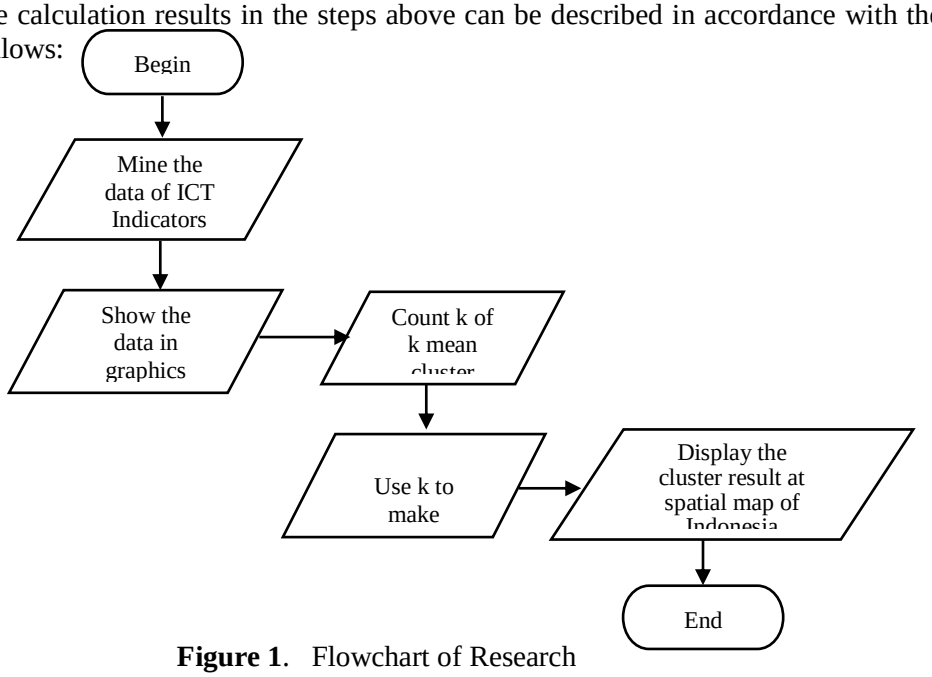

Figure 1. Flowchart of Research

4. Results and Discussion

Based on data obtained from www.bps.id [17] an internet access resume was carried out by the family in the last three months of 2018 as well as computer ownership/ownership by families in Indonesia by province, as follows.

Table 1. Data Indicators Descriptive of Information Technology Mastery

	INTERNET	USING_COMPUTERS
Count	34.00000	34.000000
Mean	63.095294	21.434412
Std	11.471216	5.856769
Min	29.500000	12.600000
25%	57.290000	17.645000
50%	63.095000	20.505000
75%	68.447500	23.220000
max	89.040000	34.990000

From descriptive data, it is known that the data consists of 2 factors: internet and computer usage. In total there are 34 data according to the number of provinces. The average internet usage 63 hours with a minimum of 29.5 hours and a maximum of 89 hours. Whereas internet usage averages 21.43 hours with a minimum of 5.85 hours and a maximum of 34.99 hours.
The K-means Cluster calculation is then performed and the calculation results are presented on a map of the provinces in Indonesia using Python Software 3.8 with the following steps.

1. Plot the Map of Provinces with Python

Plot the map of Indonesia provinces with python using the matplotlib, pandas, and numpy libraries as follows.

```
# load libraries
%matplotlib inline
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import geopandas as gpd
fi = "Indonesia/prov.shp"
map_idn = gpd.read_file(fi, encoding = "utf-8")
map_idn.head()
# now let's preview what our map looks like with no data in it
map_idn.plot()
```

Figure 2. Pseudocode to Plot the map of Indonesia provinces

The results can be presented as shown below.

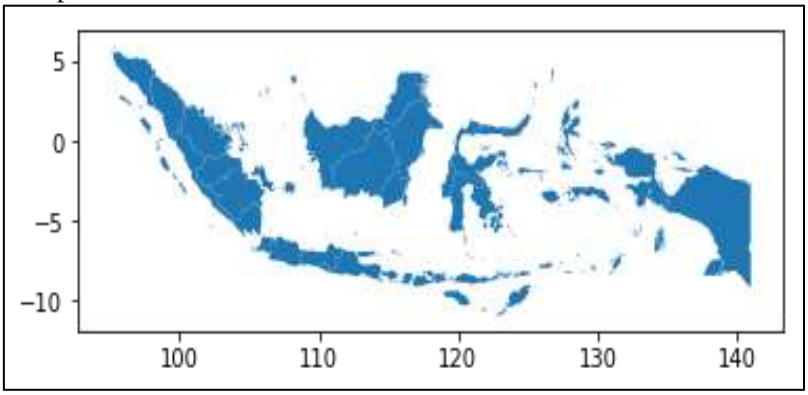

Figure 3. Plot the Map of provinces in Indonesia with shp type

In the plot, a map of the provinces in Indonesia will be used as a basis for visualizing clustering results. Furthermore, the plot is continued by clustering the data according to table 1 using k-means clustering as follows.

2. K-Means Clustering Process

First: described the position of data internet access and computer control, as follows.

```
from sklearn.cluster import KMeans
dfInternet.head()
df = dfInternet.Akses_Internet, dfInternet.Menguasai_Komputer
plt.scatter(dfInternet.Akses_Internet,
dfInternet.Menguasai_Komputer)
```

Figure 4. Pseudocode to Describe the Position of Data

Furthermore, based on the code above, the following data distribution output is generated:

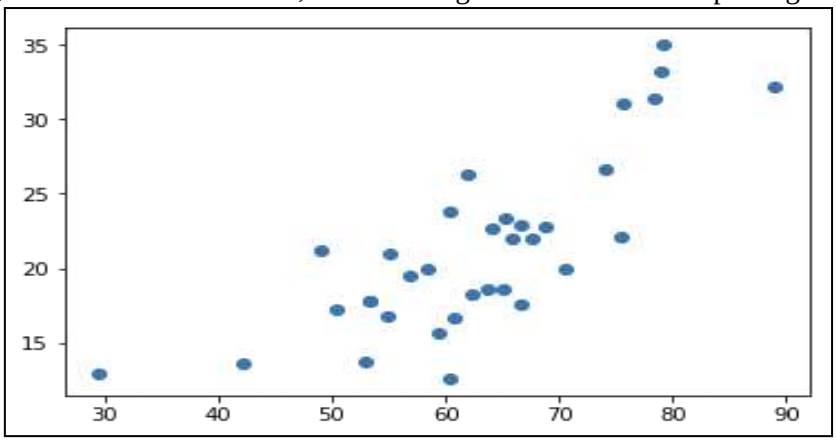

Figure 5. Plot the data indicator for Indonesian information technology

Calculate the within cluster sum of square errors (WSS) for various k values (k = 1,2,3,4,5, ...) and select k where the WSS first starts to decrease. Then plot WSS versus k plots, this is seen as a pattern changing like an elbow [16]. The steps can be explained as follows:

Calculate K-Means grouping for different K values by varying K from 1 to 10 clusters.

Calculate the total WCSS for each K.

Plot WCSS curve vs. number of K clusters

The location of bends such as the elbows in the plot is generally considered to be the most appropriate indicator of the number of clusters.

Then, the whole steps of the elbow method are written in python with numpy library using the following pseudocode:

```
import numpy as np
dfInternet.Akses_Internet[0]
a = np.arange(6)
a = a.reshape((3, 2))
X = []
for i in range (0,34):
    X.append(dfInternet.Akses_Internet[i])
    X.append(dfInternet.Menguasai_Komputer[i])
X = np.array(X)
X = x.reshape((34, 2))
from sklearn.cluster import KMeans wcss = []
```

```
for k in range(1,10):
    k_means = KMeans(n_clusters=k, init="k-means++")
    k_means.fit(X)
    wcss.append(kmeans.inertia)
plt.figure(figsize=(12,6))
plt.grid()
plt.plot(range(1,11),wcss, linewidth=2, color="red", marker
="8")
plt.xlabel("K Value")
plt.xticks(np.arange(1,11,1))
plt.ylabel("WCSS")
plt.show()
```

Figure 6. Pseudocode of the whole steps of the elbow method

In cluster analysis, the elbow method is a method used in determining the number of clusters from a set of data. The elbow method is implemented by plotting the variation in the function of the number of clusters, and selecting the k value at the curve elbow as the number of clusters to be used. The output of the optimum k determination program is as shown below.

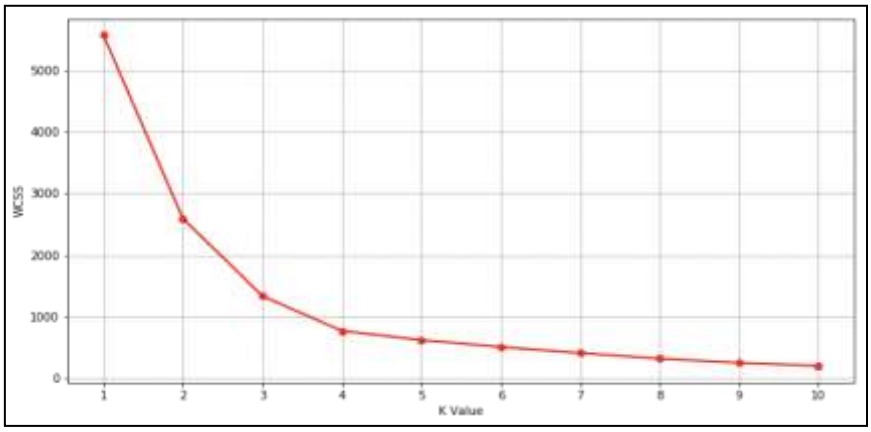

Figure 7. Graph of the Elbow Method for determining k on the k-means cluster

The elbow calculation results show that k = 2, k=3, and k=4 is the optimum value. If we use k = 2 or k = 3 then the results achieved for the analysis are not optimal because they only differentiate between 2 or 3 data, even though differences in the treatment of the results of the analysis are needed to achieve an optimal solution to the problem of using ICT in Indonesia. So, we can used clustering with k=4 or 4-mean cluster. With k = 4 a clustering plot is made as follows.

```
k_means = KMeans(n_clusters=4, init='k-means++', max_iter=300, n_init=10,
random_state=0)
prediction_y = k_means.fit_predict(X)
plt.scatter(X[:,0], X[:,1])
plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:, 1], s=10
0, c='red')
plt.show()
```

Figure 8. Pseudocode of Plot the K-Means Data Clustering

The output of the above pseudocode, can be seen on the Figure 9, as follows:

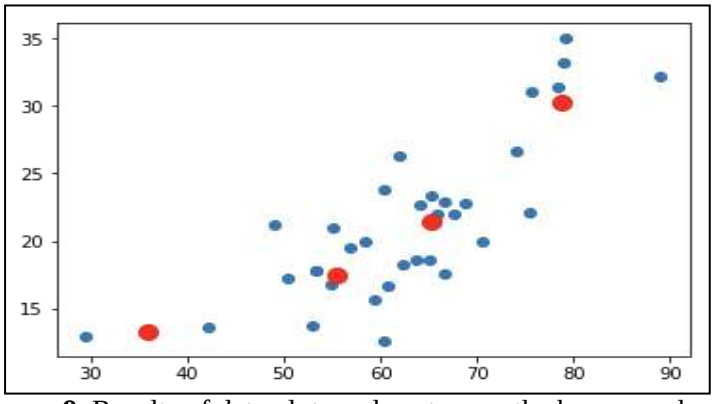

Figure 9. Results of data plots and centers on the k-means cluster

The results indicated areas in cluster 0, 1, 2, and 3. We have 7 provinces in cluster_0 12 provinces in cluster_1, 13 provinces in cluster_2 and 3 provinces in cluster_3. From the results of clustering process, we get the result as follows:

Table 2. Clustering of Indonesia indicators of ICT

Cluster	Province
0	Banten, DKI Jakarta, Yogyakarta, Bali, North Kalimantan, East Kalimantan and Riau Islands
1	Aceh, West Sumatra, North Sumatra, Bengkulu, South Sumatra, Lampung, West Kalimantan, Tenggaran Sulawesi, West Sulawesi, Central Sulawesi, Maluku, North Maluku, West Nusa Tenggara
2	Riau, Bangka Belitung, West Java, Central Java, East Java, Central Kalimantan, South Kalimantan, South Sulawesi, North Sulawesi, West Papua
3	East Nusa Tenggara and Papua

Now, we can analyse the Indonesia indicators of ICT easier than only from the table of data. Some provinces need to special attention to increase the value of ICT indicators. But, it still need to be visualized in a map to explain to Indonesia society and government to get more serious follow-up.

3. Display Clustering Result at Indonesia Map

From Figure 5 it appears that the data plot and cluster center on the four clusters shows the existence of a good cluster center distribution with existing data. Thus, further analysis can be done.

Then, labeling the data using the nearest method from the prediction point with the following steps.

```
label = k_means.labels_
print(label)
```

The result is:

[2 2 2 0 0 2 1 1 1 1 1 2 1 1 2 0 2 0 2 2 0 1 1 2 3 2 2 1 2 3 0 1 0 1]

From the result, we can see that its labelling use integers from 0 untuk 3. The same labels indicate that the provinces is in the same cluster. The order of labeling is in accordance with the order in the list of provinces that has been compiled so that the placement of provinces in the next process is not wrong. From the results of the labeling, then displayed on a map of Indonesia, as follows.

The output of the above program is:

Figure 10. Plot the spatial map of Indonesian information technology literacy indicators

Based on the results of k means cluster analysis and spatial descriptions, a description of the research findings can be presented in the table below:
From Figure 10, it can be understood that cluster 1, namely Banten, DKI Jakarta, Yogyakarta, Bali, North Kalimantan, East Kalimantan and Riau Islands is the area with the highest use of ICT so that the policies developed in that area are of a wider utilization. The second cluster namely Aceh, West Sumatra, North Sumatra, Bengkulu, South Sumatra, Lampung, West Kalimantan, Tenggaran Sulawesi, West Sulawesi, Central Sulawesi, Maluku, North Maluku, West Nusa Tenggara. Here the use of ICT is at a lower level so that it needs a higher intention to support its use. The third cluster consists of Riau, Bangka Belitung, West Java, Central Java, East Java, Central Kalimantan, South Kalimantan, South Sulawesi, North Sulawesi, West Papua. In these areas, strong support from the government and educated society is needed to optimize the use of ICT in various fields. While the fourth cluster consists of: East Nusa Tenggara and Papua. In these areas, as the most backward area in the use of ICT, it needs the most intensive education and very strong support from the government and educated society in its use. The results of this study indicate a clear mapping of information technology literacy in Indonesia from all provinces. This is the advantage of clustering and mapping in maps compared to the previous studies that have been done [7], and [8]. So that the results of this study can complement the next policy analysis.

5. Conclusions and Recommendations

Based on the results of cluster analysis, 4 clusters can be arranged optimally. Furthermore, in the mapping it appears that in the first cluster, consisting of Banten, DKI Jakarta, Yogyakarta, Bali, North Kalimantan, East Kalimantan and Riau Islands, while in the second cluster namely Aceh, West Sumatra, North Sumatra, Bengkulu, South Sumatra, Lampung, West Kalimantan, Tenggaran Sulawesi, West Sulawesi, Central Sulawesi, Maluku, North Maluku, West Nusa Tenggara. The third cluster consists of Riau, Bangka Belitung, West Java, Central Java, East Java, Central Kalimantan, South Kalimantan, South Sulawesi, North Sulawesi, West Papua. While the fourth cluster consists of: East Nusa Tenggara and Papua. Starting from the first cluster until fourth cluster, it needs the support of the government and educated society to increase the use in many areas depended on the level of cluster. So, using the k-means cluster and and presented on a statistical spatial map makes it easy to be used as an executive decision, which can be followed up with education, outreach and other floating plans in order to increase the indication of mastery of information technology [18].

References

[1] A. Shaqiri, "Impact of Information Technology in Businesses," *Acad. J. Business, Adm. Law Soc. Sci.*, vol. 1, no. 1, pp. 73–79, 2015.

[2] J. H. L. Chan and C. C. Ma, "Public Health in the Context of Environment and Housing," *Prim. Care Revisit.*, no. April, pp. 295–310, 2020, doi: 10.1007/978-981-15-2521-6_18.

[3] O. Király *et al.*, "Preventing problematic internet use during the COVID-19 pandemic: Consensus guidance," *Compr. Psychiatry*, vol. 100, pp. 1–4, 2020, doi: 10.1016/j.comppsych.2020.152180

[4] Budiman, Yusrizal, and J. Damanik, "Akses dan Penggunaan Teknologi Informasi dan Komunikasi Pada Rumah Tangga dan Individu, Access and Usage of Information and Communication Technology by Households and Individuals," vol. 15, no. 1, pp. 1–16, 2014.

[5] Y. Pujilestari, "Dampak Positif Pembelajaran Online Dalam Sistem Pendidikan Indonesia Pasca Pandemi Covid-19," *Adalah*, vol. 4, no. 1, pp. 49–56, 2020, [Online]. Available: http://journal.uinjkt.ac.id/index.php/adalah/article/ view/15394/7199.

[6] B. Saleh, "Information and Communication Technology (ICT) Literacy of Community in Mamminasata Region," *J. Pekommas*, vol. 18, no. 3, pp. 151–160, 2015.

[7] N. Kurnia and S. I. Astuti, "Peta Gerakan Literasi Digital Di Indonesia: Studi Tentang Pelaku, Ragam Kegiatan, Kelompok Sasaran Dan Mitra Yang Dilakukan Oleh Japelidi," *Informasi*, vol. 47, no. 2, p. 149, 2017, doi: 10.21831/informasi.v47i2.16079.

[8] P. Limilia and N. Aristi, "Literasi Media dan Digital di Indonesia: Sebuah Tinjauan Sistematis," *J. Komun.*, vol. 8, no. 2, pp. 205–222, 2019, doi: 10.33508/jk.v8i2.2199.

[9] B. Narang, P. Verma, and P. Kochar, "Application based, advantageous K-means Clustering Algorithm in Data Mining: A Review," *Int. J. Latest Trends Eng. Technol.*, vol. 7, no. 2, pp. 121–126, 2016.

[10] M. W. Talakua, Z. A. Leleury, and A. W. Talluta, "Acluster Analysis by Using K-Means Method for Grouping of District/City in Maluku Province Industrial Based on Indicators of Maluku Development Index In 2014," *Barekeng J. Ilmu Mat. dan Terap.*, vol. 11, no. 2, pp. 119–128, 2017.

[11] S. Shukla, "A Review ON K-means DATA Clustering APPROACH," *Int. J. Inf. Comput. Technol.*, vol. 4, no. 17, pp. 1847–1860, 2014, [Online]. Available: http://www.irphouse.com.

[12] S. Wahyuni and Y. Aryo Jatmiko, "Pengelompokan Kabupaten/Kota di Pulau Jawa Berdasarkan Faktor-Faktor Kemiskinan dengan Pendekatan Average Linkage Hierarchical Clustering," *J. Apl. Stat. KOMPUTASI Stat. Vol.*, vol. 10, no. 1, pp. 1–8, 2018.

[13] J. In and S. Lee, "Statistical data presentation," *Korean J. Anesthesiol.*, vol. 70, no. 3, pp. 267–276, 2017, doi: 10.4097/kjae.2017.70.3.267.

[14] V. Loonis and M.-P. de Bellefon, "Handbook of Spatial Analysis: Theory and Application with R," *Eurostat, INSEE*, no. October, p. 394, 2018, [Online]. Available: https://www.insee.fr/en/information/3635545.

[15] C. Yuan and H. Yang, "Research on K-Value Selection Method of K-Means Clustering Algorithm," *J*, vol. 2, no. 2, pp. 226–235, 2019, doi: 10.3390/j2020016.

[16] Ni Putu, E. Merliana, and A. J. Santoso, "Analisa Penentuan Jumlah Cluster Terbaik pada Metode K-Means," in *Prosiding Seminar Nasional Multi Disiplin Ilmu & Call for Papers Unisbank (SENDI_U)*, 2015, pp. 978–979.

[17] Badan Pusat Statistik (2019). Statistik Telekomunikasi Indonesia Tahun 2018.

[18] Z. S. Al-Lamki, "The Influence of Culture on the Successful Implementation of ICT Projects in Omani E-government An Explanatory Approach Using a Multiple Case Studies Strategy from an Information Systems Perspective," Trinity College Dublin, 2018.