SPATIAL CLUSTERING AND ANALYSIS ON HEPATITIS C VIRUS INFECTIONS IN EGYPT

Rania Fathi, Ammar Mohammed, and Hesham Hefny

Department of Computer Science, Institute of Studies and Statistics Researches (ISSR), Cairo University, Egypt

ABSTRACT

Lots of studies worldwide have been carried out to check out the prevalence of Hepatitis C Virus (HCV) in human populations. Spatial data analysis and clustering detection is a vital process in HCV monitoring to discover the area of high risk and to help involved decision makers to draw hypotheses about the cause of disease. Egypt is declared as one of the countries having the highest prevalence rate of HCV worldwide. The anomaly of the HCV infection's distribution in Egypt allowed several researches to identify the reasons that contributed to such widespread of HCV in this country. One way that can help in identification of areas with highest diseases is to give a detailed knowledge about the geographical distribution of HCV in Egypt. To achieve that goal, Data mining analytical tools integrated with GIS can help to visualize the distribution. Thus, the main propose of this paper is to present a spatial distribution of HCV in Egypt using case data obtained from the Egyptian health institute National Hepatology Tropical Medicine Research Institute (NHTMR). The visualization of the spatial analysis distribution by means of GIS allows us to investigate statistical results that are easily interpreted by non-experts.

KEYWORDS:

Data mining, clustering, k-means, spatial clustering, Geographic Information System (GIS) & Hepatitis C Virus (HCV)

1. INTRODUCTION

Data Mining is one of the most vital and motivating area of research with the objective of finding valuable information from huge data sets. Data mining is defined in [1], as the process of exploration and analysis by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules. Recently, Data Mining is becoming popular in health-care field because of the urgent needs of useful analytical methodologies for discovering unknown and meaningful information in health data.

Generally, Data Mining contributes lots of benefits in the field of health-care. For example, it can be used to detect the availability of medical solutions to the patients at lower cost, to detect the reasons of diseases and to determine medical treatment methods. Additionally, it can help health-care decision makers for planning adequate policies [2]. Moreover, it can group the patients having the same type of diseases or health problems so that health-care organizations produce adequate treatments. Usually, health-care Data is difficult to analyse as it contains a lot of details regarding patients, hospitals, medical reports, cost of treatments etc. Therefore, powerful tools and methods for analysing and extracting important information from this data are beneficial. Majority of data mining methods and techniques are grouped into supervised and unsupervised learning tasks. The first task aims at developing a model used to predict a feature data. Whereas the second task aims at model the distribution of the data to explain hidden pattern in the data.One of health-care diseases that worths investigation using data mining techniques is Hepatitis C. The latter is an infectious liver disease caused by Hepatitis C Virus (HCV). The hepatitis C is among many viruses which can invade the human liver causing inflammation and may damage to it. Since its discovery, lots of studies worldwide were carried out to check out the prevalence of HCV in human populations. According to [3],

nearly 71 million people have chronic hepatitis C worldwide with Egypt having the highest prevalence rate of HCV worldwide. The anomaly of the HCV infection's distribution in Egypt allowed several researches to identify the reasons that contributed to such widespread of HCV in this country. This anomaly also forces Egyptian health-care authorities to execute an overall screening and treatment program all over the country. One way that can help in identification of areas with highest diseases is to give a detailed knowledge about the geographical distribution of HCV in Egypt or the so- called the spatial distribution of the Hepatitis. According to [4], a spatial disease cluster may be generally defined as an area with an unusually elevated disease incidence rate. Spatial clustering can be done successfully by the integration of Clustering data mining technique and GIS. On one hand Clustering helps us to partition the data into different groups so that objects in the same group are similar and dissimilar to the objects in different clusters [5]. On the other hand, GIS supports the decision maker with effective set of tools to handle and analyse spatial information. Clustering concept is vital for spatial data visualization. The use of GIS as a visual tool allows us to investigate statistical results that are easily interpreted by non-experts. An important characteristic of spatial clustering by means of the integration of Clustering and GIS to that it allows us to identify high dimensional spatial clusters, which contains both spatial and non-spatial dimensions.

To this end, the main focus of this paper is to cluster the spatial distribution of HCV infection in Egypt using K-means clustering technique. Additionally, group analysis of the data is presented on the map.

The paper is organized as follows: Related works are shown in section 2. Background on data mining and GIS are presented in section 3. The description of the HCV data is given in section 3. Experiments and results are discussed in section 4. Finally section 6 concludes the paper.

2. RELATED WORKS

Applying data mining in healthcare is the hardest among all industries as medical data is massive, complex, inconsistent, and has no fixed rules as continuous changes and updates happens within the medicine science. This challenge always grasps many researchers in both computer science and medical fields to have the experience with data mining on disease diagnosis Differently to the experimental work presented in this study, several researchers applied data mining on other diseases or on Hepatitis C but by using other techniques. For example, In their Hepatitis C case study, the authors in [6] handled the high sensitivity of basic k-means algorithm and its flaw dealing with massive data, they suggested an Improved Genetic K-means algorithm (IGK) which selects initial clustering center purposefully using genetic algorithms (GAs). Stephane and co. [7] used Clustering and Source Attribution approaches for estimating the transmission risk factors using large HIV datasets and they found that clustering methods have higher error rates and lower sensitivity than Source Attribution method for identifying transmission risk factors but neither had provided good accuracy with identifying transmission risk ratios and the population genetic transmission modelling may be required for filling this gab. Messan Komi and co. [8] applied different five data mining techniques:

GMM (Gaussian Mixture Modelling), SVM, Logistic regression, ELM (Extreme Learning Machine),and ANN (Artificial Neural Network) to diagnose Diabetes and their experiments showed that ANN gives the highest accuracy than other techniques. Umair and co. [9] used different classification algorithms, namely Decision Tree, Neural Network and Naïve Bayes to find out heart disease patterns within 597 patients for 14 attributes. [10] Maryam and co. applied

different data mining techniques to diagnose Chronic Kidney Disease for 400 patients and 24 attributes which are shown to be highly correlated to each other. Mahmoud and Owen [11] applied k-means clustering with the help of "Elbow method" to identify the optimal clustering and also used regression for an attempt to predict elderly patient outcomes.

Regarding Hepatitis C, Huda and others [12] investigated the higher risk factors of HCV by using Binary Logistic regression technique. Also, Yousra and co. [13] had used a multivariate linear regression model to estimate the temporal trend in HCV prevalence in Egypt with data sources from PubMed and Embase databases. Multivariate linear regression technique helped them to increase the statistical power of the time trend analysis. For detecting HCV infections, Ahmed and others [14] have used the three classification techniques: Decision Trees (DT), Naïve Bayes, and Neural Networks (NN) and they found that the accuracy of the classification techniques depends on the size of the dataset, with Decision Trees technique having the highest percentage of correct predictions of patients who are likely to get HCV. Elrazek and others [15] had used both Decision Tree and Naïve Bayesian methods algorithms for prediction the vertical transmission (from pregnant mothers to new-borns) of HCV in Egypt. The analysis resulted in that the viral load, liver status, invasive techniques, HIV co infection, and immune response are among the most important predictors and they suggested further studies to determine the preventable high risk factors. As HCV spreads all over the world with its six genotypes, a recent study had been done by Ha Yeong and Tae Seon [16] comparing between the HCV genotypes using Apriori Algorithm and Decision Tree algorithm. Both algorithms revealed the distinct HCV genotypes characteristics, with genotype 1 showing very distinctive feature and as expected. From the results, this work assumed that Serine and Phenylalanine may play the key role which makes genotype 1 dominant in Korea and around the world. With focusing in 1a and 1b genotypes as being the most HCV very common infections in Korea, this study did not find the certain correlation between them and suggest that a different form of study is needed to unleash the certain relation between 1a and 1b. Somaya and co. [17] made a study for discovering the best data mining model for the predicting the advanced liver fibrosis in chronic Hepatitis C in Egypt, Their work depended on the four models: Alternating Decision Tree (ADT), Genetic Algorithm (GA), Particle Swarm Optimization (PSO), and Multilinear regression (Mreg).To refine the results, they implied feature selection and the parameters coefficients were optimized by GA, PSO, and Mreg algorithms giving different values for each one, hitherto, different accuracies. Cynthia and co. [18] introduced Single- Molecule Real-Time (SMRT) sequencing to generate full-length HCV envelope sequences at the single-molecule level, providing a data set with large sequencing depth for the characterization of intra host viral dynamics of HCV. Nazmun and Ferdous in their research [19] tried to predict HCV in the earlier stages using different Decision Tree techniques J48, LMT, Random Forest, Random tree, REPTree, Decision Stump, and Hoeffding Tree with attributes: total bilirubin, direct bilirubin, age, gender, total proteins, albumin, and globulin ratio and they found that Decision Stump had more accurate performance in HCV prediction.

From these mentioned related studies, although they applied data mining to diagnose HCV. They However, did not take the GIS into consideration.

On the other hand, there are works implied GIS in their HCV diagnosis like the study of Thomas and co. [20], it presented a new analytical model combing GIS, five-step spatial epidemiological approach, and logistic regressions to spatially identify the hotspot clusters of HCV reported infections and their characteristics in Massachusetts through the years from 2002 to 2013 displayed by dot density and choropleth maps. In a similar study, [21] had implied GIS with GWPR (Geographically Weighted Poisson Regression) model to display the spatial heterogeneity of HCV disease risk in the southern of Netherlands and to quantify the impact of socio-economic determinants on its prevalence. Also Carrie T. [22] used GIS to detect how exposure to pesticides and the age at which the exposure occurred play a vital role in increasing Breast Cancer risk.

Another related work is done by ESRI itself, ESRI [23] declared that it had developed GIS interactive map to map Zika virus infections within US. Ece [24], had discussing clustering with GIS by two different methods to classify Turkish District. Asad and others [25] processed HCV data gathered from Ghadezai Tehsil in Pakistan in GIS platform to do a spatial analysis of the infections. The maps showed HCV hotspots, the virus distribution in terms of age, gender, and marital status. Diego F. [26] and others made spatial clustering analysis on Hepatitic C prevalence in Egypt. The main point in their work is to highlight the correlation between PAT (Parenteral Antischistosomal Therapy) exposure and the HCV spread infection in Egypt for data collected 2008 with the help of Pearson correlation coefficient.

3. BACKGROUND

In this section a brief description of Data mining, clustering by k-means, GIS, and ArcGIS is introduced.

3.1 DATA MINING AND CLUSTERING

Data mining also known as a knowledge discovery is a computer-assisted process which works on a large data set scanning and analysing it to discover the hidden patterns and correlations within data. Data mining is one of the primary foundations of Machine Learning. Clustering is the most powerful and popular unsupervised learning used to reveal the natural groupings (clusters) within a large dataset based only the features (or attributes) of data points and their relationships without a pre knowledge of the clusters number. The greater the similarity within the cluster and the greater the difference between clusters, the more efficient clustering is [5].

K-means is the most widely used clustering algorithm based on partitioning. Clustering by k-means is done when the total within-cluster sum of square which is the sum of Euclidean distances between each data point and its nearest cluster centre is minimized through an iterative procedure i.e. k-means is an iterative algorithm. Hence, the main criterion used in K-means algorithm is the squared error function:

$$SSE = \sum_{j=1}^{K} \sum_{i=1}^{n} ||x_i^j - c_j||^2$$

where SSE is the Sum of Squared Error function and $||x_i^j - c_j||^2$ is the square of the distance between the data point x_i^j within the cluster c_j and its centroid.

Basic k-means algorithm steps are:

- ➢ K points are randomly selected as initial centroids of k clusters
- ➢ All data points in k clusters are assigned to the nearest (similar) centroid
- ➢ The mean distance between all the points within the cluster and the cluster's centroid is calculated and new centroids are created
- ➢ With iteration of steps 2 and 3,
- ➢ Convergence occurs as centroids do not change, and the k clusters (groups) are clearly identified.

3.2 GIS AND ARCGIS:

Geographic Information Systems (GIS) is a computational spatial analysis environment; it analyses spatial location and organizes visualized layers of information using maps [27]. Grouping of a big data set is the basic mental process to visualize it. Hence, the functionality

provided by the standard GIS packages like ArcGIS to manage and analyse data and model specific problems mainly relies on making clusters. ArcGIS provides a complete set of mapping and modelling tools. ArcGIS Maps help users to make better decisions by spotting the spatial patterns within the data set.

4. HCV DATA

The data set used in this study comprised 7587 case of HCV infections with 41 attributes shown below. The data represents HCV infections within 23 centres across 19 governorates with no available date for the rest centres. The data set is obtained from the Egyptian health institute National Hepatology Tropical Medicine Research Institute (NHTMR).

Table 1: total attributes

Centre follow up	Albumin (g/dl)
Age	ALT (IU/L)
Albumin (g/dL)	ALT (ULN)
Alcohol Intake	$ANCx10^3$
ALT Result	AST (IU/L)
ANA	AST (ULN)
ANCx	Blood Sample Storage
Ascites	Creatinine (mg/dl)
Ascites if yes	Fundus Exam
AST Result	Glucose (mg/dl)
BMI	TSH
Fib4 Calculation	Weight (Kg)
Fibrosis	W4 Rapid Virological Response
Fibrosis Stage F	HCV RNA W4
Former or on-going IV Drug User	HCV RNA W12
Fundus Exam	HCV RNA w16
Gender	HCV RNA W24
Tobacco Consumption	HCV RNA W28
Hb (g/L)	HCV RNA W36
AFP (IU/L)	Final response to cure
AFP (ULN)	

5. EXPERIMENTS

This section describes the running experiments. First, it starts by showing the data preprocessing. Then, it shows how to cluster and identify the optimal number of clusters of the data.

5.1 PREPROCESSING

Preprocessing is a necessary stage before processing data mining as the medical raw data is usually unregulated and contain missed values. Manual preprocessing is done to avoid misleading outputs. Attributes with little values are excluded and incomplete instances are also deleted. This cleaning resulted in 4638 instance and 18 attributes shown by the table 2 plus the "centre follow" up attribute:

Table 2: used attributed other than 'centre follow up' attribute and their description, type, and value.

Attributes	Description	Type	Values
Age	age of the patient	Numeric	
Albumin (g/dL)	Albumin is a protein made by the liver. It's value in the blood indicates how the liver is diseased	Numeric	
Alcohol Intake	Alcohol Intake test	Nominal	Yes, No
ALT Result	Alanine Aminotransferase plays a vital role in metabolism. If liver is inflamed, its level in blood rises	Numeric	
ANA	The Anti-Nuclear Antibody test checks any autoimmune disorders	Nominal	Negative ,Positive
ANCx	Absolute Neutrophil Count test measures the level of neutrophil granulocytes in the blood which are a type of white blood cells fighting infections.	Numeric	
Ascites	Ascites refers to abnormal accumulation of fluid in abdomen, it is a symptom of liver disease	Nominal	Yes, No
AST Result	Asperate Aminotransferase is a liver enzyme. AST test checks for liver damage.	Numeric	
BMI	Body Mass Index test measures patient's weight with respect to his/her height.	Numeric	
Fib4 Calculation	Index for Liver Fibrosis. It is a non-invasive scoring system based on several laboratory tests.	Numeric	
Former or on-going IV Drug User	Asks if patient is a former or ongoing Intravenous drug user	Nominal	Yes, No
Fundus Exam	The fundus of the eye is the interior surface of the eye opposite the lens.	Nominal	Normal, Abnormal
Gender	Male or Female	Nominal	Male, Female
Tobacco Consumption	Smoker or not	Nominal	Yes, No
Blood Sample Storage	Blood Sample Storage test	Nominal	Yes, No
HCV RNA W24	HCV RNA in treatment week no. 24	Nominal	Negative, Positive
HCV RNA W36	HCV RNA in treatment week no. 36	Nominal	Negative, Positive
Final Decision	The final result whether the patient is recovered from HCV or not.	Nominal	1:IFN/SOF/RBV, 2:SOF/RBV, 3:SOF/DAC/RBV, 4:SOF/SIM, 5:SOF/DAC.

5.2 DISCOVERING OPTIMAL PATIENT CLUSTERS

K-means algorithm requires a previous knowledge of the number of clusters. To handle this issue, the "elbow method" is used in this work to determine the best identification of the number of patients' clusters. The Elbow method looks at the total WSS (within-cluster sum of squares) as a function of the number of clusters. As K increases, total WSS drops exceedingly till a point at which adding another cluster doesn't improve much better clustering compactness. So, the best K corresponds to the angle of the graph (the elbow) at which the best modelling of data is obtained. The elbow curve is obtained by python platform as shown below in fig 1 with the result of optimal K equals three.
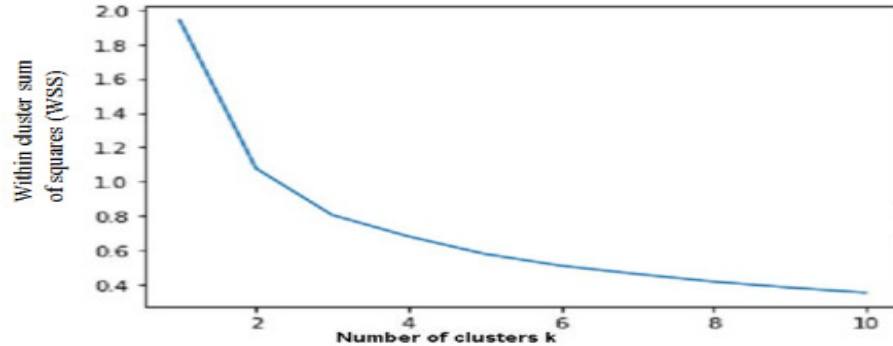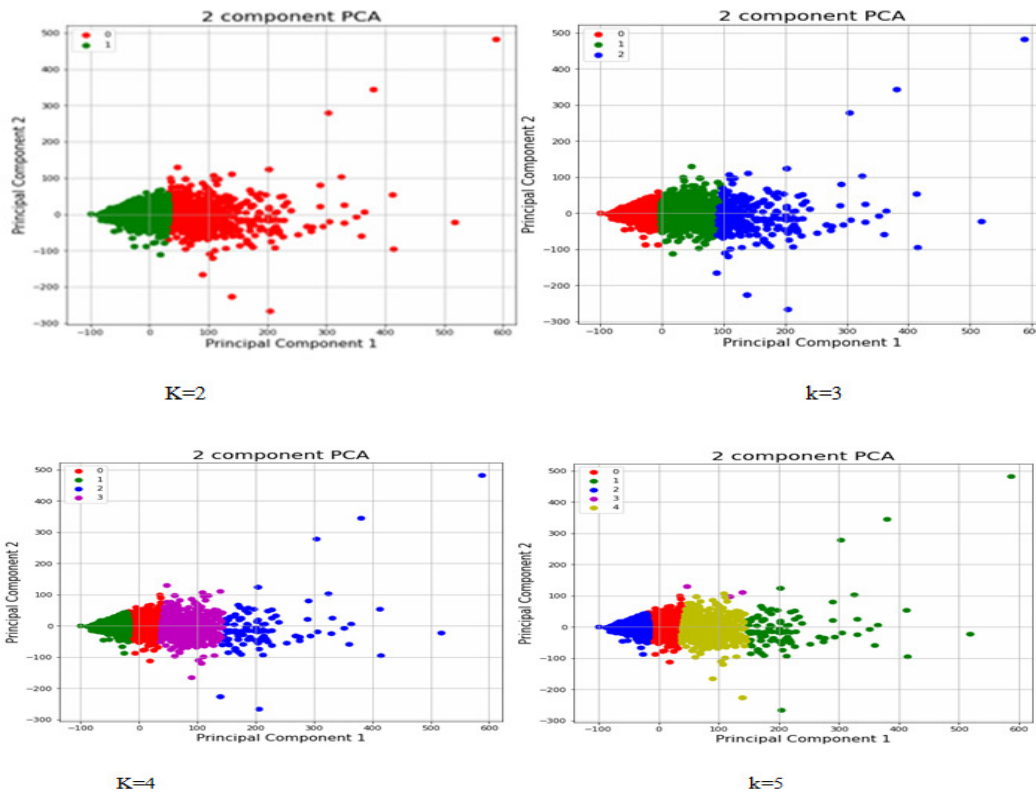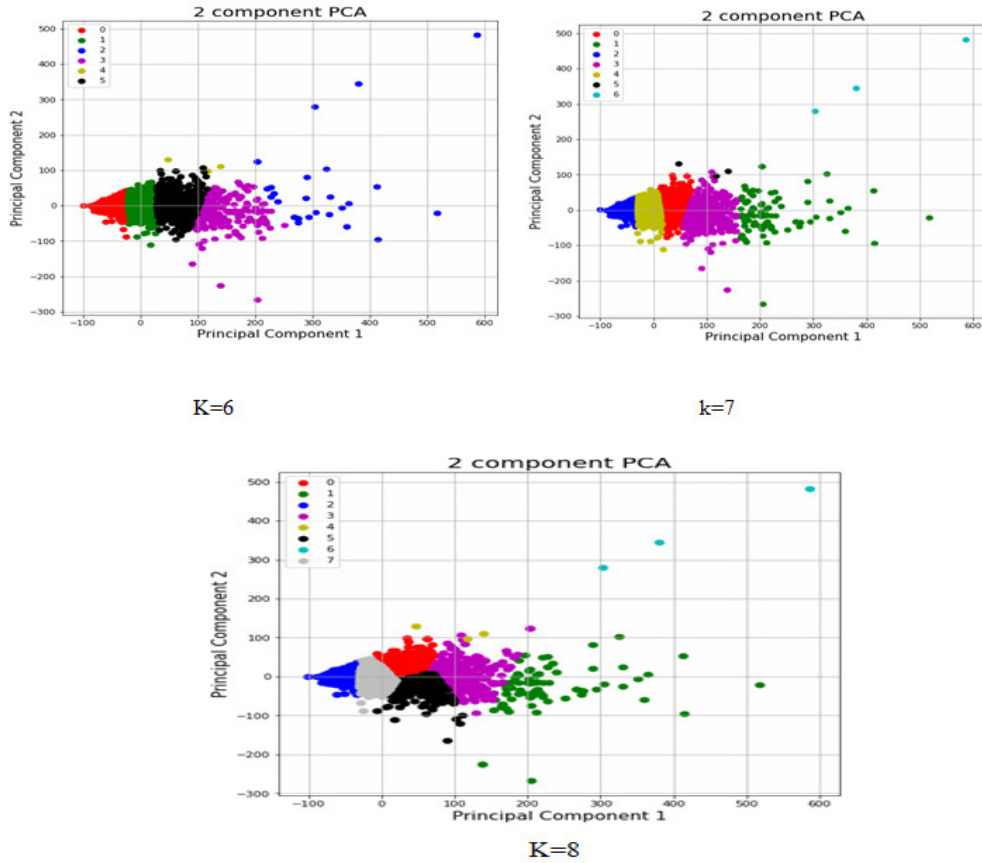

Fig 1: Elbow curve by Python

To help visualize the clusters, the figure below shows the different clustering experiments with number of clusters (K) ranging from 2 to 8 done by k-means PCA analysis within Python.

K=2

k=3

K=4

k=5

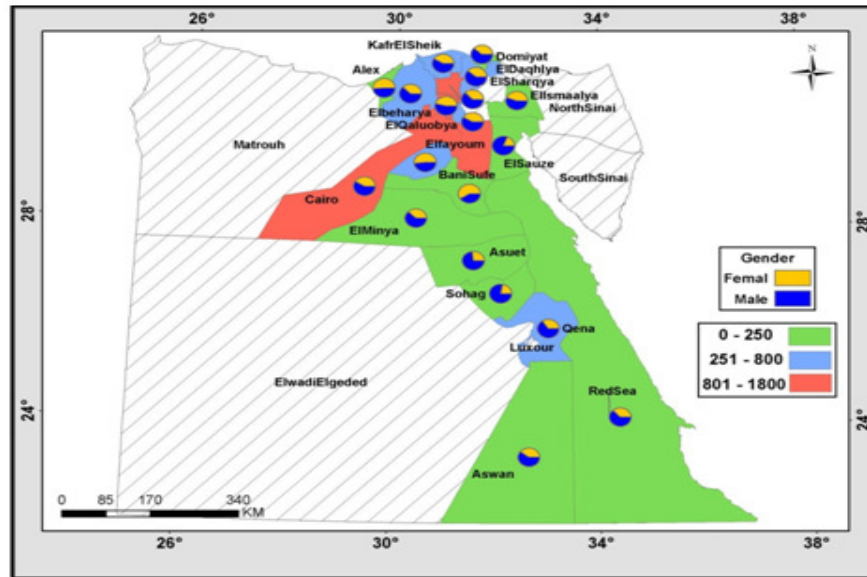K=6   k=7

K=8

6. RESULTS

After specification of the optimal number of patient clusters, we run k-means clustering by Rapid Miner program. Then, the output of the rapid miner is processed and analysed with ArcGis tool. The following maps are resulted:

Map1 shows the base map in terms of the HCV patients' count in governorates. Note that the hashed area represented no data available in these geographical regions.
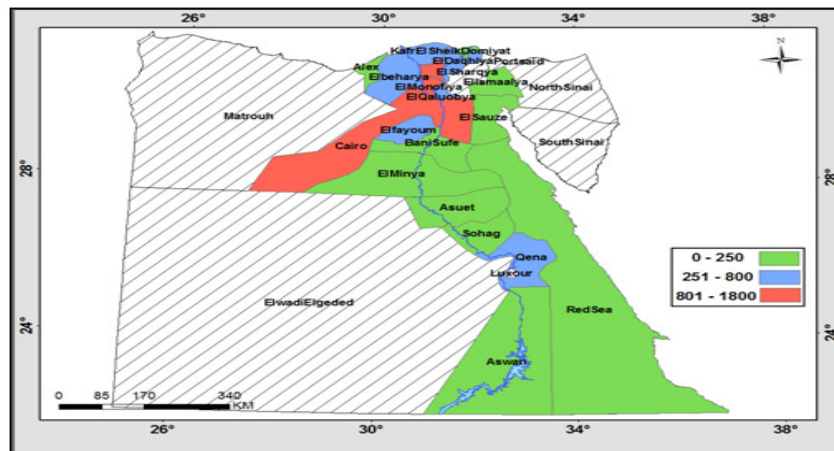

Map1: Base map by ArcGIS group analysis representing the spatial HCV distribution in Egypt

For getting this map, we initiated ArcGIS group analysis with specifying 3 clusters. Group one with green colour represents the low count governorates for number of infections less than 250 each, group two represents the intermediate infection level which is within blue regions with infections count from 251 to 800, and the third group with blue colour represents the highest spots of HCV infections in Egypt with count from 801 to 1800.

Map2 visualizes the relative sizes of the three clusters obtained by k-means across the country. This map may help in managing treatment strategy by noticing each cluster's prevalence over the state. Further analysis may be by taking the base map into account.

Map 2: distribution of k-means clusters over the ArcGis base map with numerical details shown in chart 1.

Chart 1: number of patients in each cluster for the data-covered governorates.

Map 3 shows the ratio between female and male infections in each governorate across the country with obvious higher prevalence of HCV in males than females.

Map 3: spatial distribution of HCV patients by gender

Map 4 shows 7 categories of patient's ages plotted on the geographical map:
from 11to 20 years old, from 21 to 30 years ,from 31 to 40 years old, from 41 to 50 years old, from 51 years to 60 years, from 61 to 70 years old, and finally > 70 years old.
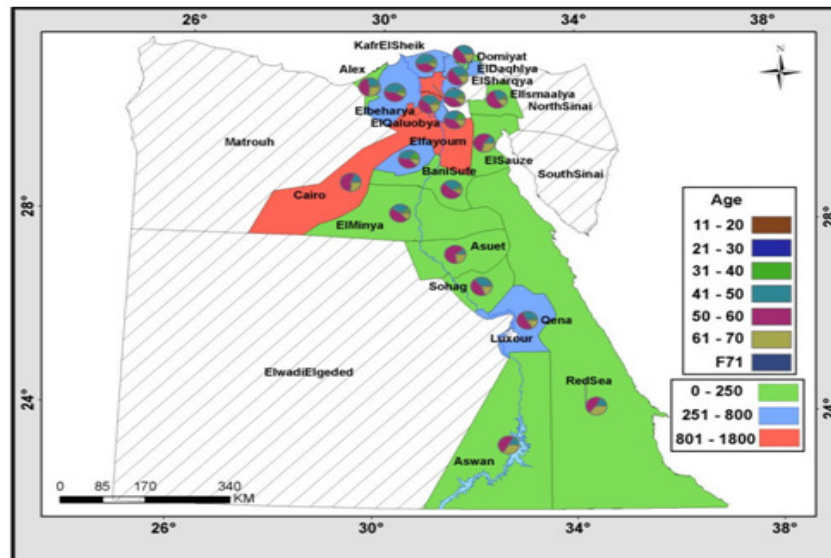

Map 4: Geographical distribution of the different age stages of the infected Egyptians across the Governments

Map 5 shows the spatial distribution treatment response (negative or positive to HCV) for males and females separately put into 4 categories:

NM: negative male with

Color green, PM: positive

Male with color yellow, NF:

Negative female with color

10

Brown, PF: positive female with color pink.
i.e. NM and NF represents male and female patients who responded to the treatment and been recovered from HCV and PM and PN represents to male and female who didn't respond to HCV treatment.
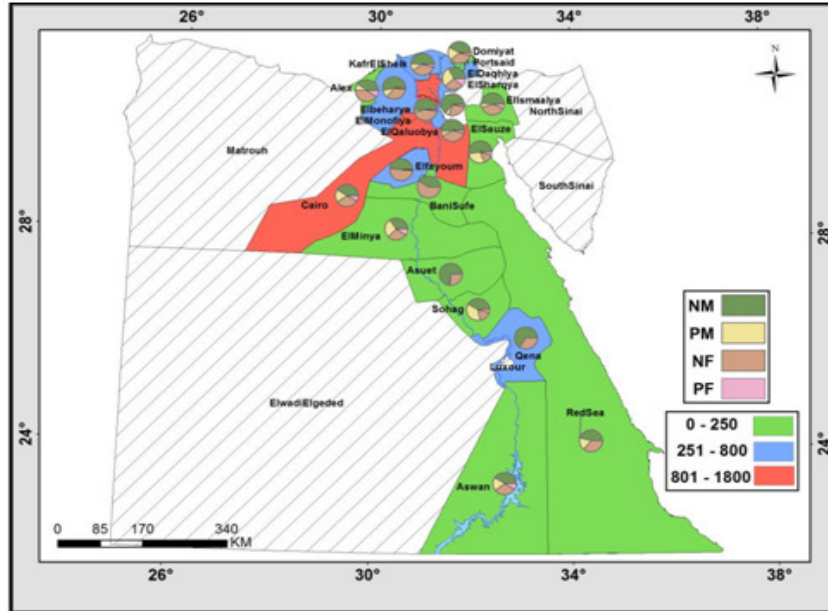
Map 5: spatial distribution of male and female final response to HCV cure program

By ArcGIS tool, another grouping analysis resulted in a spatial distribution according to recovery percentage over country shown by map 6. Red group representing the zones with lowest recovery percentage (less than 67%) over infected people there, highest percentage of recovery (more than 86%) are within green zones, and intermediate case with the blue color.
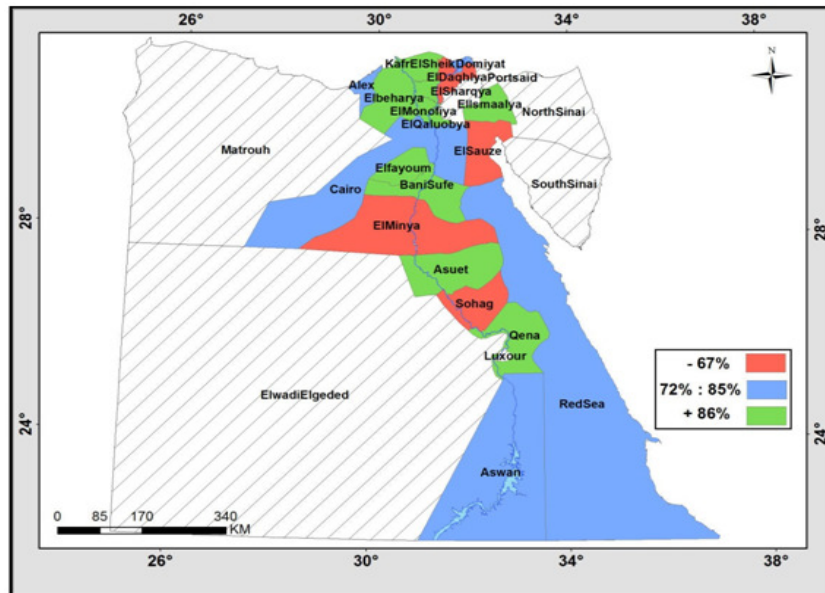
Map 6: Spatial distribution of the response degree to HCV treatment across the country

7. CONCLUSION

This paper represented spatial analysis using GIS and data mining by k-means clustering combined with the elbow method for the study of HCV infections in Egypt. The infected governorates were shown with the HCV spatial distribution by patients' gender, age, and their recovery degree to HCV cure. Our work is a trial to facilitate diagnosing HCV infections in Egypt for further actions by decision makers. The study also is a contribution to shed light on the vital importance of data visualization in healthcare field.

REFERENCES

[1] Xingquan Zhu, Ian Davidson, "Knowledge Discovery and Data Mining: Challenges and Realities", ISBN 78- 1-59904-252, Hershey, New York, 2007.

[2] Hian Chye Koh and Gerald Tan (2005), "Data Mining Applications in Health-care", Journal of healthcare information management, Vol. 19, Issue 2, Pages 64-72.

[3] World Health Organization, (2017) http://www.who.int/news-room/fact-sheets/detail/hepatitis-c.

[4] Waller LA, Hill EG, and Rudd RA (2006), "The geography of power: statistical performance of tests of clusters and clustering in heterogeneous populations statistics in Medicine", Stat Medicine Volume 25, Issue 5:853

[5] Han, J., Kamber, M., and Pei J., (2012). "Cluster Analysis: Basic Concepts and Methods" In book: Data Mining (Third Edition), Edition: Third Edition, Publisher: Morgan Kaufmann, pp.443 - 495.

[6] M. H. Marghny, Rasha M. Abd El-Aziz , and Ahmed I. Taloba, (2011) "An Effective Evolutionary Clustering Algorithm: Hepatitis C case study", International Journal of Computer Applications (0975 – 8887) Volume 34– No.6.

[7] Stéphane Le Vu, Oliver Ratmann, Valerie Delpech, Alison E. Brown, O. Noel Gillc, Anna Tostevin, Christophe Fraser ,and Erik M. Volz, (2017) "Comparison of cluster-based and source-attribution methods for estimating transmission risk using large HIV sequence databases", Epidemics Journal, 23.10.1016/j.epidem.2017.10.001, Volume 23- No 10 .

[8] Messan, Komi, Jun Li, and XXiangou Zhang, (2017) " Application of data mining methods in diabetes prediction", 2nd International Conference on Image, Vision and Computing (ICIVC), Chengdu, China, 2017.

[9] Umair Shafique, Fiaz Majeed, Haseeb Qaiser, and Irfan Ul Mustafa, (2015) "Data Mining in Healthcare for Heart Diseases", International Journal of Innovation and Applied Studies ISSN 2028-9324 Vol. 10 No. 4 Mar. 2015.

[10] Maryam Gharibdousti, Kamran Azimi, Saraswathi Hathika, and Dae H Won, (2017) "Prediction of Chronic Kidney Disease Using Data Mining Techniques", Industrial and Systems Engineering Conference, 2017.

[11] Mahmoud Elbattah and Owen Molloy, (2017) "Clustering-Aided approach for predicting patient outcomes with application to Elderly Healthcare in Ireland", The AAAI-17 Joint Workshop on Health Intelligence,WS-17-09.

[12] Huda Yasin, Tahseen A. Jilani, and Madiha Danish, (2011) "Hepatitis-C Classification using Data Mining Techniques", International Journal of Computer Applications (0975 – 8887)Volume 24– No.3.

[13] Yousra A Mohamoud, Ghina R Mumtaz, Suzanne Riome, DeWolfe Miller, and Laith J Abu-Raddad, (2013) "The Epidimiology of Hepatitis c virus in egypt: a systematic review and data synthesis", BMC Infectious Diseases Journal,2013, Volume 13:288.

[14] Ahmed A. A. Radwan, Tarek Abdel hafeez, & Heba Mamdouh, (2013) "An analysis of hepatitis c virus prediction using different data mining techniques", International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR), ISSN 2249-6831, Vol. 3, Issue 4, Oct 2013.

[15] Abd Elrazek, Mohamed Amer, Bahaa El-Hawary, Altaher Salah, Akshaya S. Bhagavathula, M. Alboraie & Samy Saab, (2016) "Prediction of HCV vertical transmission: what are factors should be optimized using data mining computational analysis", Liver International Journal, Vol.37(4), April 2016 ISSN 1478-3223.

[16] Ha Yeong Jeong and Tae Seon Yoon, (2017) "Analysis of Hepatitis C Virus using Data mining algorithm - Apriori, Decision tree", International Journal on Bioinformatics & Biosciences (IJBB) Vol.7, No.3.

[17] Somaya Hashem, Gamal Esmat, Wafaa Elakel, Shahira Habashy, Safaa Abdel Raouf, Mohamed Elhefnawi, Mohamed Eladawy & Mahmoud Elhefnawi1, (2017) " Comparison of Machine Learning Approaches for Prediction of Advanced Liver Fibrosis in Chronic Hepatitis C Patients", IEEE Journal, Volume: 15 Issue: 3.

[18] Cynthia K, Jayna Raghwani, Sylvie Koekkoek, Richard H, Jan T., Marc Van, Menno De Jong, Oliver G., Janke S., and Richard M., (2017) " Characterization of Hepatitis C Virus (HCV) Envelope Diversification from Acute to Chronic Infection within a Sexually Transmitted HCV Cluster by Using Single-Molecule, Real-Time Sequencing", Journal of Virology, Volume 91 Issue 6, 28 Feb 2017.

[19] Nazmun Nahar and Ferdous Ara, (2018) "Liver disease prediction by using different Decision Tree techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.8, No.2, March 2018. Vol.8, No.2.

[20] Thomas J. Stopka, Michael A. Goulart, David J. Meyers, Marga Hutcheson, Kerri Barton, Shauna Onofrey, Daniel Church, Ashley Donahue & Kenneth K. H., (2017) "Identifying and characterizing hepatitis C virus hotspots in Massachusetts: a spatial epidemiological approach", BMC Infectious Diseases Journal, vol 17, 2017.

[21] Boris Kauhl, Jeanne Heil, Christian J. P. A. Hoebe, Jürgen Schweikart, Thomas Krafft & Nicole H., (2015) "The Spatial Distribution of Hepatitis C Virus Infections and Associated Determinants-An Application of a Geographically Weighted Poisson Regression for Evidence-Based Screening Interventions in Hotspots", PLOS one, vol 10(9).

[22] Carrie Tayour, "A GIS-Based Model of Historical Pesticide Exposure on the Risk of Breast Cancer" article published online by American Public Health Association (APHA)142 nd Annual Meeting &Expo Nov.15-19,2014 new Orleans, LA.

[23] ESRI, (2016) "GIS Technology Proves Fundamental in Battle against Zika Virus".

[24] Ece Aksoy, (2006) "Clustering with GIS: An Attempt to Classify Turkish Data", in XXIII FIG Congress, page 8-13.

[25] Asad Ali, Muhammad Shakee, Shoaib Khalid, Tanveer Ahmed, Mubarra Noreen, and Ali Rehan Munir, (2015) "Spatial Patterns of Hepatitis C Disease in Ghadezai Tehsil District Buner, Khyber Paktunkhwa, Pakistan", Journal of Geography & Natural Disasters 5:146.

[26] Diego F., Cuadros, Adam J. Branscum, F. DeWolfe Miller, and Laith J. Abu-Raddad, (2014) "Spatial Epidemiology of Hepatitis C Virus Infection in Egypt: Analyses and Implications" Hepatology Journal, 60(4): 1150–1159, oct 2014.

[27] ESRI " https://www.esri.com/en-us/what-is-gis/overview".