

WEB GEOSPATIAL VISUALISATION FOR CLUSTERING ANALYSIS OF EPIDEMIOLOGICAL DATA

Jingyuan Zhang

A thesis submitted in fulfilment of
the requirements for the degree of
Doctor of Philosophy



College of Engineering and Science
VICTORIA UNIVERSITY
Melbourne, Australia
December 2014

Abstract

Public health is a major factor that in reducing of disease round the world. Today, most governments recognise the importance of public health surveillance in monitoring and clarifying the epidemiology of health problems. As part of public health surveillance, public health professionals utilise the results of epidemiological analysis to reform health care policy and health service plans. There are many health reports on epidemiological analysis within government departments, but the public are not authorised to access these reports because of commercial software restrictions. Although governments publish many reports of epidemiological analysis, the reports are coded in epidemiology terminology and are almost impossible for the public to fully understand.

In order to improve public awareness, there is an urgent need for government to produce a more easily understandable epidemiological analysis and to provide an open access reporting system with minimum cost. Inevitably, it poses challenges to IT professionals to develop a simple, easily understandable and freely accessible system for public use. It is not only required to identify a data analysis algorithm which can make epidemiological analysis reports easily understood but also to choose a platform which can facilitate the visualisation of epidemiological analysis reports with minimum cost. In this thesis, there were two major research objectives: the clustering analysis of epidemiological data and the geospatial visualisation of the results of the clustering analysis. SOM, FCM and k-means, the three commonly used clustering algorithms for health data analysis, were investigated. After a number of experiments, k-means has

been identified, based on Davies-Bouldin index validation, as the best clustering algorithm for epidemiological data. The geospatial visualisation requires a Geo-Mashups engine and geospatial layer customisation. Because of the capacity and many successful applications of free geospatial web services, Google Maps has been chosen as the geospatial visualisation platform for epidemiological reporting.

In summary, there are three significant contributions in this research:

- Investigation of the best algorithm for clustering analysis of epidemiological data.
- Creation of geospatial visualisation for clustering analysis of epidemiological data.
- Development of a precise, effective and intuitive web-based geospatial epidemiological data visualisation application, WebEpi.

Declaration

I, Jingyuan Zhang, declare that the PhD Thesis entitled “Web Geospatial Visualisation for Clustering Analysis of Epidemiological Data” is no more than 100,000 words in length including quotes and exclusive of tables, figures, appendices, bibliography, references and footnotes. This thesis contains no material that has been submitted previously, in whole or in part, for the award of any other academic degree or diploma. Except where otherwise indicated, this thesis is my own work.

Signature

Date

Acknowledgements

My first words of appreciation go to my supervisor, Associate Professor Hao Shi, for her full support and encouragement throughout the course of my study at Victoria University. Professor Shi is an excellent mentor. She is one of the most reliable and kindly people I have ever met. She spent a great deal of her time on my research and publications. Her guidance and advice have been the major contributors toward my PhD.

I would like to thank my co-supervisor Professor Yanchun Zhang for his support and feedback on my research study. He has been very supportive of my research. He and Professor Shi applied for a special Innovation Research Grant from the former Faculty of Engineering and Science, Victoria University for my research project and then I was offered the Faculty postgraduate scholarship to commence my PhD study. I would also like to thank the Australia Government for the Australian Postgraduate Award (APA) scholarship which supported me during the rest of my PhD studies. I would like to convey thanks to Dr Peter Wan and the Department of Health and Human Services in Tasmania, Australia for providing research data and feedback on my research results.

I wish to express my love and gratitude to my beloved parents, son and husband. My family has provided me great support, understanding and endless love throughout my study.

List of Publications and Awards

- [1] Zhang J. and Shi H. "Geo-visualization and Clustering to Support Epidemiology Surveillance Exploration" Proceedings of Digital Image Computing: Techniques and Applications (DICTA2010), 01-03 December 2010, Sydney, Australia, pp. 381-386.
- [2] Zhang J., Shi H. and Zhang Y. "Self-Organizing Map Methodology and Google Maps Services for Geographical Epidemiology Mapping", Proceedings of Digital Image Computing: Techniques and Applications (DICTA2009), 01 – 03 December 2009, Melbourne, Australia, pp. 229-235.
- [3] Shi H., Zhang J. and Zhang Y. "New WebEpi Technologies for Epidemiology Data Geo-Visualization Mashups", Proceedings of the International Conference on Modeling, Simulation and Visualization Methods, (MSV'09), 13 – 16 July 2009, Las Vegas, USA, pp. 36-41.
- [4] Zhang J., Shi H. and Zhang Y. "Geo-Mashups Automation for Web-Based Epidemiological Reporting System", Proceedings of the International Conference on Modelling, Simulation and Visualization Methods, (MSV'09), 13 – 16 July 2009, Las Vegas, USA, pp. 56-61.
- [5] Shi H., Zhang Y., Zhang J., Wan P. and Shaw K., "Development of Web-Based Epidemiological Reporting System for Tasmania Utilizing a Google Maps Add-On", Digital Image Computing: Techniques and Applications (DICTA2007), 3-5 December, 2007, Adelaide, pp. 118-123.
- [6] Zhang J., Shi H. and Zhang Y. "Web Mapping for Location Based Decision Making", International Conference on Communication Systems, Networks

and Applications (CSNA 2007) on 08-10 October 2007, Beijing, China, pp. 220 - 224.

- [7] Zhang J. and Shi H. "Geospatial Visualization using Google Maps: A Case Study on Conference Presenters ", International Multi-Symposiums on Computer and Computational Sciences (IMSCCS), The University of Iowa, Iowa City, Iowa, USA , 13 – 15 August, 2007, pp. 472-476.
- [8] Faculty Postgraduate Scholarship, Victoria University, Australia (2007-2008)
- [9] Australia Postgraduate Award, Victoria University, Australia (2008-2012)
- [10] 3rd Award, 3MT (3 Minutes Thesis Presentation), Victoria University, Australia (2011)

Table of Contents

Abstract	i
Declaration	iii
Acknowledgements	iv
List of Publications and Awards	v
Table of Contents	vii
List of Figures	xi
List of Tables	xiv
Chapter 1 Introduction	1
1.1 Background and Motivation	2
1.2 Research Challenges	4
1.2.1 Clustering analysis of epidemiological data.....	4
1.2.2 Geospatial visualisation.....	5
1.2.3 WebGIS automation application	6
1.3 Research Objectives and Contributions.....	6
1.3.1 Clustering analysis of epidemiological data.....	7
1.3.2 Geospatial processing	7
1.3.3 WebEpi.....	8
1.4 Scope of Thesis	9
Chapter 2 Literature Review	10
2.1 Introduction.....	10
2.2 Epidemiological Data	11
2.3 Clustering and Clustering Analysis	13
2.3.1 SOMs	14
2.3.2 FCM.....	17
2.3.3 K-means	21
2.3.4 Davies–Bouldin index.....	24

2.4 Geospatial Visualisation	26
2.4.1 WebGIS.....	27
2.4.2 Google Maps	28
2.4.3 Bing Maps	31
2.4.4 Comparison between Google Maps and Bing Maps	34
2.4.5 Geo-Mashups.....	36
2.5 Clustering Analysis for Geospatial Health Data Application.....	40
2.6 Summary	43
Chapter 3 WebEpi System Architecture	44
3.1 Introduction.....	44
3.2 DHHS Epidemiological Reporting System	45
3.2.1 Epidemiological data hierarchy.....	46
3.2.2 Epidemiology reporting system	48
3.3 WebEpi System Architecture	50
3.3.1 WebEpi feasibility study.....	51
3.3.2 Epidemiological data pre-processing.....	57
3.3.3 Clustering analysis of epidemiological data.....	60
3.3.4 Geo-processing of epidemiology data analysis	63
3.4 Summary	65
Chapter 4 Clustering Analysis.....	67
4.1 Introduction.....	67
4.2 Clustering Analysis	67
4.3 Epidemiological Data Analysis.....	69
4.4 Epidemiological Data Clustering.....	70
4.5 SOM Clustering Analysis for Epidemiological Data	71
4.5.1 SOM clustering algorithm	72
4.5.2 SOM cluster analysis for WebEpi data	75

4.6 FCM Clustering Analysis for Epidemiological Data.....	76
4.6.1 FCM algorithm.....	77
4.6.2 FCM cluster analysis for WebEpi data.....	79
4.7 K-means Clustering Analysis for Epidemiological Data	79
4.7.1 K-means clustering algorithm	80
4.7.2 K-means cluster analysis for WebEpi data	82
4.8 Summary	84
Chapter 5 Clustering Experiments.....	85
5.1 Introduction.....	85
5.2 Pre-Processing	85
5.3 Experiment Results.....	88
5.3.1 SOM	88
5.3.2 FCM.....	92
5.3.3 K-means	92
5.4 Experiment Evaluation.....	95
5.5 Epidemiological Data Clustering Automation.....	106
5.6 Discussion	108
Chapter 6 Geospatial Processing	110
6.1 Introduction.....	110
6.2 WebGIS	110
6.2.1 WebGIS infrastructure	111
6.2.2 WebGIS Geo-Mashups	112
6.2.3 WebGIS layer file	114
6.3 WebEpi Geo-Processing	115
6.3.1 WebEpi Geo-processing infrastructure.....	117
6.3.2 WebEpi Geo-Mashups	118
6.3.3 WebEpi geospatial layer.....	120

6.4 WebEpi Geo-processing Automation.....	125
6.5 WebEpi Case Study.....	128
6.6 Summary	134
Chapter 7 Conclusions	135
7.1 Summary of Contributions	135
7.1.1 Clustering analysis of epidemiological data.....	136
7.1.2 Geospatial visualisation of epidemiological data	137
7.1.3 WebEpi.....	138
7.2 Conclusions	138
7.3 Future Work.....	139
References.....	140
Appendices	154
A. Demonstration Files	154
A.1 Google Maps visualisation	154
A.2 Google Earth visualisation	156
B. WebEpi Guideline.....	158
C. Clustering Algorithms	164
D. CD-ROM	164

List of Figures

Fig 2.1 Tracking graphic	13
Fig 2.2 Banquet facilities maps on Google Maps	30
Fig 2.3 Housing information on Google Maps	30
Fig 2.4 Food shops on Google Maps	31
Fig 2.5 Real estate using Bing Maps	33
Fig 2.6 Movie gallery with Bing Maps.....	33
Fig 2.7 Washington state tourism map.....	33
Fig 2.8 Geo-Mashups model.....	37
Fig 3.1 Epidemiological data hierarchy	47
Fig 3.2 Epidemiological reporting system	49
Fig 3.3 Geographic information mapping system.....	52
Fig 3.4 MySQL database tables.....	53
Fig 3.5 Map feature server	54
Fig 3.6 GeoRSS conversion	55
Fig 3.7 APWeb05 presenters' mapping on Google Maps by country and region	56
Fig 3.8 APWeb05 presenters' mapping on Google Maps by city using Place Markers.....	56
Fig 3.9 WebEpi system architecture	58
Fig 3.10 Epidemiological data pre-processing	59
Fig 3.11 Epidemiological data clustering process	62
Fig 3.12 Geo-processing.....	66
Fig 4.1 SOM learning steps	73
Fig 4.2 SOM visualisation for epidemiological data.....	76

Fig 4.3 Program code of WebEpi FCM plot	80
Fig 4.4 WebEpi k-means plot function	84
Fig 5.1 Male-Standard Mortality Ratio (SMR) in 2005.....	86
Fig 5.2 MATLAB code of data pre-processing	87
Fig 5.3 SOM training results.....	89
Fig 5.4 SOM injury & poisoning.....	90
Fig 5.5 Colour coding for 29 LGAs.....	90
Fig 5.6 SOM clustering results	91
Fig 5.7 FCM clustering results	93
Fig 5.8 K-means clustering results.....	94
Fig 5.9 Comparison of clustering algorithms for Breast Cancer	97
Fig 5.10 Comparison of clustering algorithms for circulatory	98
Fig 5.11 Comparison of clustering algorithms for injury	99
Fig 5.12 Comparison of clustering algorithms for ischaemic heart.....	100
Fig 5.13 Comparison of clustering algorithms for lung cancer	101
Fig 5.14 Comparison of clustering algorithms for prostate cancer	102
Fig 5.15 Comparison of clustering algorithms for stroke	103
Fig 5.16 Comparison of clustering algorithms for 'cancer all'	104
Fig 5.17 Evaluation of the SOM, FCM and k-means.....	105
Fig 5.18 WebEpi clustering automation (Part A)	107
Fig 5.19 WebEpi clustering automation (Part B)	108
Fig 6.1 Map mashups layer model.....	113
Fig 6.2 WebEpi Geo-processing data flow diagram	118
Fig 6.3 WebEpi Geo-processing block diagram.....	119
Fig 6.4 A LGA definition in KML format.....	120

Fig 6.5 (a) Epidemiological data structures	121
Fig 6.5 (b) Epidemiological data structures	122
Fig 6.6 Epidemiological data attribute values.....	123
Fig 6.7 Colour legend.....	124
Fig 6.8 Data query	126
Fig 6.9 Geo-Mashups	127
Fig 6.10 Map feature loading	128
Fig 6.11 Sample of Epidemiology source data in Excel	129
Fig 6.12 Epidemiology KML file.....	130
Fig 6.13 Mapping layer file on Google Maps.....	131
Fig 6.14 Mapping for Males SMR in injury & poisoning.....	132
Fig 6.15 Mapping for females hospitalisation data in musculoskeletal disease	133
Fig A.1 Security settings(1)	155
Fig A.2 Security settings(2)	156
Fig A.3 File location	156
Fig A.4 WebEpi mapping	156
Fig A.5 GoogleEarth installation.....	157
Fig A.6 GoogleEarth mapping.....	157
Fig B.1 WebEpi file location	159
Fig B.4 MATLAB code(1)	161
Fig B.5 MATLAB code (2)	162
Fig B.6 Mapping file location	162
Fig B.7 KML file.....	163

List of Tables

Table 2.1 Male-standardised mortality ratio (SMR) for selected diseases in Tasmania 2003	12
Table 2.2 Map Image Comparison of Google Maps and Bing Maps.....	35
Table 2.3 Online Functionality Comparison between Google Maps and Bing Maps	35
Table 2.4 Differences between Google Maps and Bing Maps APIs.....	36

Glossary

ABS	Australian Bureau of Statistics
AJAX	Asynchronous JavaScript and XML
API	Application Programming Interface
ArcGIS	A commercial GIS software package developed by ESRI http://www.esri.com/software/arcgis
ASP .NET	A server-side Web application framework designed and developed by Microsoft, http://www.asp.net
DHHS	Department of Health and Human Services, Tasmania, Australia
Epidemiological data clustering analysis	The clustering analysis for epidemiological data
Epidemiologist	People who are conduct epidemiological studies.
ESRI	Environmental Systems Research Institute, Inc. (Esri), in Redlands, California, http://www.esri.com/about-esri/history/history-more
FCM	Fuzzy c-means http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/cmeans.html
Geocoding	Finding associated geographic coordinates
Geo-Mashups	Geospatial information mashups with geospatial related information
GeoMedia	GIS application provided by Intergraph company http://geospatial.intergraph.com/products/GeoMedia/Details.aspx
Geo-processing	Geo-Mashups and creation of geospatial layers
GeoRss	Geographically Encoded Objects for Really Simple Syndication
Geospatial layer	A geospatial file on WebGIS or GIS
Geospatial Visualisation	Information Visualisation on GIS or WebGIS
GIS	Geographic Information System

GML	Geography Markup Language
GPS	Global Positioning System
k-means	k-means http://en.wikipedia.org/wiki/K-means_clustering
KML	Keyhole Markup Language
LGA	Local Government Area
MapInfo	A commercial GIS software developed by Pitney Bowes http://www.pbinsight.com/welcome/mapinfo/
MATLAB	A high-level technical computing language and interactive environment for algorithm development http://www.mathworks.com.au/products/matlab
MySQL	My Structured Query Language
PHP	Hypertext Preprocessor, a web development language
ProMED-mail	An Internet-based reporting system http://www.isid.org/promedmail/promedmail.shtml
Silverlight	development tool for creating engaging, interactive user experiences for Web and mobile applications. http://www.microsoft.com/silverlight
SMR	Standardised Mortality Ratio
SOAP	Simple Object Access Protocol
SOM	Self-Organizing Map
SQL	Structured Query Language
SuperMap	A complete integration of a series of GIS platform software produced by GIS Software Co. Ltd http://www.supermap.com
SVG	Scalable Vector Graphics
Tiles2KML Pro	convert GIS data into KML file http://tiles2kml-pro.software.informer.com/
UML	Unified Modeling Language
URL	Uniform Resource Locator
WebEpi	Web -based Epidemiological data visualisation system designed and developed for DHHS for clustering analysis of

	epidemiological data on Google Maps
WebGIS	Web based Geographic Information System
WFS	Web Feature Service
WHO	World Health Organisation
WMS	Web Map Service
XML	Extensible Markup Language

Chapter 1 Introduction

The World Wide Web has changed almost everything, and Geographic Information Systems (GISs) are no exception (Esri Australia 2012). Web based Geographic Information Systems (WebGISs) (Fu & Sun 2010), as the combination of the Web and geographic information systems, have grown into a rapidly developing discipline. The vast majority of Internet users use simple mapping for Internet applications. For example, government agencies map the transmission of infectious diseases, real-time earthquakes and wildfire disasters online to monitor public health and safety (Fu & Sun 2010). Public health organisations' use of WebGIS not only helps to improve the community's health and wellbeing, but also helps to manage or even prevent outbreaks of infections before they occur (Australian Bureau of Statistics 2010). The research of public health epidemiological study examines the relationships between potential risk factors for diseases and their associated morbidity and Standardised Mortality Ratio (SMR). Identifying causal relationships between these risk factors and diseases is an important aspect of epidemiology (Shi et al. 2007).

Many epidemiologists use disease informatics software to conduct epidemiological data analyses. In recent decades, public health awareness has become a focus among communities. People are willing to learn and understand their community's health information. However, it is difficult for people to access this information, because of poor access to health information and the difficulty of understanding health glossaries and indicators (Australian Bureau of Statistics 2010). The actual significant achievement in this research is

the geospatial visualisation of epidemiological data. This research topic has been designed to assist in these areas. First of all, it was required to identify a data analysis algorithm which could make epidemiological analysis reports easily understood. Then, it was important to choose a platform which could visualise epidemiological analysis reports. Geospatial maps visualisation is widely used in public health surveillance; however, currently it relies heavily on expensive commercial software such as MapInfo.

1.1 Background and Motivation

The advent of electronic health records in recent decades has provided medical professionals with increasing availability of population-level health data. This has highly influenced the practice of epidemiology. Modern epidemiologists use disease informatics as a tool to identify health care needs (Australian Bureau of Statistics 2010; Shi et al. 2007). In general, the process of producing epidemiological data reports includes two components. One is the clustering analysis of epidemiological data, and the other is the geospatial visualisation of epidemiological analysis results.

The exploration of large volumes of epidemiological data to discover patterns and relationships are challenges for health research. Australian Bureau of Statistics (ABS) epidemiological data collections comprise census data, hospitalisation rates for various diseases, disease-specific death rates, cancer incidence and rates of a variety of infectious diseases notified to public health units. This data is categorised according to state, regional and local government area as well as country versus metropolitan and rural versus remote level

(Australian Bureau of Statistics 2010). Clustering algorithms can translate the massive amounts of epidemiological data into useful information.

In order to visualise epidemiological analysis results, commercial software tools such as MapInfo and ArcGIS are most widely used. These tools are sophisticated but, unfortunately, they are expensive and have limited accessibility. Also, the full mapping capability of these tools is often not required for purposes of epidemiological data visualisation (Shi et al. 2007). Furthermore the geospatial visualisation of epidemiology data is only readily understood by health researchers because of the widespread use of jargon. These factors make it highly unlikely that the public will access this data.

Originally, Department of Health and Human Services, Tasmania, Australia (DHHS) manually created epidemiological data by geospatial mapping. The steps were complicated and inefficient. During the manual mapping process, the epidemiological data was grouped manually, and sometimes grouping results could not describe the information effectively. The mapping was created using MapInfo, which can be very costly in maintenance and annual service renewal fees. However, the full mapping capability of MapInfo is not required for simple epidemiological purposes (Shi et al. 2007). The development of a precise, effective, less expensive and intuitive web-based system for the geospatial visualisation of clustering analysis of epidemiological data was necessary for the DHHS. The new system is called WebEpi.

1.2 Research Challenges

There were two major challenges in this research, one was the clustering analysis of epidemiological data; the other one was the geospatial visualisation. There are various clustering algorithms available. The selection of the best performer for clustering analysis of epidemiological data is very important. The visualisation of epidemiological data on a free geospatial visualisation platform determines the epidemiological data accessibility.

1.2.1 Clustering analysis of epidemiological data

Clustering analysis is commonly used in disease surveillance and spatial epidemiology. Clustering algorithms and clustering validation algorithms are crucial for the clustering analysis of epidemiological data. Finding an appropriate clustering algorithm and choosing a suitable clustering validation algorithm are the main challenges in the clustering analysis of epidemiological data.

Clustering algorithms which have been widely used for both epidemiological data and geospatial data analysis had to be reviewed. There are several clustering algorithms which may be suitable for epidemiological data analysis. The selection of clustering algorithms for the epidemiological data had to be based on performance.

During the clustering experiments of epidemiological data, the selected clustering algorithms may produce similar results and the results cannot be compared visually. Thus, a clustering evaluation algorithm needed to be used to validate the clustering results. The challenge was how to select the evaluation

algorithm. The evaluation algorithm had to meet the health researchers' clustering requirements for epidemiological data.

1.2.2 Geospatial visualisation

The Internet is based on standard protocols which can be accessed globally. Information has been widely shared and transferred on the Internet. Interactive Internet geospatial mapping services, which can be called a geospatial web, have been introduced as an extension of conventional stand-alone GISs in recent decades (Zhang & Shi 2007). Geospatial web services provide a better information access platform and can overcome some access limitations of stand-alone GISs. In this research, one significant task was to develop a free geospatial web service which would be applied for the geospatial visualisation of epidemiological data. However, there were certain challenges to achieving this.

The first challenge was how to combine the epidemiological clustering analysis with geospatial data. There are many techniques which have been used for data combination, but to investigate the most suitable one for free geospatial web service and clustering analysis was still a challenge.

The second challenge was how to implement a geospatial visualisation for the epidemiological clustering analysis. Geospatial visualisation describes the visualisation results on a geospatial map. Normally the clustering results are described by a plot which includes the number of clusters and the number of elements within each cluster. However, the process of the visualisation of the

clustering analysis on a geospatial map was a challenge because it had not been developed on free web mapping platform before.

1.2.3 WebGIS automation application

An effective, reliable, easy and interactive WebGIS application is the system envisioned by the DHHS. They wanted to build a fully automated web-based geospatial visualisation application for the clustering analysis of epidemiological data. However, DHHS health researchers are public health professionals, and they usually do not have sufficient IT knowledge such as programming and website development, to build a system to process the large amounts of epidemiological data. The application was required to provide a seamless transition between the clustering analysis and the geospatial visualisation of the epidemiological data.

1.3 Research Objectives and Contributions

The DHHS clustering requirements for epidemiological data and the challenges have been described in previous sections. After reviewing the clustering algorithms which are commonly used in health management and geospatial visualisation, three were selected for further investigation. Then a validation algorithm was applied to choose the best clustering algorithm of epidemiological data. Then the proposed geospatial visualisation, which comprised two processes: Geo-Mashups and geospatial layer customisation, was developed. The two major research objectives are the clustering analysis of epidemiological data and the geospatial visualisation of the results of the clustering analysis.

The successful development of the clustering analysis and the geospatial visualisation became the integral parts of the user interactive application for geospatial visualisation of the clustering analysis of epidemiological data. This system has been named WebEpi.

1.3.1 Clustering analysis of epidemiological data

Clustering experiments of epidemiological data were conducted based on epidemiological SMR. The values came from the population of the local government area. In order to solve the challenges for epidemiological clustering analysis, two steps were involved.

Firstly, three clustering algorithms for investigation were selected, i.e., Self-Organizing Map (SOM), Fuzzy c-means (FCM) and k-means. All these algorithms were applied to the epidemiological clustering experiments.

The second step was to select a clustering evaluation method. The Davies-Bouldin index was chosen to validate the clustering results of epidemiological data for the reasons described in Chapter 2 and Chapter 5. In this research this clustering evaluation algorithm was used to select the best clustering algorithm for WebEpi.

1.3.2 Geospatial processing

In order to create geospatial visualisation, the Geo-processing was developed. The development of geospatial processing for the clustering analysis of epidemiological data is based on free geospatial web services. WebEpi geospatial visualisation involves two parts.

The first part is Geo-Mashups which could be explained for this research as the combination of epidemiological data and geospatial data. Geo-Mashups had to be developed to combine the geospatial information and epidemiological clustering analysis. The Geo-Mashups engine was built to conduct mashups browsing, information classification, information rating and information formatting.

The second part is geospatial layer customisation. The reason for customising the geospatial layer is to produce an effective geospatial visualisation for the clustering analysis of epidemiological data. The colour in the map can be utilised to indicate the value of each epidemiological data cluster. In this research, the geospatial layer of coloured Local Government Area (LGA) polygons had to be created for the geospatial visualisation.

1.3.3 WebEpi

WebEpi consists of data pre-processing, data clustering and data Geo-processing. The function of data pre-processing is to convert and re-structure DHHS epidemiological data from Excel to Extensible Markup Language (XML) file format. Data clustering conducts the clustering analysis of epidemiological data in XML format. After the clustering analysis process, the clustering results are passed on for data Geo-processing. The clustering Geo-Mashups of the clustering analysis and DHHS LGA geospatial data creates a geospatial layer file. Then the geospatial layer file can be visualised on a WebGIS using a WebGIS Application Programming Interface (API). DHHS can use WebEpi to provide an open access reporting system to public.

1.4 Scope of Thesis

This thesis is divided into seven Chapters. Chapter 1 explains the research background and motivation, the research challenges and objectives. Chapter 2 is the literature review which reviews clustering analysis algorithms, WebGIS and Geo-Mashups for geospatial health data applications. In Chapter 3, the history of DHHS epidemiological reporting systems is reviewed and the WebEpi automation architecture is described from a top down approach. In Chapter 4, three clustering algorithms, SOMs, FCM, and k-means clustering processes are explained. The SOM, FCM and k-means clustering analysis results of epidemiological data are also presented in plots. In Chapter 5, clustering experiments of epidemiological data are described. The clustering analysis experimental results are validated by using the Davies-Bouldin index clustering algorithm. At this point the automation for DHHS clustering analysis of epidemiological data is realised. In chapter 6, the design and development of the Geo-processing for the clustering analysis of epidemiological data are described. At this point the automation for DHHS Geo-processing of clustering analysis of epidemiological data is implemented. In Chapter 7, the contributions of this thesis are summarised and possibilities of further research in this area are proposed.

Chapter 2 Literature Review

2.1 Introduction

The development of geospatial data and services has blossomed in recent years. Geospatial studies for epidemiological analysis have been investigated by many health researchers (Fu & Sun 2010). However, the results of these studies have limited accessibility because epidemiological analysis reports are coded in public health glossaries and indicators. Therefore, the public are not aware of these reports. The major problem of this epidemiological data access limitation is the lack of a visualisation platform. Almost all health research departments utilise commercial geospatial applications. The public are not authorised to use these expensive applications. Therefore, in order to improve public awareness and public health information accessibility, the DHHS proposed to develop a new application would integrate data clustering and web geospatial visualisation techniques.

In this thesis, the research focused on two significant components of geospatial visualisation for epidemiological analysis: a clustering analysis and a geospatial visualisation for this epidemiological data. This chapter is a literature review which reviews clustering analysis algorithms, visualisation techniques and clustering analyses for geospatial health data applications. The epidemiological data are explained in Section 2.2. The related works in clustering analysis are introduced and reviewed in Section 2.3. Then the geospatial visualisation

applications are discussed in Section 2.4. The clustering analyses for geospatial health data applications are explored in Section 2.5.

2.2 Epidemiological Data

The epidemiological data include mortality by disease, sex and year. Disease includes cancer incidence, death, hospitalisation and notified infectious. Cancer incidences include 'cancer all', colorectal, lung and prostate. Death includes all cause, breast cancer, 'cancer all', circulatory disease, injury & poisoning, ischaemic heart disease, lung cancer and prostate cancer. Hospitalisation includes accidental falls, acute respiratory infections, all cause, asthma, 'cancer all', circulatory disease, diabetes, injury and poisoning, musculoskeletal disease, pneumonia and influenza prostate cancer, respiratory disease, stroke and transport accidents. Notified infectious include all cause, campylobacter, chlamydia, hepatitis c and salmonella. Sex includes male, female and people. In Australia, epidemiological data are collected from different sources such as ABS, Department of Health and Ageing, Pharmaceutical Benefits Scheme, Australian Institute of Health and Welfare, and Department of Human Services. State, territory and local governments also collect residency health and wellbeing data for monitoring a number of disease sources (Australian Government Department of Health and Ageing 2013). Australian federal government produces enormous health reports which include

- the number deaths from cancers or accidents, and types of cancers or accidents
- the number deaths from heart disease, stroke, and diabetes and other diseases

- residential areas, and
- gender and age groups

Table 2.1 Male-standardised mortality ratio (SMR) for selected diseases in Tasmania 2003

Male-standardised mortality ratio (SMR) for selected diseases in Tasmania 2003		
Disease	LGA	SMR
All causes	Break ODay	110
All causes	Brighton	135
All causes	Burnie	104
All causes	Central Coast	95
All causes	Central Highlands	102
All causes	Circular Head	99
All causes	Clarence	93
All causes	Derwent Valley	122
All causes	Devonport	90
All causes	Dorset	102
All causes	Flinders	123
All causes	George Town	132
All causes	Glamorgan/Spring Bay	87
All causes	Glenorchy	107
All causes	Hobart	98
All causes	Huon Valley	100
All causes	Kentish	106
All causes	King Island	96
All causes	Kingborough	93
All causes	Latrobe	86
All causes	Launceston	105
All causes	Meander Valley	89
All causes	North Midlands	101
All causes	Sorell	94
All causes	Southern Midlands	108
All causes	Tasman	103
All causes	Waratah/Wynyard	105
All causes	West Coast	143
All causes	West Tamar	92

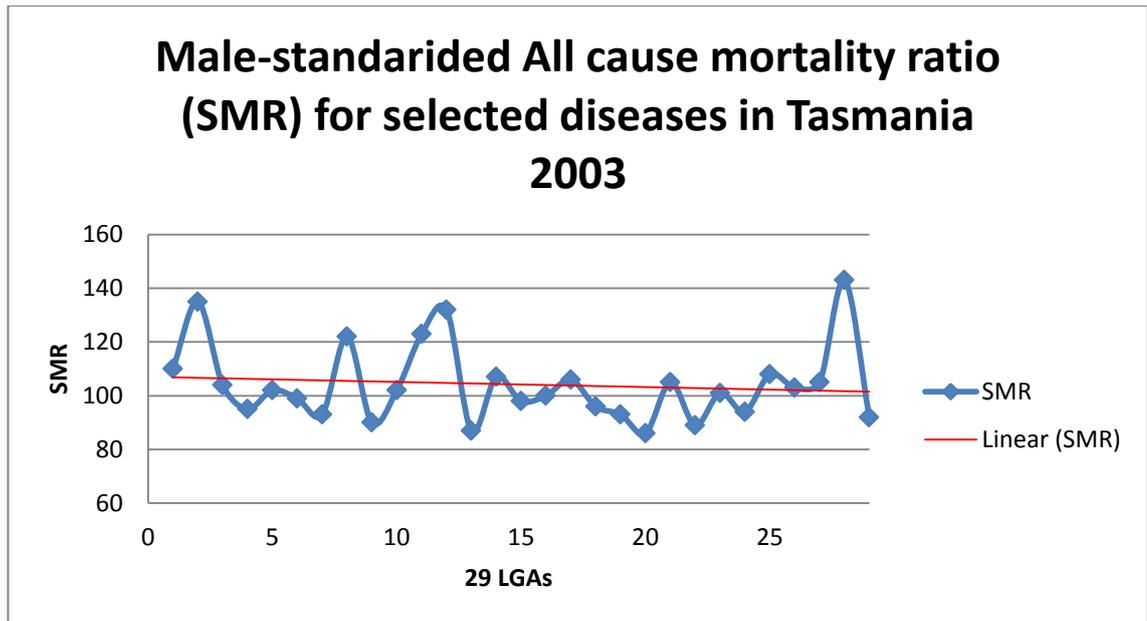


Fig 2.1 Tracking graphic

At DHHS, the epidemiological data statistics are based on region population, proportion of people and disease. DHHS produces epidemiological reports with tables and tracking graphs. In a table a mortality rate for a specific disease is listed by year and sex as shown in Table 2.1. A tracking graph plots all the numbers in the table and creates a trend line to calculate the percentage increase or decrease of mortality rate as shown in Fig 2.1 (Department of Human Service and Health Tasmania 2003). Unfortunately the epidemiological reports could only be understood by health professionals. Without any description, SMR just means purely number to public.

2.3 Clustering and Clustering Analysis

The term Cluster Analysis (CA) was first introduced by Tryon (1939) in his monograph. Cluster Analysis or Clustering is well defined by Wikipedia contributors (2013) as “the task of grouping a set of objects in such a way that objects in the same group (called cluster) are more similar (in some sense or

another) to each other than to those in other groups (clusters)". Clustering analysis aims to assign numbers of objects into clusters by the similarity and dissimilarity between the objects. Similar objects are grouped as one cluster. In the research, clustering analysis of epidemiological data focused on statistical data analysis.

The proposed clustering of epidemiological data was by grouping the specific epidemiological attributes by their SMR. However, there were two criteria for clustering analysis of DHHS epidemiological data. Firstly, the nature of the epidemiological data should strongly influence the choice of the cluster measure. Secondly, the choice of measuring should depend on scale of the data. A clustering evaluation algorithm should also be chosen. There are many clustering algorithms available, and, according to the criteria, three clustering analysis algorithms are commonly used and reviewed in this chapter: SOMs (Self Organising Maps), FCM (Fuzzy c-means), and k-means.

2.3.1 SOMs

Many researches of data analysis are applying artificial neural networks. Although the basic idea of artificial neural networks was formed in 1976 but their research did not begin until in early 1981 (Kohonen 1997). SOM was introduced by Teuvo Kohonen (1991). In his study, by adopting unsupervised learning, SOMs were able to map multidimensional training data sets into lower dimensional spaces (Kohonen 1997). SOMs are typical artificial neural networks which have been widely used in artificial intelligence unsupervised learning (Hung & Huang 2011).

Qiang, Cheng & Li (2010) suggest that having some basic knowledge of the human brain will improve the understanding of SOM processes. The development of SOMs simulates the topological maps of human brains. The design of a SOM algorithm is based on local neighbourhoods of interconnected networks. The SOM integrates dimensional plans of neuron system and complex spatial mapping (Qiang, Cheng & Li 2010). The implementation of a SOM can project high dimensional inputs, which are extracted by instance attributes of input signals, into lower dimensions. The projection results on lower dimensions still maintain the topological map of the input objects. In other words, the multidimensional data is projected into low dimensional space. The low dimensional space is able to match the input objects' topological map. Therefore, a SOM can produce better low dimensional projections for multidimensional data (Qiang, Cheng & Li 2010).

In Kohonen's SOMs, all the data processing layers can be visualised, the output layer contains restructured neurons' plan map. After the calculation of distances between neurons, the weights of the neurons will be updated. Accordingly, other neurons near the one which has been updated, have to undergo the updating process as well. Therefore one of the most significant features of a SOM is the distance relationship between the nearby neurons. However, in Kohonen's original theory, the distances between the neurons were fixed during the learning process. Therefore, the structure of neural network is not applicable for some data structures, such as liner or mesh structures (Chang, Yu & Heh 1998).

Like other artificial neural networks, the SOM mathematical model has two phases, in Jin Shuai's research, the two phases are explained as the training

phase and the mapping phase. In their SOM model, the SOM was having been trained, it was used to find and map its best clusters. (Jin et al. 2011)

The procedure of the training phases is described below (Jin et al. 2011):

- 1) Input dataset
- 2) Initialise the weight vectors of nodes as tiny random numbers, initialise the iteration count as 1
- 3) Traverse each node in the origin map and process:
 - By selecting the index of the vector to measure the distance between the input vector and weight vector. The distance can describe their similarity.
 - Calculate the distances between nodes to find out the shortest distance.
- 4) Compare the distances, update the neighbourhood of shortest distance unit by moving the neighbours close to the winning unit.
- 5) Go to step 3) unless the cluster centre differences reduced to 0.0001 or less, or the number of iterations reaches 200.

SOMs have been applied to a diversity of problems in artificial intelligence and image processing (Jin et al. 2011; Zhu & Zhu 2010a & 2010b). Usually, SOMs are defined in metric vector spaces. When using a SOM, the number of nodes and the plan map of the nodes are initialised. The request of plan map initialisation causes a dramatic limitation on the final output. Another negative aspect of utilising a SOM as a clustering algorithm is that the structure of input objects is not predictable. This negative aspect results in the difficulty of determining the initial size. Furthermore, it is harder to verify the best cluster structure for the objects (Zhu & Zhu 2010a & 2010b). Zhu and Zhu (2010a)

presented an approach to cluster the data from programming without a manual process. They created syntax trees for programming. The similarities between the syntax trees were computed in order to get a generalised mean for a SOM model. They then used programming to extract the data and present it to their SOM for visualisation. By contrast with traditional SOMs, their work can be used for data set clustering and visualisation. A similarity measurement between the programming codes can be then defined (Zhu & Zhu 2010b).

SOMs have been utilised as a classic tool in MATLAB software. A SOM toolbox is provided by MATLAB and it has been build based on neural network theory. Many classic activation functions such as a linear function are provided by the SOM toolbox (SOM 2000; Yin & Gang 2010). Users can customise the number of clusters and times of iteration by changing the SOM function reference or writing a calling function in MATLAB. Furthermore, if the user does not know how to customise the function, the user can directly call the function or sub functions in the SOM toolbox. Therefore, MATLAB SOM toolbox is very helpful in clustering analysis tasks (Yin & Gang 2010).

2.3.2 FCM

Clustering is a process involving mathematical calculations. It aims to determine the relationships between the objects, and then group the objects which are close to each other as one cluster. The difference between the artificial neural network clustering and fuzzy clustering is that in fuzzy clustering, one object can belong to one or more clusters, and have different relationships. The boundaries of fuzzy clusters are not pre-determined. In the fuzzy clustering

process, the way of discovering the relationships between objects is by rating the similarity and dissimilarity between the objects (Wang 2010).

The FCM clustering algorithm was proposed by Dunn (1973) and extended by Bezdek in 1981 (Bezdek 1981). In FCM clustering process, the vectors that have more similarity are assigned to the same cluster. Each vector presents the location of the object in the vector. Also information about the objects is analysed by mapping the objects into d-dimensional vectors. The measurements of the d-dimensional of the object are chosen as a basis for comparison with the rest of the objects. As result, the vectors location represents the relationships between the input objects (Windham 1982).

FCM is one of the most widely used and investigated clustering algorithms. It was designed for handling numerical data (Rong & Fan 2009). The FCM algorithm has been recognised as the most suitable clustering algorithm for image segmentation. In addition, FCM enables robust characteristics for ambiguity which aims to allow for elements to be in more than one cluster (Sathiracheewin & Surapatana 2011). The FCM technique is the combination of grouping of similar data (Sathiracheewin & Surapatana 2011). The combination process calculates a degree of membership of each data point to every cluster's centre. The grouping process combines the points determined by which has a high degree of membership to a cluster's centre (Sathiracheewin & Surapatana 2011).

The FCM algorithm enables multi-membership of the input objects. One object might be assigned to different clusters according to their relationships. Therefore, there might not be a single, absolute, relationship for an individual

object (Santhakshmi & Bharathi 2011). FCM finds a good partition of an image by searching a suitable prototype that minimises the object function (Rong & Fan 2009). However, the FCM clustering algorithm is sensitive to the initialisation and easily falls into a local minimum or a saddle point during iteration. The object function can be described as

$$J_m = \sum_{i=1}^N \sum_{j=1}^C u_{ji}^m d^2(x_i, v_j) \quad (\text{Santhakshmi \& Bharathi 2011})(\text{Equ. 2.1})$$

The above function has to be run iteratively to get the best solution. The iteration procedure is conducted as the following steps: (Santhakshmi & Bharathi 2011)

- 1) Initialise the value for c, m , and ε .
- 2) Initialise the fuzzy partition matrix $U^{(0)}$
- 3) Initialise the iteration counter $b = 0$
- 4) Calculate the ' c ' cluster centres $v_j^{(b)}$ with $U^{(b)}$
- 5) $v_j^{(b)} = \frac{\sum_{i=1}^N (u_{ji}^{(b)})^m x_i}{\sum_{i=1}^N (u_{ji}^{(b)})^m}$ (Santhakshmi & Bharathi 2011)(Equ. 2.2)
- 6) Calculate the membership matrix $U^{(b+1)}$

$$u_{ji}^{(b+1)} = \frac{1}{\sum_{k=1}^c \left(\frac{d_{ji}}{d_{ki}}\right)^{\frac{2}{m-1}}} \quad (\text{Santhakshmi \& Bharathi 2011})(\text{Equ. 2.3})$$

If $\max \{U^{(b)} - U^{(b-1)}\} < \varepsilon$ then stop else set $b = b - 1$ and go to calculation of the cluster centre step, step 4.

Where

x is the dataset which located in m -dimensional space.

N is the total number of data items,

c is the total number of clusters, the interval value is from 2 to N .

ε is the threshold value of clustering.

U_{ji} is the degree value of relationship between x_i in the j^{th} cluster,

m is the weighting exponent of the degree of relationship,

v_j is the initial location of the cluster centre,

$d^2(x_i, v_j)$ is a distance measured between the object and the cluster centre (Santhalakshmi & Bharathi 2011).

FCM is based on the minimisation of an objective function. The initial step of the FCM algorithm is to decide the number of clusters, and initialise the membership of matrix U . The initialisation of matrix U is conducted randomly, and the cluster centres are selected by using matrix U . The matrix U contains all datasets in each cluster (Srinivasa et al. 2005).

The points which are close to a cluster centre are assigned a high degree of membership. The membership value is the impact factor of the centre's calculation process. After all the cluster centres are calculated, the cluster centres are reselected. Consequently, the matrix of cluster centre membership

is updated according to the cluster centres which have better membership value. In order to get better cluster centres, the membership calculations not only consider distance to particular clusters, but also take into account the distance of the point from all other cluster centres. The difference of the membership matrix before the change and after the change is calculated. If the value of the difference is less than the initialised threshold, the cluster centre updating process will be terminated, otherwise the updating process will continue. The membership matrix will be renewed as well. The process of finding better cluster centres will be terminated depending on the minimum changes in the membership matrix (Srinivasa et al. 2005).

FCM as a clustering allocation algorithm has many advantages, such as: it is adaptable to many areas of application, able to handle large amount of data and is efficient in calculation (Chu et al. 2010). FCM has been applied to many areas of data analysis. Chi utilised FCM for forecasting bus incidents. This application used MATLAB to verify the effectiveness of the method (Chi et al. 2010). FCM has been used in health data analysis as well. Yun Chi Yeh and Hong-Jhih Lin applied FCM for classifying cardiac arrhythmia on ECG signals. Using FCM, the total classification accuracy was approximately 93.57% (Yeh & Lin 2010).

2.3.3 K-means

Clustering is a process of discovering the similarity of a set of objects. A great deal of research has been conducted on clustering and finding the relationships between objects. There are many ways to discover clusters. Finding dissimilarity between objects is also considered as an effective option of

discovering clusters. The dissimilarity of clusters can be explained as finding the disjoint clustering. K-means has been recognised as the most reliable disjoint clustering algorithm (Yu, Soh & Bond 2005).

This algorithm was originally developed by MacQueen in 1967 (MacQueen 1967). K-means has been considered the most efficient unsupervised learning approach (Zhou & Liu 2006). K-means aims to discover the similarity and dissimilarity between objects and clusters. Objects which have similar features are clustered, and objects with dissimilarities are assigned to other clusters. Therefore, compared with other clustering algorithms, k-means also considers dissimilarities between objects and clusters (Zhou & Liu 2006).

K-means clustering algorithm is efficient and effective. This clustering algorithm adopts a segmentation cluster approach. K-means clustering divides objects into k clusters, therefore k needs to be initialised before the clustering process. The centre of a cluster is the average value of the objects in the cluster. The process of clustering is according to the changes of the centre (Wu & Yao 2010). The number of clusters is taken as a parameter in the k-means algorithm. After getting the number of clusters, k-means will distribute the input objects into the clusters. The distribution of objects is based on the similarity between the objects and other objects within the same cluster. As a result, the measurement of similarity within the cluster is calculated by the mean value of all the objects within the same cluster (Deng & Mei 2009).

K-means algorithm process is as follows:

- 1) Firstly, the value of k has to be pre-defined, and the cluster centres for all the clusters are randomly selected.

- 2) Objects are assigned to their closet cluster, according to their similarities to the cluster centres (Wu & Yao 2010).
- 3) Re-calculate the average values of all data objects within each cluster, and set the new average value as a new cluster centre, and go to step 2 (Wu & Yao 2010).
- 4) The new centre is compared with the original centre, if the cluster centre needs to be changed, return to step 2. The iterative process stops when the difference reaches 0.0001 or less (Deng & Mei 2009).

In the k-means clustering process, each object must belong to one cluster, it does not allow two or more clusters to contain the same object. K is the number of clusters, and the centres of the k clusters are selected from the input objects. In order to get better results, it is recommended to run k-means many times (Wang et al. 2009). K-means requires having initial cluster centres and cluster centres which are sensitive to clustering results. The k-means classification process uses the smallest distance to determine the maximum degree of membership categories. It is a gradually iterating algorithm (Yang & Deng 2010).

There are many variations of the k-means clustering mathematical function. Some of them have become very popular in recent years. Bradley and Fayyad (1998) presented a technique for clustering initialisation. Their technique was based on estimation of the distribution in order to optimise the iterative algorithm. Khan and Ahmad (2004) proposed a solution to get better initial cluster centres before running the k-means clustering algorithm. Their solution was based on two aspects, one is the similarity between the objects, the second is the reasons why objects are located in the same cluster. However, all these algorithms are computationally intensive (Gu, Zhou & Chen 2009).

K-means has been used in the geospatial area as well. Xiao JiaoHuo et al utilise k-means for GIS map copyright protection. They used a k-means clustering algorithm to create a watermarking schema, based on a polygon type of ESRI (Environmental Systems Research Institute, Inc.) which is one of the most professional GIS development and GIS service support companies. Huo et al. (2011) distributed polygons into clusters and then utilised a watermark bit to calculate the mean distance between polygons within a cluster. They improved the performance by updating coordinates according to the mean distance between polygons. In the end, their research resulted in outstanding invisibility (Huo et al. 2011).

2.3.4 Davies–Bouldin index

For epidemiological data clustering, three clustering algorithms have been chosen: SOMs, FCM and k-means. The three clustering algorithms could produce very similar results, therefore a clustering evaluation algorithm had to be applied for the selection of the best performing clustering algorithm for epidemiological data clustering. Clustering evaluation methods were selected based on an internal criterion which assigns a high value to the clustering algorithm that can produce clusters which have high degree of membership within the same cluster. On the opposite side, clustering evaluation assigns a low value to clusters which have low similarity within clusters. However, there is one disadvantage of using an internal criterion and that is that the measurement of a high degree of membership might not be helpful in discovering the best clustering results (Carvalho & Tome 1999).

However, according to DHHS requirements, the closest distance between elements in clustering was the most important element. The Davies–Bouldin index is an interval criterion, and was selected for evaluation purposes. It aims to discover the clusters which have high similarity by calculation of the interval distance of objects within cluster.

Davies–Bouldin index algorithm can not only be allocated a criteria to select better clustering results for objects, but can also discover the differences between different clustering algorithm computation results. The clustering validation results are not affected by the total number of clusters and the type of algorithm which produces the cluster results. It can also be used to assess a cluster seeking algorithm. The following distance functions: dispersion measurement (Equ. 2.4) and characteristic vector (Equ. 2.5) were chosen: (Davies & Bouldin 1979).

$$S_i = \sqrt[q]{\frac{1}{T_i} \sum_{j=1}^{T_i} |X_j - A_j|^q} \quad (\text{Davies \& Bouldin 1979}) \text{ (Equ. 2.4)}$$

X_j is a dimensional feature vector where T_i is the number of vectors in cluster i .

A_j is the centroid of cluster i

$$M_{i,j} = \sqrt[p]{\frac{1}{T_i} \sum_{j=1}^{T_i} |a_{k,i} - a_{k,j}|^p} \quad (\text{Davies \& Bouldin 1979}) \text{ (Equ. 2.5)}$$

$M_{i,j}$ is a measure of separation between clusters where $a_{k,j}$ is the k th component of the n -dimensional vector. The WebEpi SOM, C-means and k-means clustering results were validated using the Davies–Bouldin index.

2.4 Geospatial Visualisation

GISs consist of geospatial application tools, geospatial data analysis, geospatial file production and none geospatial data integrated web services (Colantonio et al. 2011). Currently, many popular GIS software platforms are available. The selection of platform plays an important role in mapping data. At present, popular GIS platforms include ArcGIS, MapInfo, SuperMap and GeoMedia. They all have advantages and disadvantages. ArcGIS was developed by ESRI (<http://www.esri.com/>) (Esri 2012). It contains several products such as ArcGIS for desktops, ArcGIS for mobiles, ArcGIS for servers and ArcGIS online. MapInfo is owned by the Pitney Bowes Software company (<http://www.pbinsight.com>). MapInfo has been developed for many platforms. Based on Mapinfo, Pitney Bowes developed many applications such as Enterprise Asset Management, Enterprise Address Management and Customer Communication Management. SuperMap GIS was developed by the SuperMap Company (<http://www.supermap.com/>). They provide a SuperMap GIS platform, based on Java and .Net, which enables GIS developers to choose their own GIS platforms. The SuperMap Company also developed applications for Desktop GIS, Service GIS, Component GIS and Mobile GIS. GeoMedia Intergraph is owned by Intergraph Pty Ltd (<http://www.intergraph.com/global/au/geomedia/>). Intergraph has products such as GeoMedia 3D, GeoMedia Image, GeoMedia Map Publisher etc.

Although many commercial GIS map applications are available right now, ArcGIS and MapInfo are considered as the most popular GIS platforms in the Australian health industry. This is because ArcGIS and MapInfo target is large

companies or organisations. Their customers are from both professional and national or multinational companies and organisations. At the same time, ArcGIS and MapInfo compete with each other in mapping functionality, map visualisation and geospatial data management. They also support each other by the significant achievement of data conversion between two the different platforms (Wang et al. 2011). These tools are sophisticated but expensive and have limited accessibility. Also, the full mapping capability of these tools is often not required for simple epidemiological purposes. An alternative that is less expensive and more accessible to health professionals and the public would, therefore, be useful (Shi et al. 2007).

2.4.1 WebGIS

The web provides a platform which enables data sharing and web applications. Data is presented to people in a visualised way such as images and text. At the present time, however, images and text alone are not enough for information presentation (Markovic & Kloos 2009). With the popularisation of the Internet, GISs are moving from isolated, standalone systems to Internet applications which can be called WebGISs. The advantages of WebGISs include real-time accessibility, public accessibility and low cost. WebGISs integrate GISs and Internet technologies.

Unlike traditional GISs, WebGISs describe the differences between different hardware, software application, communication protocols and data storage. WebGISs create integration between different applications and data storage. In addition, WebGISs not only enable data sharing between multiple data sources, but also construct a framework of geospatial data sharing and non-geospatial

data sharing. The development of WebGISs offers an effective and efficient methodology to conduct geospatial research. However, the implementation of WebGISs places higher demands on hardware and GIS software, particularly the software support GIS system architecture and GIS application development (Han et al. 2010). Google Maps and Bing Maps are becoming the most popular, free WebGIS applications on the Internet. However, which one performs better is still under investigation, as each company works on the improvement of their WebGISs services. In addition, WebGIS services are being continually increased and improved.

2.4.2 Google Maps

Google Maps which was launched by Google in February 2005. It is a free electronic map service (Su 2011). Google Maps provides a high resolution Satellite view, Map view and a Hybrid view of the Satellite and Map views. Google Maps integrates the Internet map services with third-party applications. Google Maps has become more and more popular with wide scale service applications and has free accessibility (Fu et al. 2010).

Google Maps is based on the Asynchronous JavaScript and XML (AJAX) technology. Like other professional web mapping services, Google Maps also implements zooming and dragging map tools. In addition, Google Maps has improved map loading speeds by reducing the times of web page reloading. Google Maps has also developed functionalities such as spatial search by location information (Tan et al. 2008). Google also launched the Google Maps APIs (<http://code.google.com/apis/maps/index.html>) to allow the customisation of map output for specific data applications on Google Maps. By using the APIs,

developers develop their client side script to access Google Maps server applications. Google Maps is a functional platform of geospatial data, because Google Maps can unite different sources of web service interface for information visualisation (Tan et al. 2008).

In order to extend the functionality of Google Maps, the GIS file is the most important data source. The GIS file uses a standard format for encoding geographical information. Most GIS files are produced by government or professional GIS companies. Meta data of a GIS file often includes elevation data in either raster or vector form. Map layers are usually expressed as point, line and polygon combined with coordinate system description. Google Maps is associated with XML-based Keyhole Markup Language (KML), Geographically Encoded Objects for Really Simple Syndication (GeoRss) and Scalable Vector Graphics (SVG). They serve as modelling languages for geographic information systems and also produce an open standard format for geospatial data transactions (Shi et al. 2007).

A number of business and social organisations use Google Maps to enhance their projects such as in the hospitality, real-estate, food and areas as shown in Fig 2.2, 2.3 and 2.4 respectively. Another example of a Google Maps application is SchoolFinder (<http://www.schoolfinder.us/>). The United States of America (USA) Government SchoolFinder application uses satellite views to provide general information for over 130,000 public and private schools across the USA (Shi et al. 2007). Healthmap is an example of the use of Google Maps for health applications (<http://healthmap.org/en/>). It is the result of collaboration between Harvard-MIT Division of Health Sciences and Technology, the Children's Hospital Informatics Program and Harvard Medical School. The

application provides global disease alert mapping. The Healthmap project aims to enhance surveillance of infectious diseases through the integration of a number of datasets, including World Health Organisation (WHO) data, ProMED-mail and Google News (Shi et al. 2007). The Healthmap application extends the usefulness of each individual dataset. The data can be used for both research and interventions to improve public health communication (Shi et al. 2007).

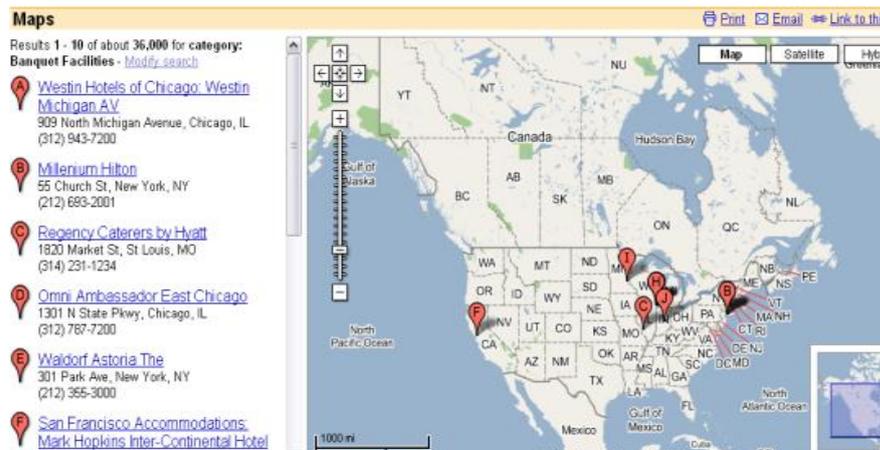


Fig 2.2 Banquet facilities maps on Google Maps

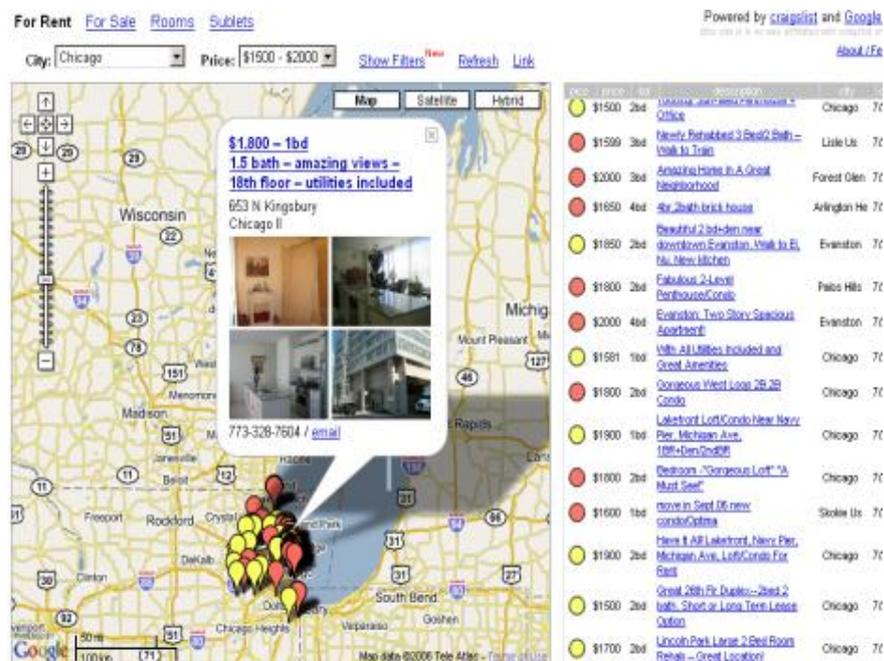


Fig 2.3 Housing information on Google Maps

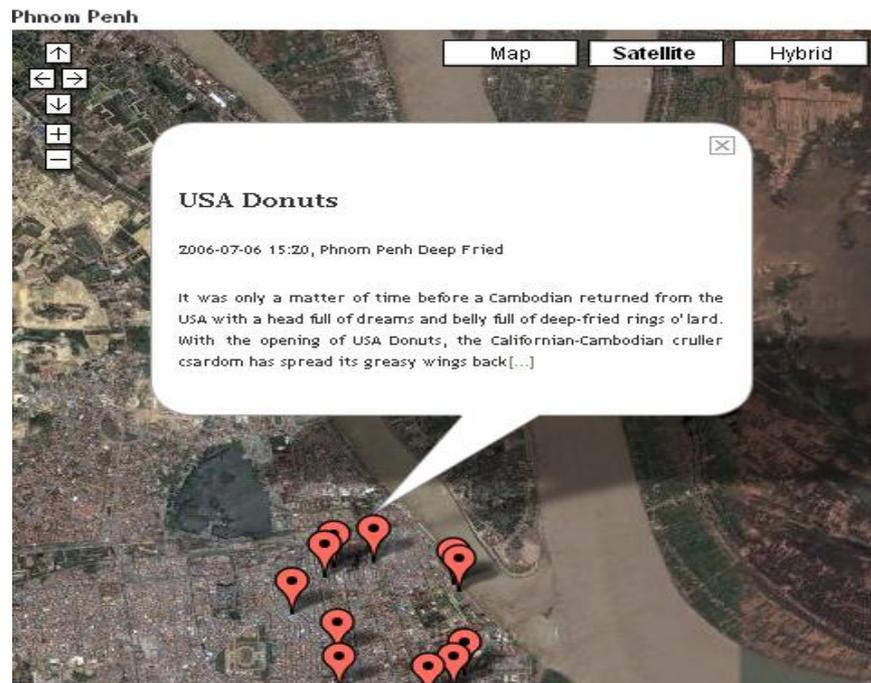


Fig 2.4 Food shops on Google Maps

2.4.3 Bing Maps

Bing Maps was developed by Microsoft, and was released in December 2005. Bing Maps provides Aerial, Road and Hybrid views. The Aerial view is the visualisation from the sky, and Road view can be described as road map visualisation. Hybrid view is the combination of Aerial and Road views. Bing Maps provides a 3D view for buildings. The Road view can show the name of roads in cities round the world and also provides a search function by road, city and country. Bing Maps Aerial view resolution is not very good. Users can directly find house geographical locations by providing street number, street, city and country. Users can also add pushpins on Bing Maps, and these pushpins can be automatically saved on the local hard disk. When users launch the web page next time these pushpins are loaded into the map. Bing Maps also enables drawing path and polygon (Zhang, Shi & Zhang 2007).

Bing Maps has been embedded into many web applications to support geospatial visualisation. Bing Maps supplies APIs for WebGISs developers to develop their own applications. APIs allow developers to overlap their customised geospatial layers onto Bing Maps. Bing Maps APIs support Geography Markup Language (GML), KML and GeoRSS geospatial file types. By comparing with Google Maps, Bing Maps also supports time series animation or geospatial visualisation. Time series animation is extremely helpful for GIS decision making. The comparison of time difference is the principle mechanism for historical data analysis. Furthermore, Bing Maps APIs also have polymorphism. It can be used in different web development techniques such as Silverlight, AJAX and Simple Object Access Protocol (SOAP). Bing Maps has fully demonstrated the characteristics of object oriented development (Qu et al. 2011).

Many businesses and organisations have begun to use Bing Maps for their business services or promotions. Bing Maps has also been very popular with real estate, retail and travel agency businesses. For example, Sperry Van Ness is commercial real estate investment brokerage firm in the USA. It provides real estate services across 35 states in the USA with geographic location maps on Bing Maps as shown in Fig 2.5. Movie Gallery uses Bing Maps road map for customers to search online for movie rental store locations as shown in Fig 2.6. It allows their customers to make decision about which video stores they would like to visit. Washington State Tourism has embedded Bing Maps into their websites so that web users can search for Points Of Interest (POI) such as hotels, parks and landmarks as shown in Fig 2.7 (Zhang, Shi & Zhang 2007).

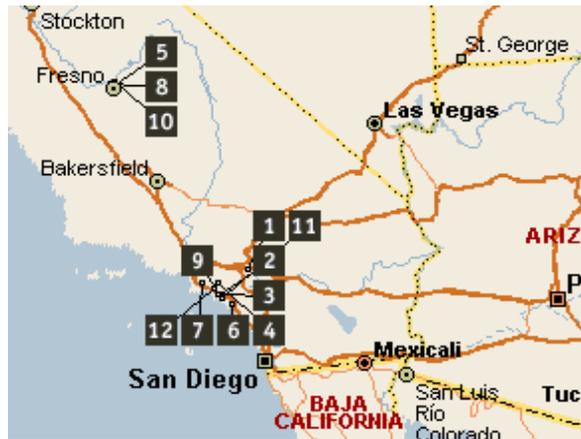


Fig 2.5 Real estate using Bing Maps

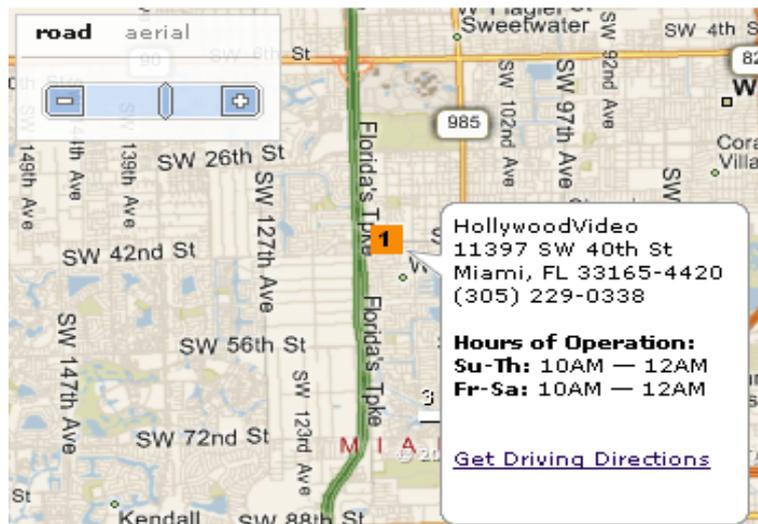


Fig 2.6 Movie gallery with Bing Maps

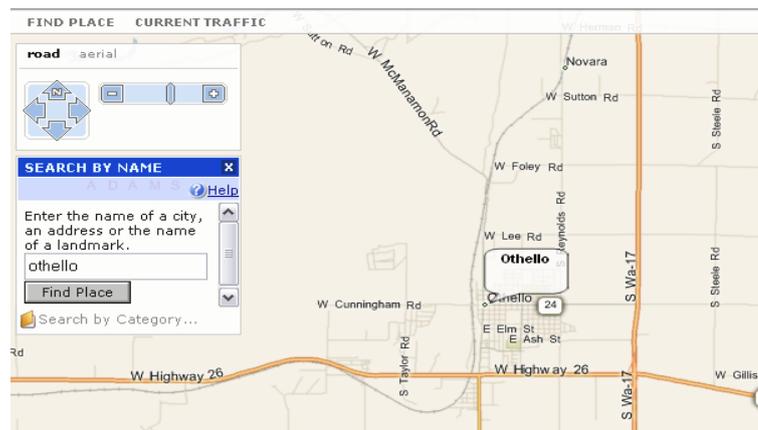


Fig 2.7 Washington state tourism map

2.4.4 Comparison between Google Maps and Bing Maps

The two free WebGISs servers, Bing Maps and Google Maps have attracted a great deal of attention from web users, developers and researchers. Each server has its unique way to present the data and its own approach to business solutions. The similarity and difference between Bing Maps and Google Maps in the areas of Map Visualisation, Online Functionality and API usability are discussed as follows (Zhang, Shi & Zhang 2007).

Google Maps provides three types of web images, i.e. Map view, Satellite view, and Hybrid view. Satellite view presents photos taken from the sky, Map view is the road map of a city, and Hybrid view overlaps the Map view over the Satellite view. Google Maps is continually updating its images, and more and more cities round the world have become clearer; even cars on roads and local intersections are clearly visible (Zhang, Shi & Zhang 2007).

On the other hand, Bing Maps provides Aerial view, Road view and Hybrid view. The Aerial view is similar to Satellite view in Google Maps while Road view is more or less same as Map view in Google Maps. The road map can show the name of road in many cities, it also provides a search function by road, city and country which is also available in Google Maps. However, Google Maps provides better zooming in images than Bing Maps. The comparisons are summarised in Table 2.2 (Zhang, Shi & Zhang 2007).

Table 2.2 Map Image Comparison of Google Maps and Bing Maps

Web Map	Satellite View	Road View	Hybrid View	3D View	Max Zoom in Satellite View	Max Zoom in Road View
Google Maps	✓	✓	✓	✓	✓	✓
Bing Maps	✓	✓	✓	✓	x	✓

The Google Maps website provides search functions such as search by address, by business and by direction between two places. Users can create their local business marks on Google Maps although at this stage the service is only available for twelve countries such as Canada, Australia, Japan, the United Kingdom and the United States. Searching based on street number and Get Directions function are both working on Google Maps in Australia (Zhang, Shi & Zhang 2007).

Bing Maps also has an address search function by specifying street, city and country. Users can custom tag the location by adding pushpins on Bing Maps, and the pushpins' coordinate locations can be saved on the user's hard disk. Bing Maps has a drawing polygon layer function. A functionality comparison between the two servers is shown in Table 2.3 (Zhang, Shi & Zhang 2007).

Table 2.3 Online Functionality Comparison between Google Maps and Bing Maps

Web Map	Search place by road, city, and Country (Australia)	Add place mark by pushing point	Draw Path/Polygon	Search business place
Google Maps	✓	✓	x	✓
Bing Maps	✓	✓	✓	✓

Both Google Maps and Bing Maps provide APIs for business and personal applications. Their APIs are compared based on attributes like API Key Request, Geocoding Request Limit, Route and Drive Direction, Geo Layer File Size Limited, and SVG support (Zhang, Shi & Zhang 2007).

The Google Maps server asks for an API key when the client sends a query to convert address to the geographic coordinate. There is also a limit on the maximum number of Geocoding requests for a particular API key: 50,000 per day. Route and Drive Direction are available on both Google Maps and Bing Maps. Both Google Maps and Bing Maps provided APIs can load a Geo layer file, but, once the layer file size exceeds 4MB, the webpage will generate an unexpected error. SVG is a kind of XML file which describes two dimensional graphics and is supported by both Google Maps and Bing Maps. Table 2.4 illustrates the major difference between Google Maps and Bing Maps APIs (Zhang, Shi & Zhang 2007).

Table 2.4 Differences between Google Maps and Bing Maps APIs

API	API Key Request	Geocoding Request Limit	Route and Drive Direction	Geo Layer File Size Limit	SVG Support
Google Maps	✓	✓	✓	✓	✓
Bing Maps	✗	✗	✓	✗	✓

2.4.5 Geo-Mashups

Web mashups (Markovic & Kloos 2009) are web applications generated by combining contents, presentations or application functionalities from disparate web sources. There is a large amount of geospatial information available

through the Internet, but geospatial data exploring and conversion are challenges. Web mashups for geospatial information can also be called Geo-Mashups (Zhang, Shi & Zhang 2009).

Geo-Mashups enable the remixed geospatial layers or features from different sources to be combined into an integrated geospatial file. Mashups technology promotes communication between geospatial data access and the web geospatial data source. Geo-Mashups can be in the form of map-based applications and services mashups. There are some free Geo-Mashups web services available, such as Google Maps, Yahoo Maps and Bing Maps. These free Geo-Mashups services support map APIs with built-in AJAX, JavaScript, and Flash software. The map APIs allow users to build amazing map applications (Zhang, Shi & Zhang 2009). Healthmap is a typical WebGIS application utilising Geo-Mashups which enhances the surveillance of infectious diseases through global disease alert mapping (Zhang, Shi & Zhang 2009).

The general form of a Geo-Mashups model can be summarised as shown in Fig 2.8

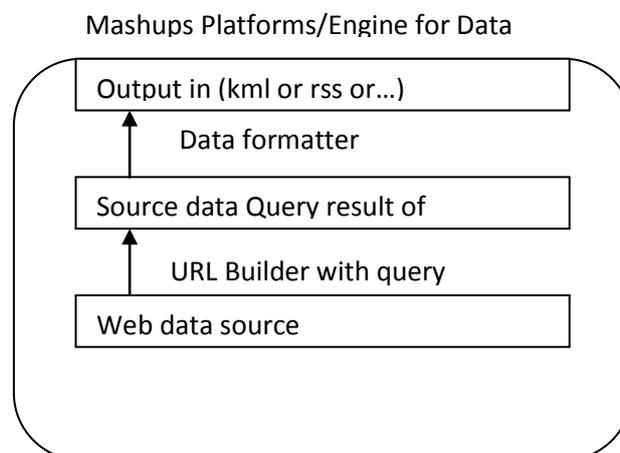


Fig 2.8 Geo-Mashups model

First the web data source has to be presented and a Uniform Resource Locator (URL) builder is constructed. The results of the source data are retrieved from the URL Builder using a semantic query. The results may contain different formats such as XML, EXtensible Stylesheet Language (XSL) and Really Simple Syndication (RSS). A uniform output can be achieved by using a data formatter to convert results between different data formats. The output is then well formatted and is ready to be used in the next stage.

Because of the advantages of Geo-Mashups, many applications utilise them for data presentation and decision making. An application presented by Rui Zhu et al utilise Google Earth API based mashups for visualisation of a weather induced disaster warning information system (Zhu et al. 2011). The reason for their use of the Google Earth API is that it provides a 3D visualisation. The Google Earth API integrates Satellite view, Map view, Hybrid view and 3D images for the major cities. The Google Earth API offers basic functions of 3D GIS efficiently (Zhu et al. 2011).

In Zhu's system, its Google Earth and Central Meteorological Observatory servers have been used as the API publisher and the data publisher (Zhu et al. 2011). In their research, they built a website for this application, they used the Apache HTTP Server, commonly referred to as Apache, to publish the website for Geo-Mashups. (Zhu et al. 2011). In their publication they explained that their servers consisted of three servers connected by the Internet. The communication programming between server was developed based on a Geo-Mashups program in HTML and JavaScript. (Zhu et al. 2011) They also used their client side website for disaster forecasting and warning. The client-side browser used JavaScript to access the Google Earth and Central

Meteorological Observatory servers. Their example demonstrates how a Geo-Mashups application could be created to generate results in the client-side browsers (Zhu et al. 2011).

A research related visualisation of Geo-Mashups was conducted by Wood et al (2007). They demonstrated a Geo-Mashups case study. Google Earth is utilised as a geospatial visualisation tool. The source data is located in a My Structured Query Language (MySQL). Hypertext Preprocessor (PHP) is applied as the middle layer between MySQL and a geospatial data processing server. LandSerf (<http://www.soi.city.ac.uk/~jwo/landserf/>) was selected as the geospatial map processor which aims to conduct spatial data calculations. The KML data type can be visualised on Google Maps and Google Earth. The KML data type was adopted as the final geospatial data type for the application. The integration of geospatial application and a different source of data enables geospatial researchers to query, store, process and create of geospatial maps. The successful implementation of this case study proves that Google Earth can be utilised as professional geospatial data analysis tool.

Wood et al (2007) used Geo-Mashups to implement better interaction and geospatial visualisation for their WebGISs application. By using KML as their geospatial layer file type, it saved a lot of effort in the development of their geospatial visualisation tool and GIS functionalities, because KML is an open source language. Many applications have already been developed by others. Instead of writing new GIS functionalities, GIS developers can use pre-written application source code and manipulate it to suite their own applications. From the existing implementations, it is clear that an open source geospatial language will save a lot of time for developers (Wood et al. 2007).

2.5 Clustering Analysis for Geospatial Health Data

Application

The combination of clustering analysis and geospatial studies is becoming very common in health data management, eg. Pathak et al's reference point location based geospatial analysis of heart disease incidence in Florida metropolitan areas (Pathak et al. 2011). The USA collects heart disease death incidence for non-transported and transported patients. This data is analysed by a geospatial clustering algorithm. In the clustering results, they demonstrated that there are significant differences between clusters that are close to a hospital and those which are far from a hospital. In their research, they used residential addresses as geospatial references to record the incidence of heart related deaths. They utilised ESRI's ArcGIS street name database as their map base reference (Pathak et al. 2011).

Geospatial research on the incidence of heart disease deaths in Florida has been successfully conducted. In this research, the geospatial data analysis was based on a street level map base. It also included route calculation between a patient home address and the hospital address. The distance calculation results have been clustered as well. In their research, they claim that they use probability weights for the inferential testing of clusters. This method has been implemented on public health literature research. In their future research, they are going to conduct geographical epidemiology using the same methodology (Pathak et al. 2011).

Another research project based on the detection of clusters of a rare disease over a large territory was conducted by Goujon-Bellec et al. Their research

proved that clustering analysis facilities could be combined with a geospatial application (Goujon-Bellec et al. 2011). They used circular and elliptic scan methods for detecting regularly shaped clusters. The circular scan method was based on a circular window that scanned the entire map from one area to another. The elliptic scan method was based on the screen scan. The size of window was set by the length of its semi major axis. The shape ratio of the window was set by semi-major and minor axes. The angle degree was defined by horizontal line and its semi-major axis. In order to include a large geospatial area, the major axis length was set differently. In their research, they utilised living zones as geospatial data and mapping regions (Goujon-Bellec et al. 2011). They achieved a significant conclusion from the research; it is that a circular scan does not fully support the determination of clusters. When compared with the circular scan, the elliptic scan performed better in large geospatial area cluster analysis and it was compatible for liner clustering detection. The elliptic scan also performed better over large geospatial areas (Goujon-Bellec et al. 2011).

Basara and Yuan (2008) utilised a SOM as their community health data clustering algorithm to produce geospatial visualisation for their clustering results. They used the SOM to classify 511 communities into five clusters based on 92 environmental variables. ArcGIS is a very expensive, commercial geospatial mapping software package and has been used by many organisations. They used ArcGIS as their geospatial visualisation tool. Their research methodology enabled health researchers to combine the clustering algorithm and geospatial visualisation application for decision making. However, their research paper did not present the SOM clustering source code and

detailed processes for geospatial visualisation using ArcGIS (Basara & Yuan 2008).

Basara and Yuan demonstrated the significant relationship between SOM classifications and the geographic visualisation of population-adjusted rates for selected diseases (Basara & Yuan 2008). Their research presented an important relationship between environmental conditions and health reports. The reports described the area environment as a major factor in public health (Basara & Yuan 2008). The research showed that it was possible to combine environment analysis and a geospatial system. The research focused on population infectious diseases at a community level. Population health analysis was based on geospatial environmental conditions related to health outcomes. Various environmental assessments might have been utilised as alternative attributes for 'practice-based' health assessments in cases where data was limited (Basara & Yuan 2008).

K-means has been used in GIS health data management as well. Lebel et al. conducted research on health inequalities in Quebec, Canada. Before they adopted attributes associated to the clustering algorithm criteria, the data analysis results were not compatible for geospatial visualisation, because the analysis results did not include the possibility of overmatching and maximising inside clusters. Therefore, they decided to use k-means for finding clusters. Their research aimed to produce maps which included population and social conditions for the geospatial area. K-means clustering algorithm clusters geospatial locations by finding the similar attributes. The results also described the differences between adjacent clusters. The results of their research have proved that different sources of data and various data analysis methodologies

can be integrated together for geospatial data analysis. This concept can be expanded to other applications (Lebel, Pampalon & Villeneuve 2007).

2.6 Summary

The clustering algorithms SOMs, FCM and k-means have been thoroughly reviewed in the chapter because of their popularity in health data management and geospatial visualisation. It is very hard to decide which one performs the best in epidemiological data clustering analysis. Therefore, in order to find the best performance clustering algorithm, it was necessary to conduct experiments on the three clustering algorithms using the Davies-Bouldin index clustering validation algorithm. The DHHS required the proposed epidemiological data report system to utilise a free, popular and good quality image WebGIS service. Google Maps is now becoming the most popular web based geospatial information browsing service. According to the web geospatial visualisation review, Google Maps seems to be the most suitable geospatial visualisation platform for epidemiological data. It meets all the DHHS geospatial visualisation criteria.

Chapter 3 WebEpi System Architecture

3.1 Introduction

One of the research tasks was to develop an automated process for epidemiological data clustering analysis on a web-based geospatial system. In the past, the DHHS used MapInfo, a commercial software package, for their epidemiological data reporting services. This was costly and time consuming as it required manual analysis of the epidemiological data. In order to improve public accessibility, the performance of epidemiological data clustering analysis and geospatial visualisation, the DHHS were very interested in developing a new application for Web based geospatial visualisation, this application is called WebEpi.

In this chapter, the existing epidemiological reporting system and the proposed WebEpi system are introduced and described from a top-down approach. Section 3.2 reviews the previous DHHS epidemiological reporting system, including the epidemiological data architecture and the epidemiological reporting system architecture. In Section 3.3, WebEpi is presented. There are three major processes: data pre-processing, data clustering and geospatial visualisation.

3.2 DHHS Epidemiological Reporting System

The Population Health Epidemiology Unit of the DHHS uses Cancer Incidence, Death, Hospitalisation, and Notified Infectious data to conduct monitoring and surveillance of the health of the Tasmanian population (Shi et al. 2007). The data is coded according to LGA. Traditionally, commercial mapping software tools such as ArcGIS and MapInfo have been used to map this type of epidemiological data. However, these tools are very expensive, demand high computer resources and have very limited accessibility (Shi et al. 2007). The epidemiology data geographical mapping was obtained manually and the mapping steps were complicated and inefficient. The manual mapping process produced the proposed epidemiology layer files and all the records were manually clustered, but it was very time consuming in exporting source data, converting data format, rating data and creating a layer file. Normally it took 4 weeks to create one group of epidemiology data. Owing to the large amount of data, the data clustering procedures required a special process (Zhang, Shi & Zhang 2009).

For epidemiological data analysis, the ratio value was classified into five categories from low to high. The classification was done by manually setting a range value. In order to visualise the data, a commercial software tool MapInfo was used, although the full mapping capability of this tool was often not required for simple epidemiological purposes. The DHHS were eager to find an alternative that was less expensive and more accessible to health professionals and the public (Shi et al. 2007). Therefore, a new system WebEpi was proposed to replace the existing DHHS Epidemiological reporting system. It aims to

integrate free web mapping services with epidemiological data analysis (Shi et al. 2007).

3.2.1 Epidemiological data hierarchy

The epidemiological data provided by the DHHS is formatted in a data structure which is shown in Fig 3.1. The epidemiological data comes with four epidemiological groups: Cancer Incidence, Death, Hospitalisation and Notified Infectious. The data is collected annually. Each epidemiological group of data contains three people groups: Males, Females and Persons. There are different disease categories for each of the four epidemiological groups and their people groups. Each category has an SMR for specific conditions for the current category disease. The SMR is assigned to its corresponding LGA. There are twenty-nine LGAs in Tasmania. LGAs are the bottom level of the DHHS epidemiological data hierarchy. Above the LGAs are disease categories. People groups are the parent class of disease categories. There are, in total, five levels in the epidemiological information hierarchy, as shown in Fig 3.1.

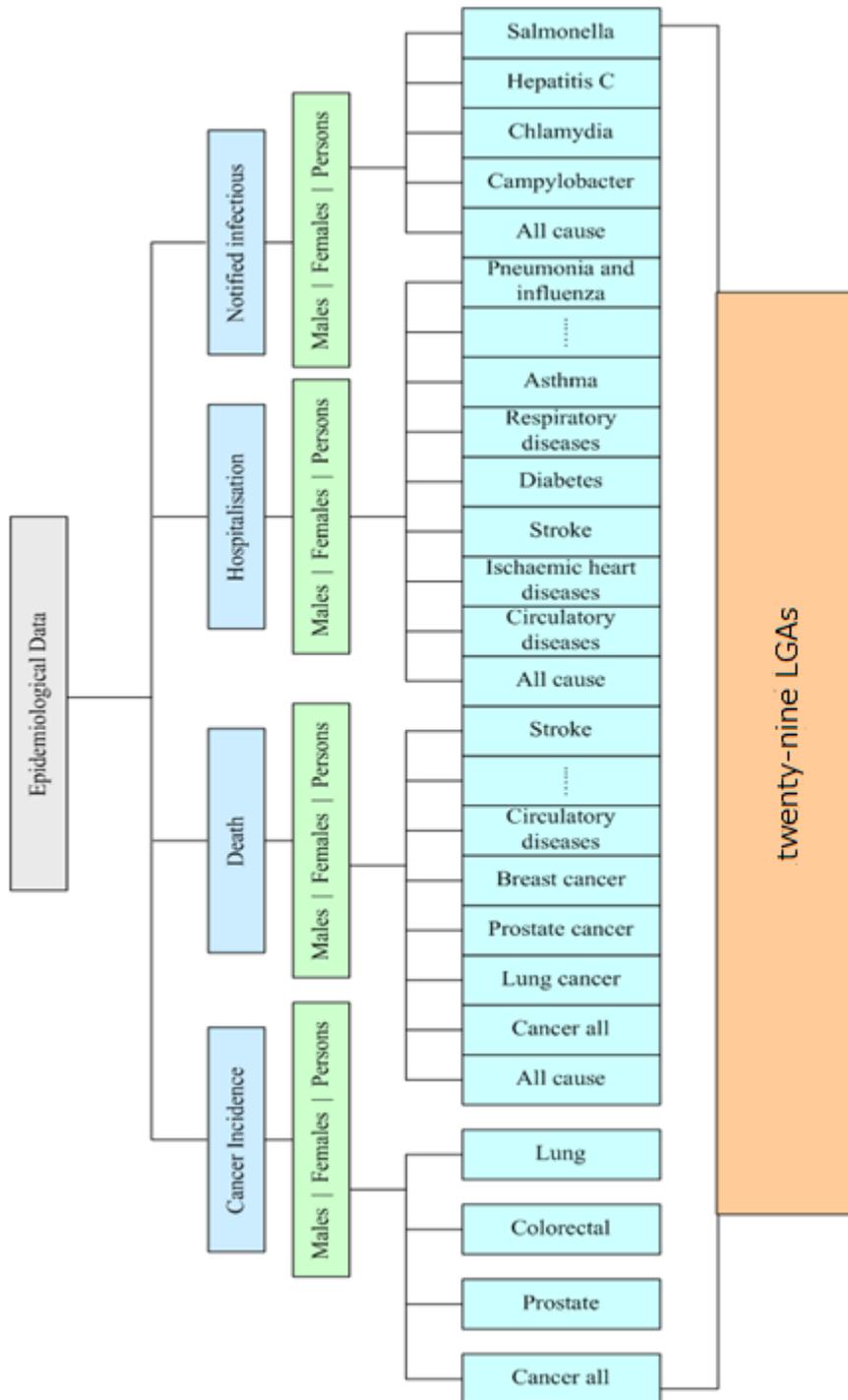


Fig 3.1 Epidemiological data hierarchy

3.2.2 Epidemiology reporting system

Health researchers manually entered epidemiological data into Excel files. The columns and dataset format of the Excel file were customised for the MapInfo mapping process. After all the epidemiological data was entered into the customised Excel files, all these files were passed on to the GIS analysis staff. GIS analysis staff could then start the mapping process. MapInfo requires its users to have technical knowledge of GIS, such as GIS queries, geospatial object combination and the auditing of geospatial data tables. MapInfo provides macro facilities, but if users want to write their own mapping macros they have to know the Mapbase program language. In addition to this, maintaining MapInfo is very costly, as the license needs to be renewed every year. After all the maps for epidemiological data were created, these maps would be sent back to health researchers for further study. DHHS MapInfo maps could not be shared on the Internet, as health researchers could not access the visualisation map inside the DHHS intranet. The system architecture diagram for the epidemiological reporting system is shown in Fig 3.2.

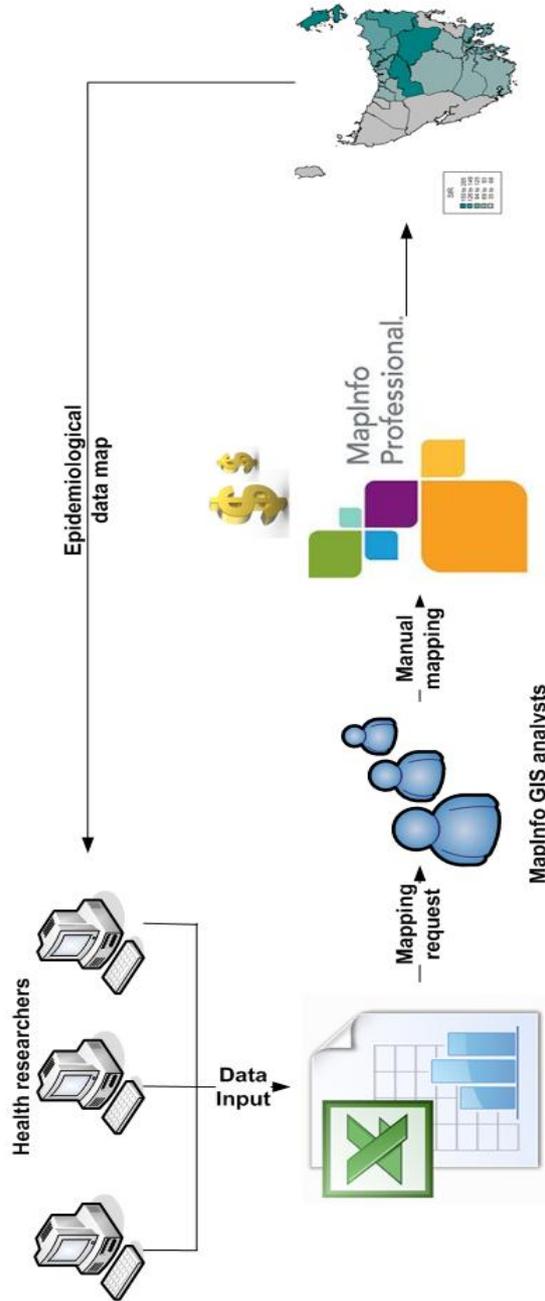


Fig 3.2 Epidemiological reporting system

3.3 WebEpi System Architecture

WebEpi is a system proposed to replace the existing epidemiological reporting system. The aim of WebEpi is to use free web mapping services for geospatial visualisation and to improve the efficiency of the clustering analysis of epidemiological data. WebEpi has been designed to transform the boring and massive epidemiological data into useful epidemiological information, so that it can provide decision support for public health researchers. The transformation of epidemiological data into epidemiological information is comprised of two tasks. One is to discover the information of epidemiological data, which involves data extraction and data clustering. The second task is information visualisation. Information visualisation deals with representing concepts and data in a meaningful way (Spence 2007). Seeing is believing. A picture is worth a thousand words. Therefore, it is better to visualise the epidemiological information as an image. In this thesis, the popular Google Maps has been used as the epidemiological data visualisation platform.

In order to replace the powerful and expensive MapInfo platform with Google Maps, a feasibility study had to be carried out before the commencement of WebEpi development. The details of this study are described in Section 3.3.1. In general, the WebEpi system architecture includes three major processes: the first is epidemiological data pre-processing, as presented in Section 3.3.2; the second is epidemiological data clustering analysis, and it is addressed in Section 3.3.3 and the last is Geo-processing of the epidemiological data analysis results, as described in Section 3.3.4.

3.3.1 WebEpi feasibility study

Google Maps is one of the most popular tools for mapping geographically-referenced health data on the Internet. The interactive web mapping services provided by Google Maps have the power to deliver and visualise data online, without cost. Typically, Google Maps supports Map view, Satellite view, and Hybrid view. Google Maps is also free and allows a reusable web mapping service (Shi et al. 2007). It is a highly responsive visual interface which uses AJAX technology. It contains detailed street and aerial image data and has open APIs to allow customisation of map output for specific data applications (Shi et al. 2007).

Before the development of WebEpi, a feasibility study of Geo-Mashups with Google Maps APIs was carried out based on a case study of mapping the APWeb05 (Asia Pacific Web 2005) conference presenters' geographical locations. It utilised Geo-Mashups and Google Maps and was composed of a web interface, a web server and a Google Maps server as shown in Fig 3.3. Web users can browse the map information layer. A Web server is used to maintain the website, query the database and generate the map features. A MySQL database maintains geographical location information for each country and region. A Google Maps server is used as the map server and provides geospatial visualisation on Google Maps (Zhang & Shi 2007).

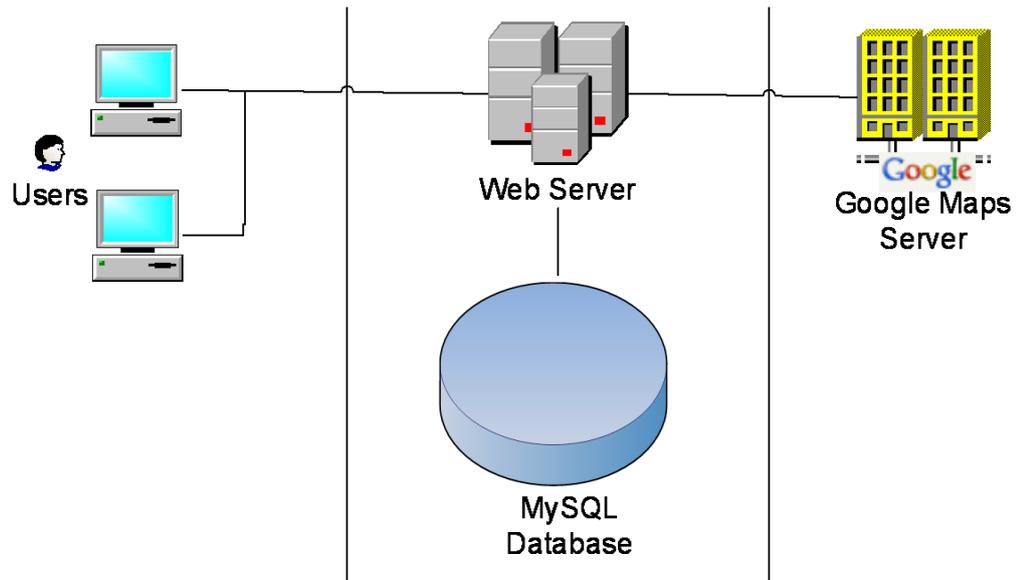


Fig 3.3 Geographic information mapping system

The system allowed web users to query the world wide location by executing the developed PHP scripts on the web server. The summary of presenters' location information, based on country and region, was then saved in a MySQL database. The countries were then divided into five categories. Each category number was assigned to a specific geographic feature and stored in the web server. Then the country region map was rendered in different colours, based on the category number. Finally, geospatial visualisation was realised on Google Maps (Zhang & Shi 2007).

This geospatial information mapping system had three modules: database mining, map feature creation and GeoRSS generation. In the MySQL database, the conference presenters' information table was generated for the geospatial information mapping system. The address information of presenters was then calculated by country region in the database. Each country was assigned a category number according to the number of presenters from the country (Zhang & Shi 2007).

The category number was used to determine the colour coding of the country and region on Google Maps. A PHP script was used on the web server to produce location information in the PresenterInformation table. The data mining results were stored in Country and PlaceMark tables as shown in Fig 3.4. Each PlaceMark had to be assigned to the same category number as its Country because the category number was used to determine the colour coding of PlaceMarks on Google Maps (Zhang & Shi 2007).

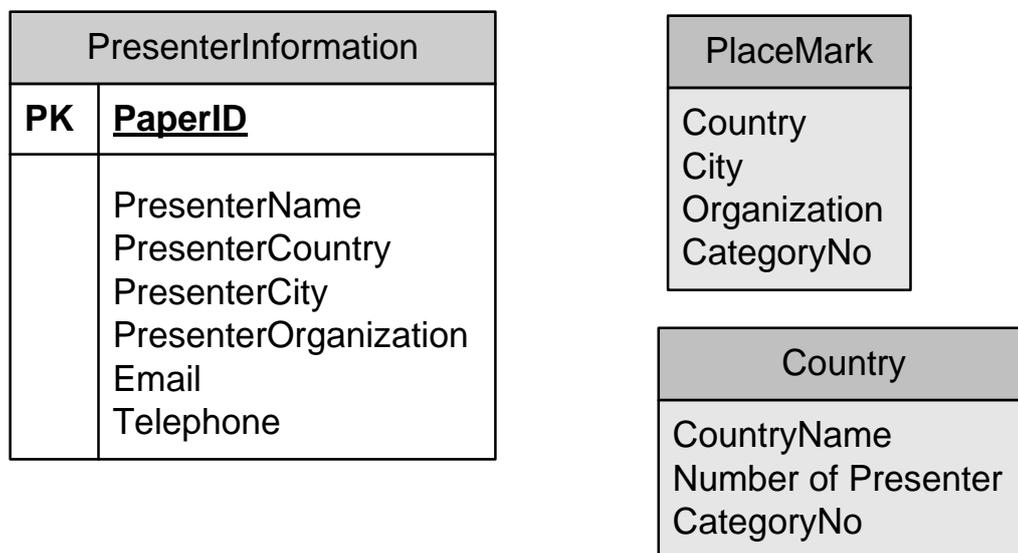


Fig 3.4 MySQL database tables

A map features schema was used to create geospatial feature data and non-geospatial data. It had two components: Country Region Features and PlaceMark Feature as show in Fig 3.5. All this data was organised in the GeoRSS data format. The GeoRSS file contained colour and name features. Each country had its own GeoRSS file which maintained polygon data of the country and region. In the GeoRSS Country Region file, the polygon data was tagged by <georss:polygon>. Each country also had its own GeoRSS

PlaceMark file which contained several items. Each item described one PlaceMark in the country. The coordinate information for each PlaceMark item was tagged by <georss:point>. All the items in the Country share the same coloured icon (Zhang & Shi 2007).

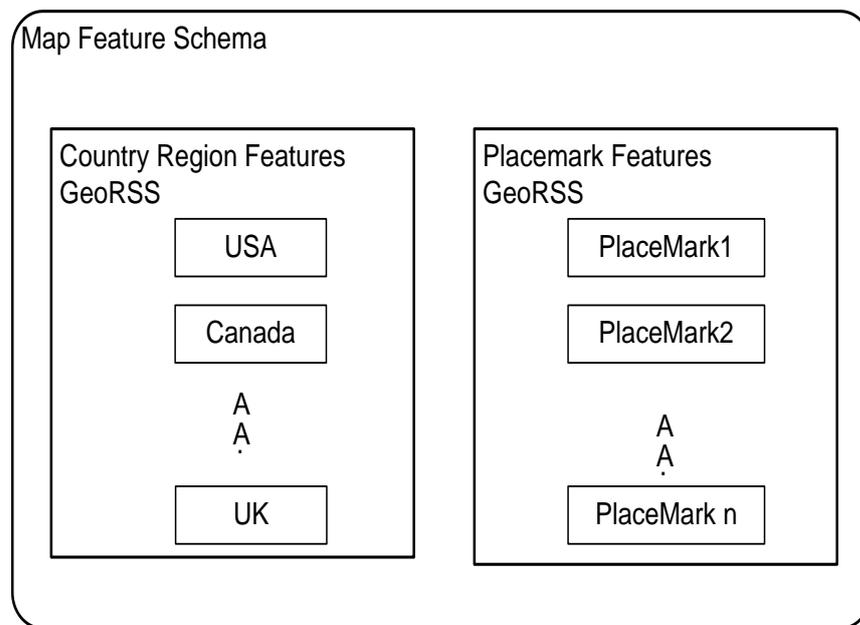


Fig 3.5 Map feature server

After the GeoRSS file had been created, the geographic coordinates needed to be recorded in the file. As mentioned before, a Google Maps API has geocoding capability, i.e. the process of finding associated geographic coordinates. Therefore, the address coordinate information could be obtained by sending HTTP requests from the web server to the Google Maps API server. The address was extracted from a GeoRSS tag <Country> <City> <Organisation> which composed part of the HTTP request. If the Google Maps server could not find the geographic coordinates of an organisation, the web server would resend the country and city information (Zhang & Shi 2007). The web server could then save the coordinate data into a GeoRSS file. Saving the coordinates

from the map feature server had two benefits: one is that it would not run out of the geocoding request limit of Google Maps and secondly, the other equally important benefit, was that it would minimise the web site browsing time when rendering the same layer file on Google Maps. The last module was GeoRSS manipulation. GeoRSS tag attribute values were generated according to Country Region and PlaceMark coordinates and category number (Zhang & Shi 2007). The details of how the new GeoRSS file was generated are shown in Fig 3.6.

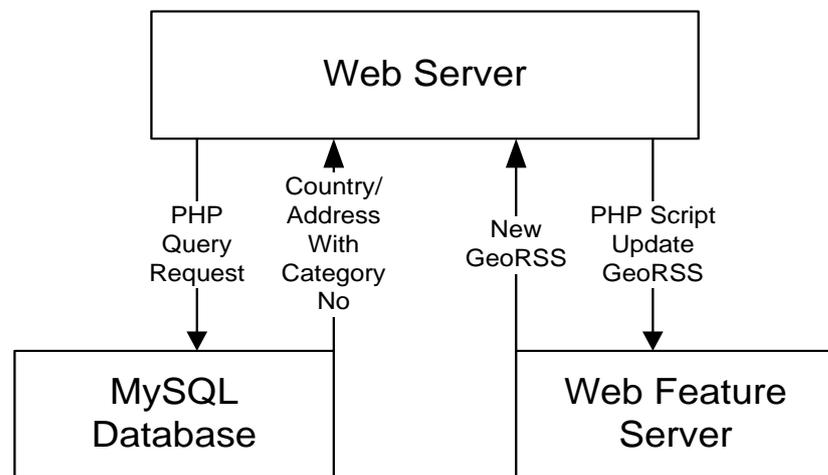


Fig 3.6 GeoRSS conversion

The map features were created in GeoRSS format and users could choose to select by the country and region layer as shown in Fig 3.7 or Country and the Place Marker layer as shown Fig 3.8. The visualisation of the country and region layer shows that the highest number of presenters came from China, which had more than 27 presenters (Zhang & Shi 2007).

The Europe area had the smallest number of presenters in APWeb05. It is clear that most of the APWeb05 conference presenters came from Asia and the Pacific Area, such as China, Australia, Korea and Japan as shown in Fig 3.8.

Once the conference organisers analysed the geographic locations of their conference presenters, attendees and program committee members, they could make decision where and how to market and promote APWeb conferences more effectively in the future (Zhang & Shi 2007).



Fig 3.7 APWeb05 presenters' mapping on Google Maps by country and region

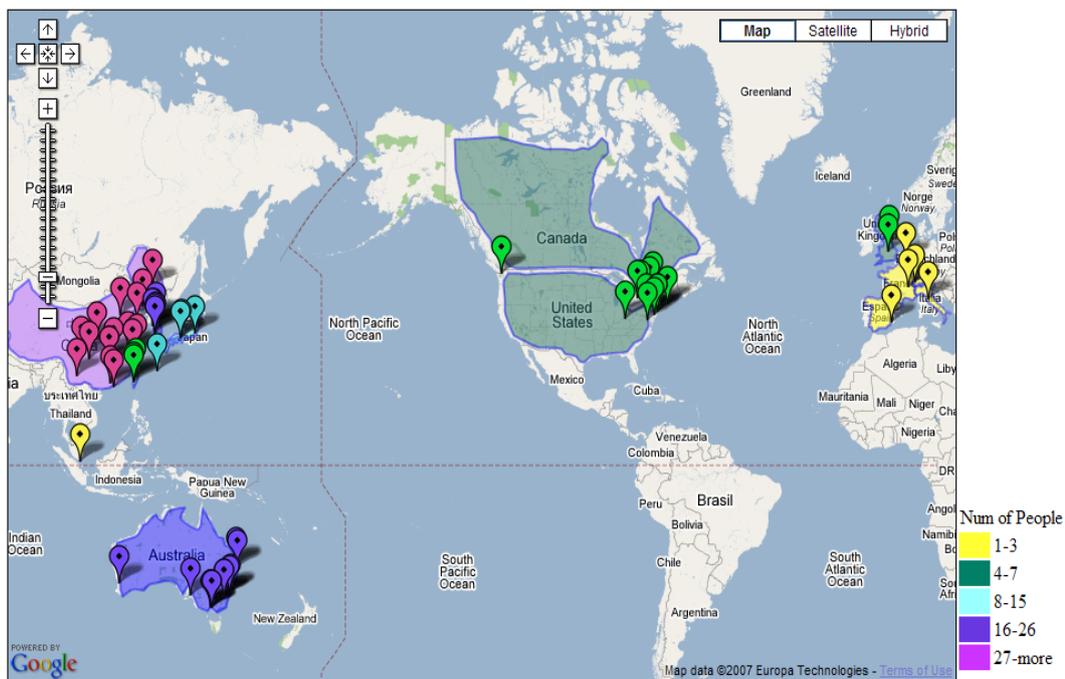


Fig 3.8 APWeb05 presenters' mapping on Google Maps by city using Place Markers

From the feasibility study, it has been shown that interactive web mapping services can be used for the geospatial visualisation of data on Google Maps. It has been demonstrated that the web mapping services can provide a new approach to presenting information geographically (Zhang & Shi 2007). With the enhancement of web-based mapping services, it can be concluded that this advanced technique can be used for the mapping of epidemiological data on Google Maps for the DHHS. Based on the study, the proposed architecture of the WebEpi development is shown in Fig 3.9.

3.3.2 Epidemiological data pre-processing

The epidemiology data provided by DHHS comes as a Microsoft Excel file. The column names and data set locations in the Excel file are not constant. During pre-processing, MATLAB R2008b is employed to process the data and unify the epidemiological data so that it can later be loaded for the clustering algorithm. The automation of the epidemiological data pre-processing algorithm is illustrated as a data flow diagram in Fig 3.10, more details of which will be explained in Chapter 5. The processing consists of four parts:

- a) Load epidemiology data into MATLAB, and classify the data into four categories as Cancer Incidence, Death, Hospitalisation and Notified Infectious.
- b) Re-group data into different diseases according to their categories see also Fig 3.1.
- c) Re-structure the data by gender
- d) Create records for each LGA

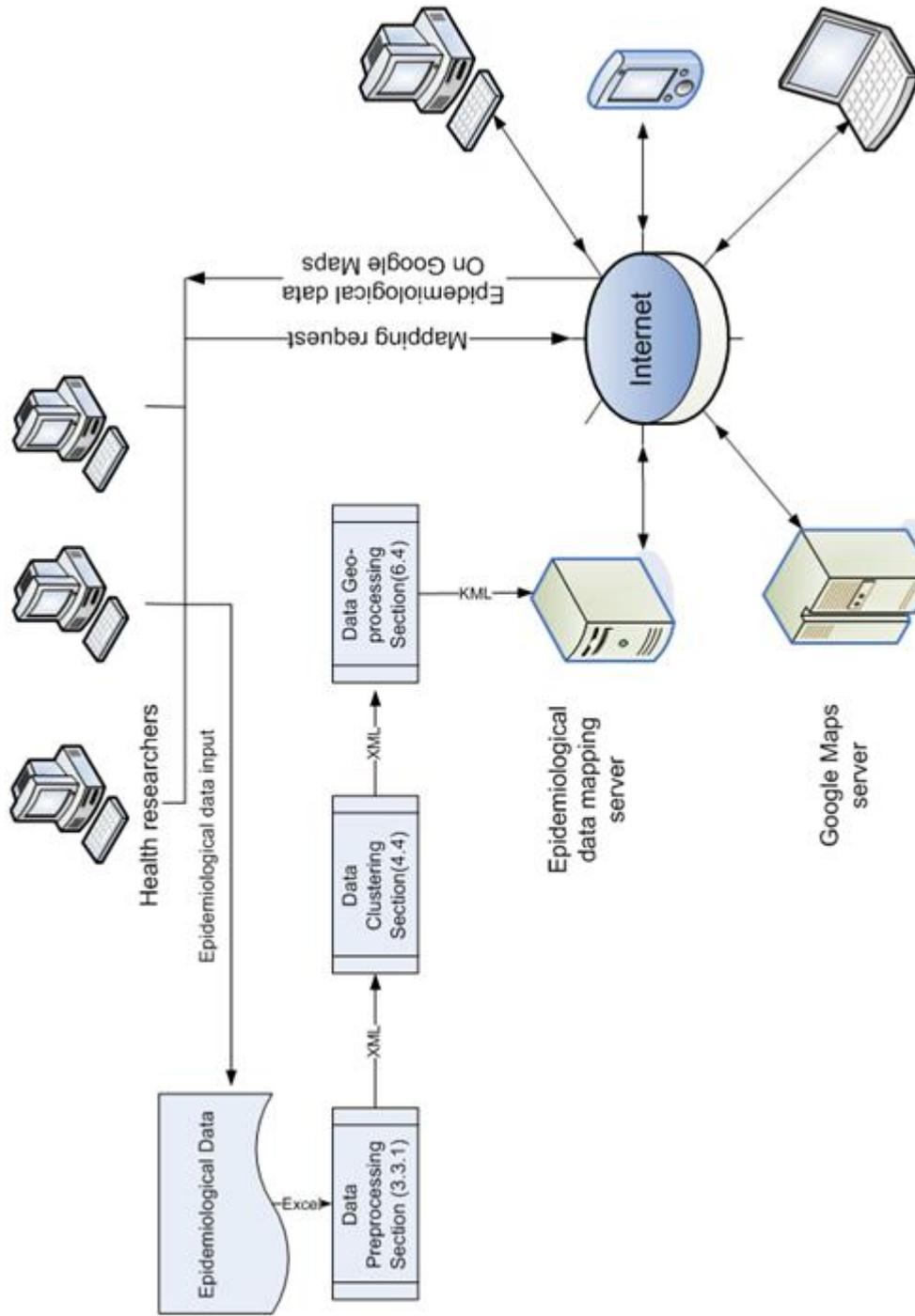


Fig 3.9 WebEpi system architecture

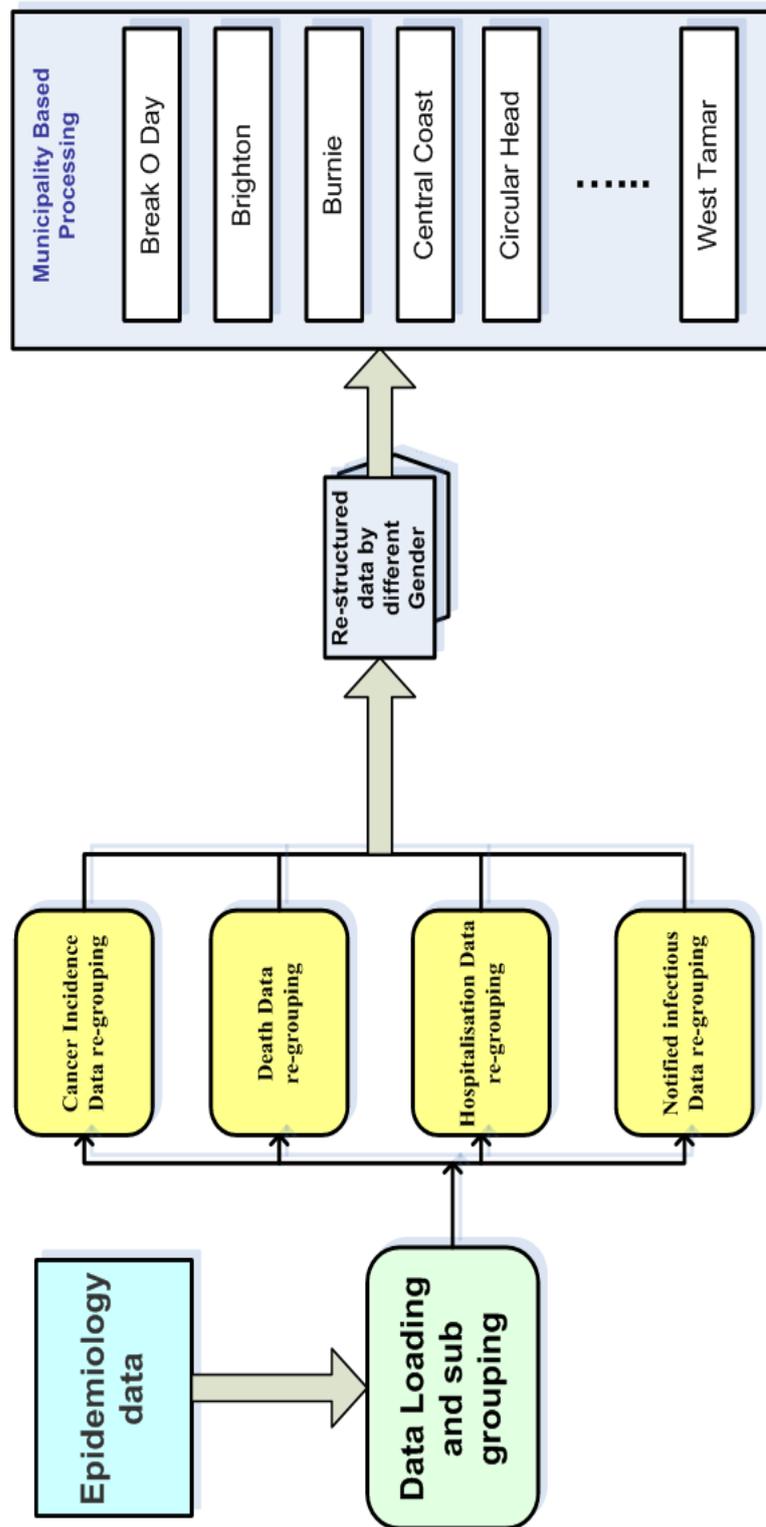


Fig 3.10 Epidemiological data pre-processing

3.3.3 Clustering analysis of epidemiological data

The grouping of objects which have similar attributes is an everyday occurrence. It is also a process of self-learning. Rousseeuw (1990) describes object classification as a process of child learning. Children can distinguish boys and girls, chicken and ducks. Children learn classification continuously. Children also expand the learning of classification in their future study (Rousseeuw 1990). The aim of epidemiological data analysis is to form different clusters according to their geospatial information and different disease SMR. Clustering analysis is the art of finding clusters in data.

The first objective was to study how many types of DHHS data needed to be clustered, and then to investigate the methods for pre-processing the data to make it ready for clustering analysis. The information itself might not be ready for clustering because of its structure or format. Therefore, investigation in preparation for pre-processing of the data was necessary. The data had then to be restructured or reformatted to make appropriate for clustering analysis.

The second objective was to decide which clustering algorithm was to be selected. The selection of clustering algorithm depends both on the information hierarchy and on the aims of the clustering analysis. Several clustering algorithms were applicable for epidemiological data. However, one selection criterion is not sufficient for the selection of the clustering algorithm. Just analysing the clustering algorithm itself is not enough. Therefore, it was better to run all the selected algorithms and analyse the clustering results. Even if there was a winning algorithm, it was highly recommend that geospatial visualisation

experiments on the clustering results should also be conducted for better information visualisation (Rousseeuw 1990).

In order to conduct epidemiological data clustering experiments, three commonly used clustering algorithms were chosen: SOMs, FCM and k-means. The flow chart of the epidemiological data clustering process is described in Fig 3.11. The comparison of the clustering analysis performance will be addressed in Chapter 4. The clustering training process is the core of the clustering analysis. There are four steps in the clustering training process:

- a) Choose input vectors.
- b) Calculate the distances between every input vector and cluster centres.
- c) Select the cluster centre which is the closest to the input vector of the training network.
- d) Discover the winner neuron and its weight vector. Lastly the cluster centre is updated by moving it closer to the input vector.

When all the data has been run through the clustering analysis process, the clustering results would be ready for Geo-processing.

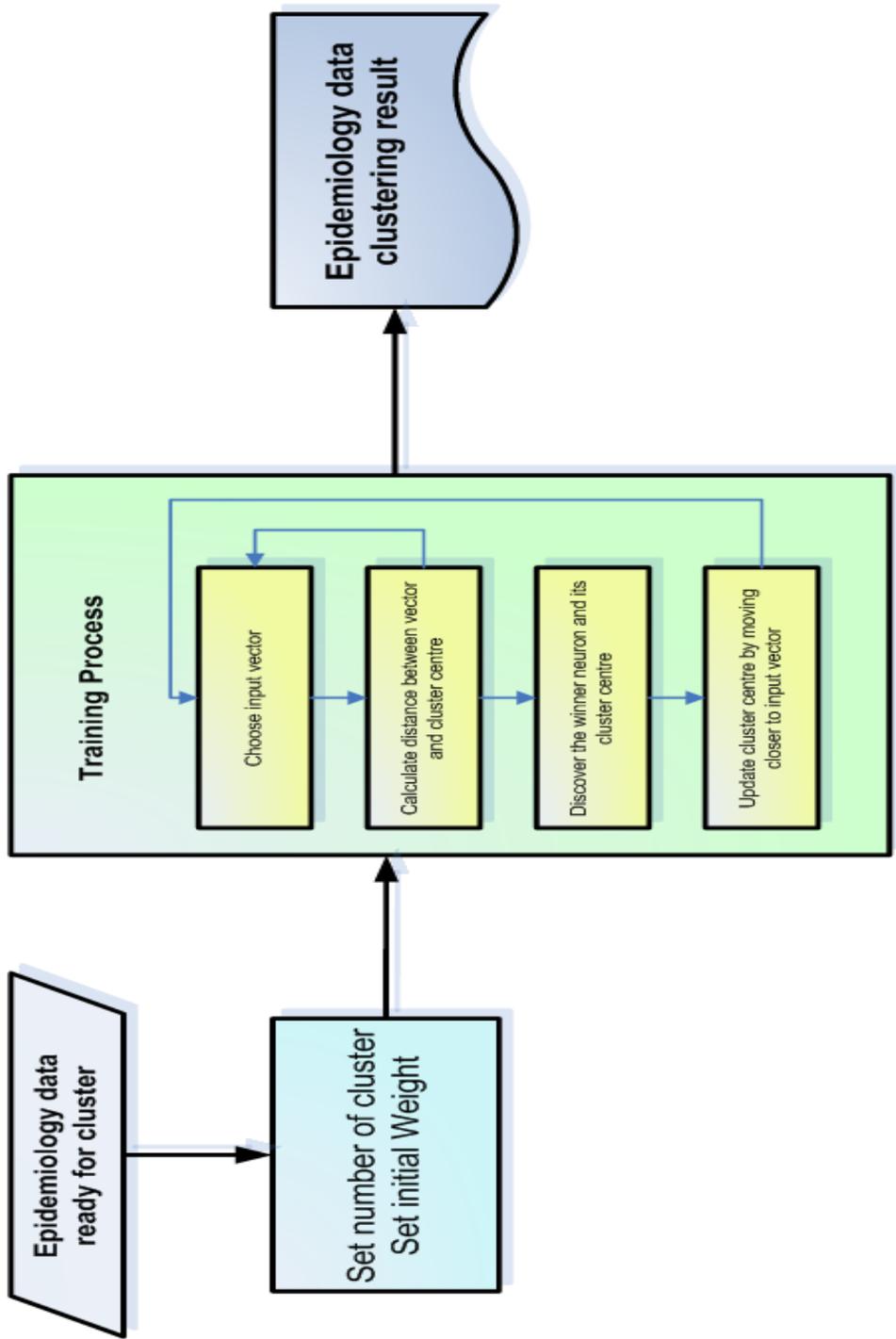


Fig 3.11 Epidemiological data clustering process

3.3.4 Geo-processing of epidemiology data analysis

The geospatial web is a web service which integrates geospatial maps and Internet data sources together. This service provides data query, data sharing and data management functions. Geospatial technology can be applied to many web applications because it provides an open standard for data transactions and file formatting (Jain & Wu 2007).

The geospatial web is extremely well supported for a wide range platforms and network protocols. The reason is that the geospatial web has adopted XML as its standard for data formatting and data management. XML is compatible with many platforms and data sources (Jain & Wu 2007). This thesis investigates the technologies that enable the sharing of epidemiological information via the geospatial web and also the background information relevant to these technologies.

The first objective for Geo-processing for WebEpi was to select a common language capable of expressing geographic information. There are several languages available for geospatial web information management, such as GML, KML and SVG. GML is based on XML. It aims to encode geospatial data and other related data. As long as the non-geospatial data can be embedded together with geospatial data and standardised by GML, all the non-geospatial data can be queried by its geospatial location (Jain & Wu 2007).

GML was developed by the Open Geospatial Consortium. KML is based on GML by the addition of a geospatial information tag (Du, Yu & Liu 2009). As a

web standard of graphic visualisation, SVG has been widely used for web mapping in numerous applications. (Dong & Li 2009).

KML was developed by Google Inc. The syntax of KML is very similar to that of GML and XML. KML was developed to support geospatial visualisation of Google Earth and Google Maps. However, Google Earth and Google Maps are also compatible with other data formats such as GML and SVG. They have a common background, in that they are all based on XML. The DHHS require the clustering analysis results to be presented on Google Maps, so KML was the preferred format for geospatial web information.

The second main objective of building the WebEpi architecture was to enable the visualisation of the epidemiological data analysis on the geospatial web. Smart spatial information extraction and retrieval in repositories of electronic documents are the essential goals for the geospatial web. The proposed WebEpi framework explores ways to effectively extract patterns using clustering analysis techniques and to present the clustering results using graphical representations for health decision making support.

The geospatial web framework includes spatial data analysis, data extraction and data mining. These components can be implemented by the application of suitable interfaces or customised program development. Geospatial applications enable users to explore maps of the datasets from different points of view and also to visualise geospatially related data on selected attributes of the epidemiology data. The Geo-processing for WebEpi is illustrated in Fig 3.12. After clustering analysis, the clustering results are stored in a data repository.

Once a web user selects the epidemiology data attributes and sends the clustering result mapping request, this request is passed on to the web server. The web server analyses the current mapping request and sends a data extraction request to the data repository. The data repository responds by sending a data report back to the web server. The web server subsequently processes the data and sends a further mapping request to the map server. Once the map server gets the request, it sends the map back to the web server to fulfil the original spatial data request as shown in Fig 3.12. The details of the WebEpi Geo-processing are presented in Chapter 6.

3.4 Summary

In this chapter, first the feasibility study has proved that Google Maps can be used as a web service to generate the geospatial visualisation of epidemiological data analysis. Based on the study, the WebEpi system architecture was designed to satisfy DHHS requirements. It consists of three major processes: epidemiological data pre-processing, epidemiological data clustering analysis and Geo-processing of epidemiology data analysis results. The performance comparison of clustering analysis algorithms will be addressed in Chapter 4, while the detailed explanation of the epidemiological data pre-processing algorithm will be documented in Chapter 5.

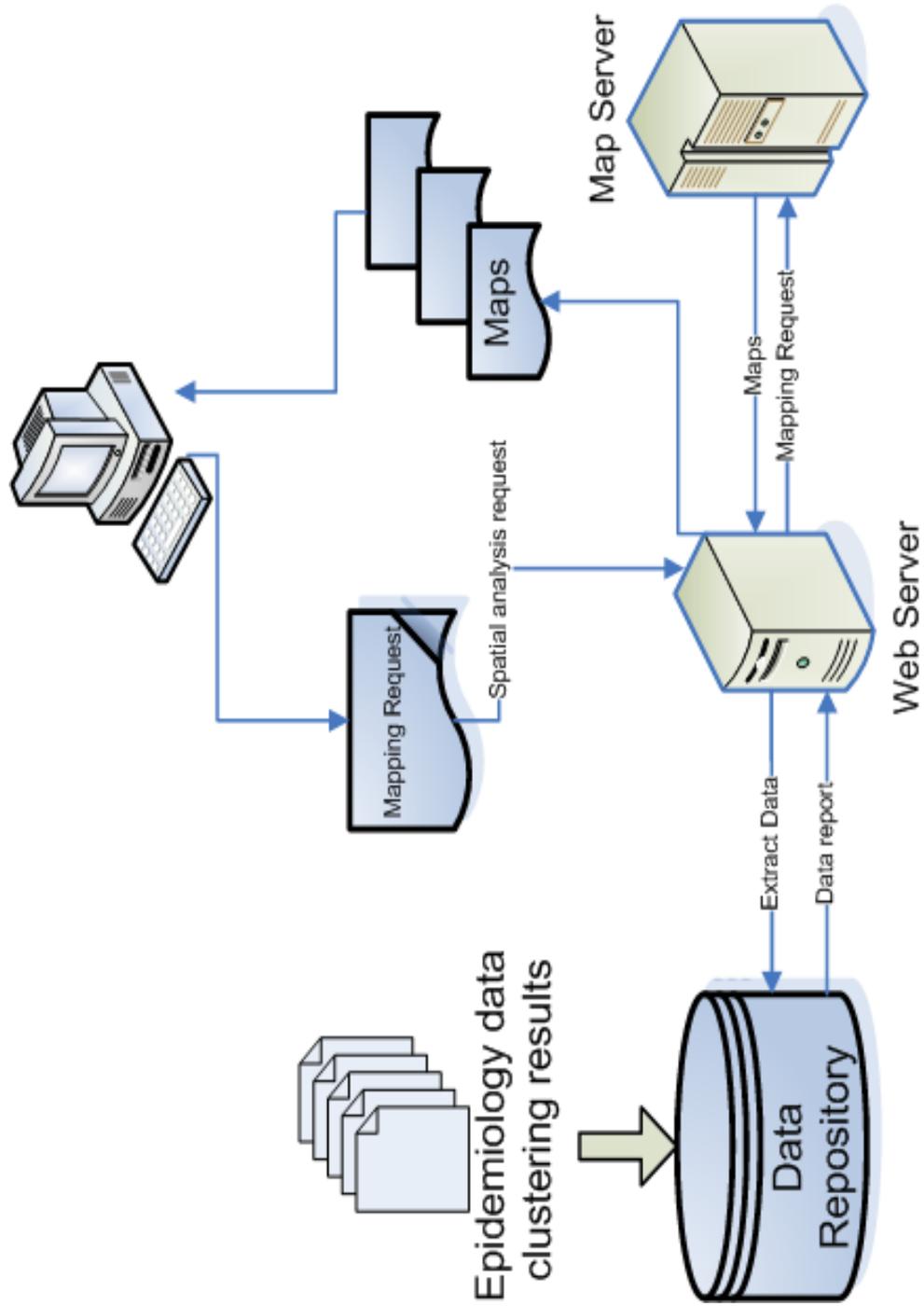


Fig 3.12 Geo-processing

Chapter 4 Clustering Analysis

4.1 Introduction

In this chapter, the general clustering analysis process is described. Section 4.2 introduces the principle of clustering analysis. Section 4.3 presents the reasons and aims for carrying out epidemiological data analysis. As part of epidemiological data analysis, DHHS epidemiological data clustering analysis criteria are elaborated in Section 4.4. Three clustering algorithms, SOMs, FCM and k-means are reviewed. Then epidemiology data clustering analysis is conducted using these three clustering algorithms and described in Sections 4.5, 4.6 and 4.7, respectively.

4.2 Clustering Analysis

Cluster analysis is essentially about discovering clusters in data. Clustering methods should not be confused with assignment methods, which are called supervised learning in artificial intelligence. The aim of supervised learning is to construct rules for classifying new individuals into one group or another, which are known before the classification (Everitt, Landau & Leese 2001). Clustering is a process of classifying a set of objects into different groups by their similarity attributes. In multi-dimensional feature space, the objects are treated as points which are allocated in the space. Objects that have similar attributes are grouped into the same cluster. Therefore, objects within the same cluster

should have high similarity. On the other hand, objects in different clusters should have high dissimilarity. Better cluster results should have high similarity within the cluster and low similarity outside the cluster. However, the definition of cluster is very general. Actually, there are many kinds of clusters, such as spherical clusters, liner clusters, drawn-out clusters and so on (Lai et al. 2007). Different kinds cluster methodologies are suitable for different data requests. Some cluster tasks need to find out similarities, some of them might request to find out dissimilarities (Lai et al. 2007).

Clustering analysis attempts to identify clusters within the presented data. The process of clustering analysis can be described as the discovery of how 'close' individuals are to each other, or how far apart they are. Many clustering investigations have as a starting point an $n * n$ one-mode matrix. It is used to describe the relationship between the objects. The word "Proximities" has been introduced to explain the values in the matrix. There are two ways to find out the proximities, they are similarity and dissimilarity. Similarity is used to measure how close objects are to each other. Dissimilarity is used to measure how far apart objects are from each other. Therefore, proximities can produce two different measurement results (Everitt, Landau & Leese 2001).

Cluster analysis techniques are used to assess whether or not data can be summarised meaningfully in terms of a relatively small number of clusters. The cluster contains objects which are similar to each other. The objects in one cluster are different in some aspects from the objects in other clusters. A vast variety of clustering methods have been developed over the decades (Everitt, Landau & Leese 2001).

However, how to perform a large number of data clustering analyses using complex clustering algorithm formulae becomes a challenge. Developing an automatic data clustering process will eventually solve the problem. Automatic data clustering is still a novel scientific discipline in the geospatially based analysis of epidemiological data. Since 1990, automatic classification has established itself as an independent scientific discipline (Rousseeuw 1990). Automatic data clustering improves clustering performance in both time and accuracy.

4.3 Epidemiological Data Analysis

The DHHS manage information about the health data of all residents in all the LGAs (Zhang, Shi & Zhang 2009). The existing methodology for health data analysis can cope with general disease outbreaks. In the event of infection, a large number of health factors have to be included in the health management research. The DHHS are responsible for the state wide public health management which includes health resource management and government decision making support (Zhang, Shi & Zhang 2009). Epidemiological study, as an area of public health research, plays a significant role in public health surveillance. The epidemiological data is presented by geographic boundaries such as LGA and health service areas. Epidemiological data is useful in monitoring epidemic disease trends, allocating public health resources and also generating hypothesis reports.

The management of state wide health data needs a universal strategy. It should include health emergency services, health source management, public

residence health services and health decision support (Zhang, Shi & Zhang 2009). The prediction of the potential effects of disease spread is still a challenge in health research. It is important information for health policy decision making support (Zhang, Shi & Zhang 2009). It is possible to summarise historical epidemiological data and to analyse it by considering other attributes such as air quality, age group of the population and community health services.

4.4 Epidemiological Data Clustering

Clustering analysis of epidemiological data is one of the most important components in an epidemiological study. Epidemiological data clustering can include many attributes for the clustering process. The aim of clustering analysis is to find data patterns for the epidemiological data and the relationships between the attributes, in order to support health decision making. As described earlier, clustering automation can improve the clustering performance. Therefore, the development of epidemiological data clustering automation will help health researchers to produce more effective analysis reports. Before the development of epidemiological data clustering automation, the algorithms for clustering had to be selected first. Different research needs different clustering algorithms because every clustering algorithm has its strengths and weaknesses. In this research project, WebEpi epidemiological data clustering follows DHHS clustering criteria (Shi et al. 2007).

Before clustering algorithm selection, it was necessary to evaluate the clustering results for the epidemiological data and identify the clustering criteria.

There are two criteria for measuring the epidemiological data proximities. Firstly, the nature of the epidemiological data should strongly influence the choice of the cluster measure. Secondly, the choice of the measurement should depend on the scale of the data. The closer the objects are together, the larger the proximity similarities become. Finally, the clustering evaluation algorithm should be chosen (Everitt, Landau & Leese 2001).

In Chapter 3, the clustering analysis of epidemiological data was described, three clustering algorithms were selected for the epidemiological clustering process, SOMs, FCM and k-means. All of them belong to partitioned clustering methods. They conduct the learning process iteratively. After each iteration, the results are projected in n-dimensional space. Within the space, the distances of objects within same cluster can be summed. As they are partitioned clustering methods, the distance measurement function is already included in the clustering process. The second advantage of partitioned clustering methods is that they are sensitive to input objects. Objects can be dynamically assigned to different clusters to reduce the times of distance calculations. This advantage also improves the efficiency of clustering (Wilson, Boots & Millward 2002). The reason for selecting SOMs, FCM and k-means for investigation is that they all meet the DHHS epidemiological data clustering criteria.

4.5 SOM Clustering Analysis for Epidemiological Data

A SOM is an unsupervised learning algorithm. It consists of several processes such as initialisation, training, visualisation and analysis. SOMs consist of neurons organised on a regular low-dimensional grid (Kohonen 1990) and fit

into the neural network class of methodologies. Some are tolerant of non-normally distributed data. Therefore, the multidimensional epidemiological dataset can be processed to separate the dataset into clusters (Basara & Yuan 2008).

4.5.1 SOM clustering algorithm

A SOM is one of the classic spatial clustering algorithms (Aksoy 2006). It is an unsupervised algorithm which achieves a mapping from a higher dimensional input space to a lower dimensional map space. SOMs provide a compact representation of the data distribution, and have been widely applied in the visualisation of high dimensional data (Logeswari & Karnan 2010).

The SOM competitive learning algorithm can be summarised into four steps as shown in Fig 4.1. The first step is to initialise the weight values by random selection, the interval of the weights is from 0 to 1, and to assign a positive value for the learning rate parameter. The second step is activation and similarity matching. Activation refers to activating the SOM network by applying the input vector x . Similarity matching is finding the winner by taking the best matching neuron at each iteration. The method of finding the best matching neuron is the minimum-distance criterion. The third step is learning, update the weight value by the competitive learning rule. All the neurons located inside the topological neighbourhood are activated simultaneously. The relationship among these neurons is independent of their distances from the winners. The last step is iteration. The iteration counter is increased by one and the process then goes back to step two and continues the iteration until the shortest

distance is not changed or changes less than 0.001, then the whole process is finished (Negnevitsky 2002).

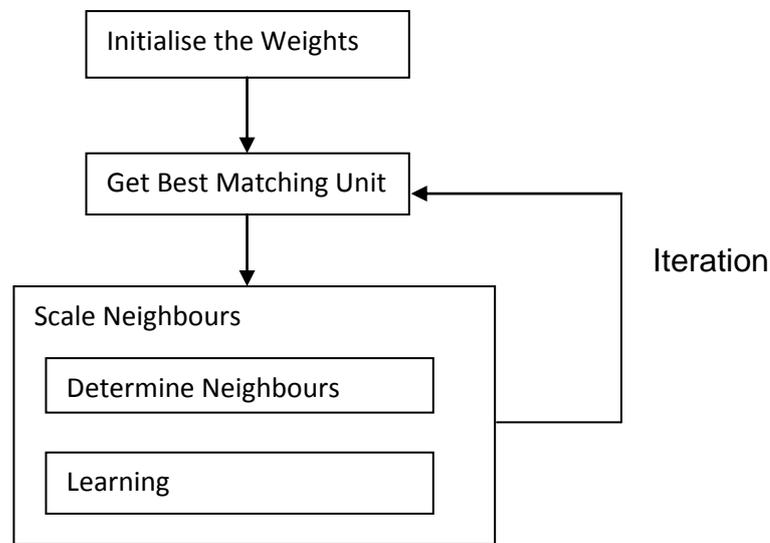


Fig 4.1 SOM learning steps

In each training iteration process, the input vector x is chosen randomly, and the distances between the input vector and the weight vector are calculated. After the iteration training process, the neuron closest to the input vector x and its weight vector are selected. The winner (the closest one) of the training results is denoted by $\|x - m_c\|$. $\{\|x - m_i\|\}$ describes the distance measure, and indicates the training iteration count (Zhang, Shi & Zhang 2009).

$$\|x - m_c\| = \min\{\|x - m_i\|\} \quad (\text{Zhang, Shi \& Zhang 2009}) \quad (\text{Equ 4.1})$$

After discovering the winner neuron and its weight vector, the weight vector is updated by moving it closer to the input vector to optimise the presentation of the input vector. The winning weight vector is updated according to the following algorithm (Zhang, Shi & Zhang 2009).

$$m_i(t + 1) = m_i(t) + \alpha(t)h_{ci}(t)[x(t) - m_i(t)] \quad (\text{Zhang, Shi \& Zhang 2009})(\text{Equ 4.2})$$

Where

t denotes iteration training count,

i is the input dataset,

$x(t)$ is an input vector randomly drawn from the input dataset,

$h_{ci}(t)$ represents the neighbourhood kernel around the winning neuron unit c

$\alpha(t)$ is the learning rate

MATLAB (MathWorks 2010) provides the SOM function to classify input objects. Objects are loaded into a topology map and retain their physical location. There are three functions which are used as object allocation functions; they are grid top, hex top and rand top. Objects can be distributed into a grid, hexagonal or random topology map. Distances of the objects are calculated using the distance function. The SOM in MATLAB includes four distance functions, they are dist, boxdist, linkdist and mandist. The neuron of the topology map which has smallest distance is selected as the winner. The neurons around winner will also be moved closer to the winner (MathWorks 2010).

The final results of the SOM are presented as a map. Similar objects are located on the same point or points near to each other. The map is presented as a rectangular or hexagonal layout (Dykes, MacEachren & Kraak 2005). A set of objects is allocated to individual clusters. SOMs are able to locate one or more objects onto same neuron or its neighbouring neurons inside the cluster. This is part of the training process of the SOM (Dykes, MacEachren & Kraak

2005). SOMs and spatial analysis have been integrated in some geographical information systems to provide a novel approach for the analysis of health reports from different point of views (Zhang, Shi & Zhang 2009).

4.5.2 SOM cluster analysis for WebEpi data

WebEpi aims to combine epidemiological data and geospatial information. The geo-visualisation of WebEpi involves clustering data with similar feature values. The SOM sequential training process is conducted iteratively until the distance between the neuron weight vectors and the input vector cannot be minimised further (Zhang, Shi & Zhang 2009). The SOM analyses the WebEpi epidemiology data, which comes as four groups: Cancer Incidence, Death, Hospitalisation and Notified Infectious. The data is collated annually. One set of data contains three groups: Males, Females and Persons. Each group has different categories of disease. For example, there are seven categories of Notified Infectious and thirteen categories of Cancer Incidence. Each category has attributes such as disease name, LGA information, death figure, expected death rate and SMR. There are twenty-nine LGAs in Tasmania (Zhang, Shi & Zhang 2009).

After processing using the SOM clustering method, input vectors with similar values are mapped on to the same neuron or nearby neurons. The mapping results of the input vectors represents their attribute relationships. The projection of high dimensional data space into low dimensional space reflects the input vector's topological relationships (Zhang, Shi & Zhang 2009). The SOM visualisation for epidemiological data is shown in Fig 4.2

```

function create_c_net(inputs)
% create SOM clustering function for EPI data

dimension1 = 1; % Adjust as desired
dimension2 = 5; % Adjust as desired

% cluster EPI data into 5 clusters

inputs=inputs'
% change inputs data dimension from 92*1 to 1*29 for SOM

net = newsom(inputs,[dimension1 dimension2]);
% create SOM network for inputs

[net, tr] = train(net, inputs); %input data to training network

[outputs, x, y] = sim(net, inputs); % get SOM clustering result

class1=outputs(1,:); % get dataset in cluster 1
class2=outputs(2,:); % get dataset in cluster 2
class3=outputs(3,:); % get dataset in cluster 2
class4=outputs(4,:); % get dataset in cluster 2
class5=outputs(5,:); % get dataset in cluster 2

c11=logical(class1); %convert cluster to logical array
c12=logical(class2); %convert cluster to logical array
c13=logical(class3); %convert cluster to logical array
c14=logical(class4); %convert cluster to logical array
c15=logical(class5); %convert cluster to logical array

% Plot
plot(inputs(c11),'marker','*','color','r'); % plot cluster
hold on;
plot(inputs(c12),'marker','*','color','r'); % plot cluster
plot(inputs(c13),'marker','*','color','r'); % plot cluster
plot(inputs(c14),'marker','*','color','r'); % plot cluster
plot(inputs(c15),'marker','*','color','r'); % plot cluster
hold on

hold off

```

Fig 4.2 SOM visualisation for epidemiological data

4.6 FCM Clustering Analysis for Epidemiological Data

Clustering has been considered as a common issue by many researchers.

Clustering algorithms aim to partition data sets into different clusters by their similarity. Data items which have more similarity are grouped into one cluster

(Wang et al. 2007). A large number of clustering algorithms have been proposed for various applications. They may be roughly categorised into two classes: hard and fuzzy clustering (Wang et al. 2007). During the fuzzy clustering process, one object might not entirely belong to one cluster, it might also partially belong to another cluster. Clustering plays an important role in data analysis results. According to this characteristic, FCM has a greater possibility of clustering (Hoppner, Klawonn & Kruse 1999). Many fuzzy clustering algorithms are available. Until now, the FCM algorithm has been applied in many areas such as image processing, pattern recognition and feature analysis. FCM clustering algorithm is considered as one of the most useful ones. (Wang et al. 2006).

4.6.1 FCM algorithm

FCM aims to analyse numerical data and cluster it into groups. The word “fuzzy” means uncertainty or not precise (Rong & Fan 2009). The input objects in FCM are treated as fuzzy data. That is the reason why this algorithm is called FCM. The advantage of fuzzy data is that the probability of finding correct clusters is increased, because one object can be included in more than one cluster (Rong & Fan 2009).

In the FCM clustering algorithm, the distance function is utilised to allocated objects into one or more clusters (Balaji & Zacharias 2007)

$$J_q(U, V) = \sum_{k=1}^n \sum_{i=1}^c (u_{ik})^q d^2(x_k, v_i) \quad (\text{Huang et al. 2009}) \text{ (Equ 4.3)}$$

where $X = \{x_1, x_2, \dots, x_n\} \in R^p$, n is the total number of objects, c is the total number of clusters, u_{ik} is the degree of relationship of object x_k in the cluster i^{th} , J_q is the fuzzy membership, q is the exponent of the membership. v_i is the centre of cluster i , $d^2(x_k, v_i)$ is a distance measure function, x_k is the object and v_i is the cluster. All the steps of the algorithm can be explained as an iteration process, which is shown below (Huang et al. 2009):

- 1) Initialise c, q and ε .
- 2) Initialise matrix U for fuzzy partition.
- 3) Initialise loop counter $b = 0$.
- 4) Calculate the c cluster centres $\{v_i^b\}$ with U^b :

$$v_i^b = \frac{\sum_{k=1}^n (u_{ik}^b)^q x_k}{\sum_{k=1}^n (u_{ik}^b)^q} \quad (\text{Huang et al. 2009}) \quad (\text{Equ 4.4})$$

Using FCM cluster for Histogram-Based Colour Image Segmentation

- 5) Calculate Matrix U for fuzzy membership, initialise $k = 1$ iterate Matrix calculation from 1 to n , calculate the following:
 - $I_k = \{i | 1 \leq i \leq c, d_{ik} = \|x_k - v_i\| = 0\}$ (Huang et al. 2009)
 - $\tilde{I}_k = \{1, 2, \dots, c\} - I_k$ (Huang et al. 2009)
 - Calculate membership values for the k^{th} column of the matrix:
 - if $I_k \neq \emptyset$, then
 - $u_{ik}^{b-1} = \frac{1}{\sum_{j=1}^c \left(\frac{d_{ik}}{d_{jk}}\right)^{2/(q-1)}}$ (Huang et al. 2009) (Equ 4.5)
 - else $u_{ik}^{b-1} = 0$ for all $i \in I_k$ and $\sum_{i \in I_k} u_{ik}^{b-1} = 1$; next k (Huang et al. 2009)
- 6) If $\|U^b - U^{b-1}\| < \varepsilon$ terminate, else, set $b = b + 1$ and then go back to step d (Huang et al. 2009).

4.6.2 FCM cluster analysis for WebEpi data

FCM was used to conduct the same data experiment as the SOM. First, the program loaded epidemiology data, and classified the data by gender and disease. Then the FCM clustering algorithm was executed for the classified data. The data was clustered into five groups. For each cluster, the minimum and maximum values were calculated. The last step was to plot clusters in five different colours. Fig 4.3 describes the program used to plot epidemiological data FCM clustering results. The plotting process includes getting the threshold for each cluster and colouring cluster elements. The reason for plotting the results is to allow comparison between the SOM and FCM clustering results.

4.7 K-means Clustering Analysis for Epidemiological Data

The k-means algorithm has the advantages of efficient calculation and ease of application. It has been used in cluster analysis for different kinds of data, such as text and images. The algorithm does, however, tend to terminate the iterative process too quickly resulting in only partially optimal results being returned. This feature of the k-means algorithm can result in clustering result fluctuations because of the random selection of the initial iterative centre points. Because of this fact, users cannot always judge the quality of data clustering results, as the clustering results are not consistent. It is very important for multidimensional data to increase the clustering results stability and decrease the calculation time (Li 2010).

```

function [ center,U,obj_fcn ] = cmeans( data )
% create cmeans plotting function for EPI data

[center,U,obj_fcn] = fcm(data,5);
%cluster data into 5 clusters

maxU = max(U);
% Get the data points with highest grade of membership

index1 = find(U(1,:) == maxU);
% Find the data points with highest grade of membership in cluster 1

index2 = find(U(2,:) == maxU);
% Find the data points with highest grade of membership in cluster 2

index3 = find(U(3,:) == maxU);
% Find the data points with highest grade of membership in cluster 3

index4 = find(U(4,:) == maxU);
% Find the data points with highest grade of membership in cluster 4

index5 = find(U(5,:) == maxU);
% Find the data points with highest grade of membership in cluster 5

%plot the 5 clusters dataset into one figure
plot(data(index1,1),'marker','*','color','b');
hold on;
plot(data(index2,1),'marker','*','color','b');
plot(data(index3,1),'marker','*','color','b');
plot(data(index4,1),'marker','*','color','b');
plot(data(index5,1),'marker','*','color','b');
% Plot the cluster centers
plot(center,'*','color','k')
hold off;

```

Fig 4.3 Program code of WebEpi FCM plot

4.7.1 K-means clustering algorithm

The k-means methodology is a commonly used clustering technique. K-means has been used as an unsupervised learning algorithm in many areas of data clustering. This algorithm was developed by MacQueen in 1967 (MacQueen 1967). He defined his algorithm as the k-means. The whole idea is about calculating cluster centres and moving objects (Aksoy 2006):

- Before running the algorithm, the number of clusters has to be set as k, Objects are distributed into k clusters (Aksoy 2006).

- Move objects to the closet cluster centre, and then recalculate the cluster centre for the new objects within the cluster (Aksoy 2006).
- Back to the second step until the movement of cluster centres do not move or they reach their minimal values.

In this clustering analysis, processing starts with the initialisation of the total number of clusters which is K . K has to be fixed from the start to the end. According to the object locations, initial centres are set for each cluster. The target of k-means is to update centres by distance calculation. After the centres are renewed, the objects inside clusters might have to move out or stay in. Then the process goes back to cluster centre distance calculation. The distance is the mean value of objects to their cluster centre (Lai et al. 2007; Vijayabhanu & Radha 2010).

The k-means clustering algorithm is described by *Equ 4.6* below: n is the input object expressed as a vector. c is the value of total number of clusters. The cluster centres are calculated by distance function which was originally developed from the Euclidean distance (Liu et al. 2009),

$$J(u, x) = \sum_{i=1}^c J_i = \sum_{i=1}^c \sum_{j=1, i \neq j}^n (u_{ij}) \|x_j - c_i\| \quad (\text{Liu et al. 2009}) \text{ (Equ 4.6)}$$

where u_{ij} is the degree value for vector x_j . If x_j belongs to cluster $G_i (i = 1, 2, \dots, c)$ then $u_{ij} = 1$ otherwise $u_{ij} = 0$. n objects are distributed into different clusters according to their distance to the clusters. The algorithm is calculated as shown below: (Liu et al. 2009)

$$u_{ij} = \begin{cases} 1, & \text{if } k \neq i \text{ and } \|x_j - c_i\| \neq \|x_j - c_k\| \\ 0, & \text{other} \end{cases} \quad (\text{Liu et al. 2009})$$

$x_j - c_i$ aims to find out the shortest distance to the cluster centre. If c_i is the winner, x_j is assigned to cluster i . Each object can be assigned to only one cluster, these criteria are explained in the Equations 4.6 to 4.10 (Liu et al. 2009).

$$\sum_{i=1}^c u_{ij} = 1, (\forall j = 1, 2, \dots, n) \quad (\text{Liu et al. 2009}) \text{ (Equ 4.7)}$$

and

$$\sum_{i=1}^c \sum_{j=1, i \neq j}^n u_{ij} = n \quad (\text{Liu et al. 2009}) \text{ (Equ 4.8)}$$

if u_{ij} has not been updated during the calculation, the mean value of the objects in cluster i is the minimum value of the iteration function, this is

$$c_i = \frac{1}{|G_i|} \sum_{k, x_k \in G_i} x_k \quad (\text{Liu et al. 2009}) \text{ (Equ 4.9)}$$

where

$$|G_i| = \sum_{j=1}^n u_{ij} \quad (\text{Liu et al. 2009}) \text{ (Equ 4.10)}$$

is the modulus of G_i . It is obviously that clustering process is an iteration process with conditions (Liu et al. 2009).

4.7.2 K-means cluster analysis for WebEpi data

One set of epidemiological data was chosen for running the k-means analysis. The data was selected from epidemiology data for different genders and disease categories. The 2005 epidemiological data, based on the 29 LGAs of

Tasmania, was used as the WebEpi k-means experimental data. The epidemiological data includes area, disease categories, gender, ratio value for incidence and other information. The data was classified into 5 clusters. The WebEpi k-means clustering algorithm processes in 4 steps. Suppose there are N instances, where $N=29$ which are to be classified into K clusters where $K=5$.

Firstly, k-means randomly selects k representatives as the initial cluster centres. Secondly, it starts the iteration. It assigns cluster centres that are close to rest of the objects. The distances are calculated from the distance between cluster centres and objects (Lai et al. 2007) as shown in Equation 4.11

$$\min \left\{ \sum_{k=1}^K \sum_{n=1}^N [\delta(x_n, c_k) \|x_n - c_k\|^2] \right\} \quad (\text{Lai et al. 2007}) \text{ (Equ 4.11)}$$

Where $k = 1, 2, 3, \dots, K$ and $n = 1, 2, 3, \dots, N$. $\delta(x_n, c_k)$ means to locate instance n in the k cluster.

$\|x_n - c_k\|^2$ means to calculate the distance between the instance n and k cluster centre point. Thirdly, k-means re-calculates the centre for each cluster as shown in Equation 4.12

$$c_k = \frac{\sum_{n=1}^N \delta(x_n, c_k) x_{nk}}{\sum_{n=1}^N \delta(x_n, c_k)} \quad (\text{Lai et al. 2007}) \text{ (Equ 4.12)}$$

where c_k means the cluster k centre point. Lastly, it returns to the membership assignment iteration. The iteration will continue until the membership change rate is less than or equal to the threshold (Lai et al. 2007; Hong & Kwong 2009). The minimum absolute coordinate change for any centre is set as 0.001, and the maximum iteration is set as 100. In other words, when the cluster centre

points are not changed or the number of iteration reaches 100, the program terminates. Otherwise, once all the cluster centre points are set constantly, the iteration is terminated. The final clusters were created by the measurement of distance between the points and the cluster centre points. Fig 4.4 contains the plotting function of the k-means clustering results in MATLAB.

```
function [ IDX, C ] = kmeans_epi( data )
%create kmeans plot function for EPI data

[IDX, C, sumd, D] = kmeans( data, 5 );
% cluster data in to 5 clusters

%plot 5 clusters datasets into one figure
plot( data (IDX==1,1), 'marker', '*', 'color', 'g' );
hold on
plot( data (IDX==2,1), 'marker', '*', 'color', 'g' );
plot( data (IDX==3,1), 'marker', '*', 'color', 'g' );
plot( data (IDX==4,1), 'marker', '*', 'color', 'g' );
plot( data (IDX==5,1), 'marker', '*', 'color', 'g' );

%plot 5 clusters centres
plot( C, '*', 'color', 'k' );

hold off
```

Fig 4.4 WebEpi k-means plot function

4.8 Summary

In this chapter, clustering algorithms SOMs, FCM and k-means were explained. Their clustering processes are similar to each other and can be summarised as four general steps: choose the input vectors, calculate the distance between the vector and the cluster centre, discover the winner neuron and update the centroid, moving it closer to the input vector. SOMs, FCM and k-means clustering plot visualisation functions were also developed and detailed.

Chapter 5 Clustering Experiments

5.1 Introduction

The aim of this chapter is to find out the most suitable algorithm for the clustering analysis of geospatial epidemiological data. A number of epidemiological data clustering experiments were conducted based on WebEpi data from DHHS. First, the WebEpi data was pre-processed and formatted in a standard format for data clustering analysis which is described in Section 5.2. Section 5.3 describes data clustering which was carried out by selected clustering algorithms, SOM, FCM and k-means respectively. Since the SOM, FCM and k-means have similar results, a clustering validation algorithm was required to identify the most suitable clustering algorithm for DHHS WebEpi epidemiological data.

The Davies-Bouldin index clustering validation algorithm is based on interval criteria and is discussed in Section 5.4. This validation algorithm was used to assess the clustering results and find the most suitable clustering algorithm for DHHS WebEpi epidemiological data. The automation of the clustering analysis of epidemiological data for WebEpi is described in Section 5.5.

5.2 Pre-Processing

DHHS epidemiological data has four epidemiological groups: Cancer Incidence, Death, Hospitalisation and Notified infectious. Each LGA's epidemiological data

is saved as an individual Microsoft Excel file. Each epidemiological group's data contains three people groups: Males, Females and Persons. There are different disease categories for these four epidemiological groups and three population groups. Each category has an SMR value for specific conditions for the current category disease as shown in Fig 5.1.

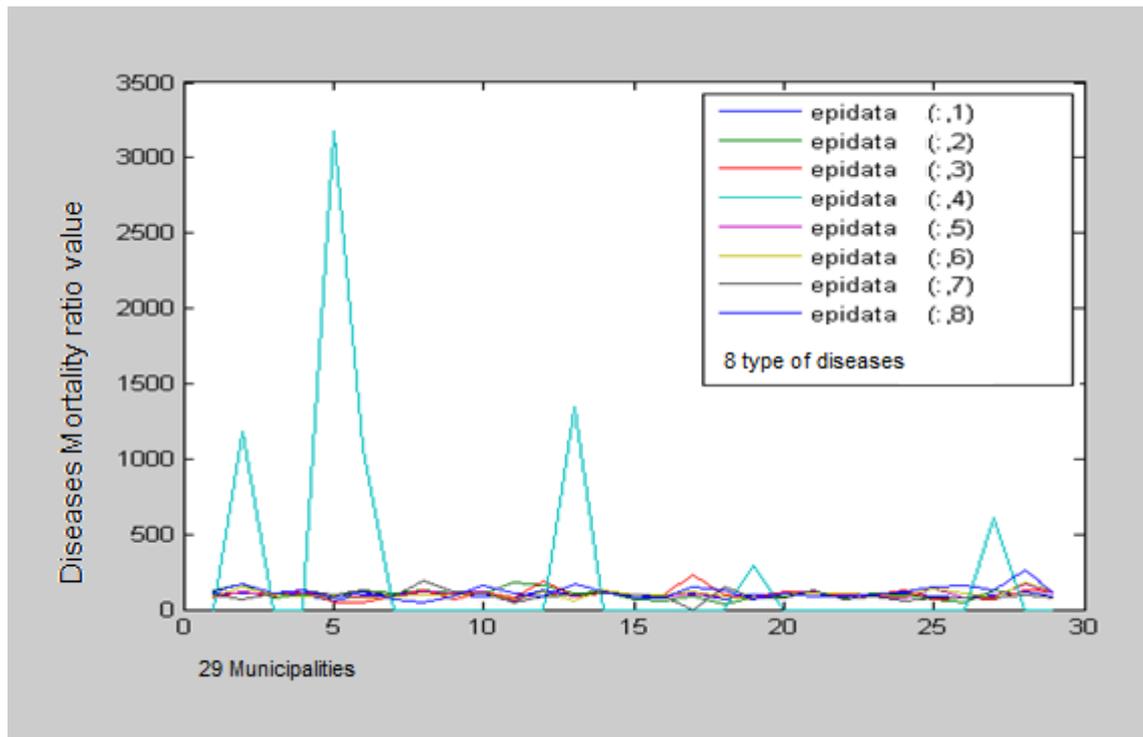


Fig 5.1 Male-Standard Mortality Ratio (SMR) in 2005

The ratio value is assigned to its corresponding LGA. The XML file format was adopted as the formalised epidemiology data format as it is more flexible for data extraction and is also compatible with many geospatial file formats, such as GML, KML, and the ESRI .shp file. The epidemiological data experiments were carried out in a MATLAB workspace.

Because each different epidemiological group contains data for different diseases, MATLAB functions were created for each individual epidemiological

group. The sample MATLAB function code, e.g. *webepi_cancerincidence.m file in the CD-ROM*, for data pre-processing is shown in Fig 5.2. After pre-processing, the epidemiological data was saved in an XML file, and organised by LGA parser and disease parser.

```

function webepi_cancerincidence(epifile)
%create pre-processing function for cancerincidence EPI data

xmlmdir='c:\webepi\cancer_incidence\';
%set source data directory

%load input file from three excel sheets males, females and persons
[epi_males,LGA_males]=xlsread(epifile,'males');
[epi_females,LGA_females]=xlsread(epifile,'females');
[epi_persons,LGA_persons]=xlsread(epifile,'persons');

%load LGA coordinate file
[LGA_num,LGA_GIS]=xlsread('C:\Users\YuanYuan\Documents\MATLAB\LGA.xls','LGA

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Pre-processing data, all types of diseases in Cancerincidence excel file
% are saved on the same sheet, each type of disease SMR has 29 cells
% because there are 29 LGA area. from row 1 to row 29 is cancer_all, row
% 30 to row 58 is Prostate, row 59 to row 87 is colorectal...

function xmlfile=Cancer_all(filename,epi_cancerall,LGA,LGA_GIS)
xmlname=[filename,'\cancer_(all_sites)'];
cancerall=epi_cancerall(1:29);
[IDX]=kmeans(cancerall,5);
xmlfile=Epi_xml(xmlname,IDX,LGA,LGA_GIS);
end

function xmlfile=Prostate(filename,epi_prostatecancer,LGA,LGA_GIS)
xmlname=[filename,'\prostate'];
prostatecancer=epi_prostatecancer(30:58);
[IDX]=kmeans(prostatecancer,5);
xmlfile=Epi_xml(xmlname,IDX,LGA,LGA_GIS);
end

function xmlfile=Colorectal(filename,epi_colorectal,LGA,LGA_GIS)
xmlname=[filename,'\colorectal'];
colorectal=epi_colorectal(59:87);
[IDX]=kmeans(colorectal,5);
xmlfile=Epi_xml(xmlname,IDX,LGA,LGA_GIS);
end

```

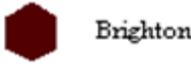
Fig 5.2 MATLAB code of data pre-processing

5.3 Experiment Results

The DHHS require the epidemiological data to be grouped into five clusters. The aim of clustering analysis is to classify epidemiological data into five clusters from low to high value according to DHHS requirements. The classification thresholds are automatically adjusted to accommodate the maximum and minimum of a data set. The cluster number is used to determine the colour coding for each LGA area in the WebEpi geospatial layer file. The experimental data was clustered by SOM, FCM and k-means. The clustering results are visualised by a MATLAB program, the source code is in the CD-ROM experiment folder, in Appendix D.

5.3.1 SOM

The WebEpi SOM clustering method formed the five most significant clusters based on the input vector values with their corresponding geographical information. In Fig 5.3 diagram (a) illustrates the Unified distance matrix (U-matrix) (Zhang, Shi & Zhang 2009). The U-matrix diagram represents the distances between neighbouring neurons, and visualises the cluster structure of the training results (Zhang, Shi & Zhang 2009). The diagrams shown as Fig 5.3(b) to (i) describes the cluster training results of different attributes of the input vectors from Cancer and Injury and Poisoning (Zhang, Shi & Zhang 2009). Fig 5.4 is an enlarged version of Fig 5.3(i) with legends, e.g. it describes the LGA information of Males Injury and the Poisoning SMR in Tasmania in 2005. Fig 5.5 list the colour code for all 29 LGAs of Fig 5.4. For example



represents Brighton which has a numerical value of 151 (Zhang, Shi & Zhang 2009).

The projection of the Male SMR on the SOM output space illustrates the most significant clusters and the least number of neurons clusters. The size of clusters reflects the number of pattern appearances on the SOM (Zhang, Shi & Zhang 2009). The topology of the input vectors in its multi-dimensional input space is mapped on the SOM two dimensional space (Zhang, Shi & Zhang 2009). The input vector describes the relationship between SMRs and their corresponding spatial locations. Multiple views of epidemiology data give better data exploration and interpretation of patterns which help health researchers to discover hidden information (Zhang, Shi & Zhang 2009).

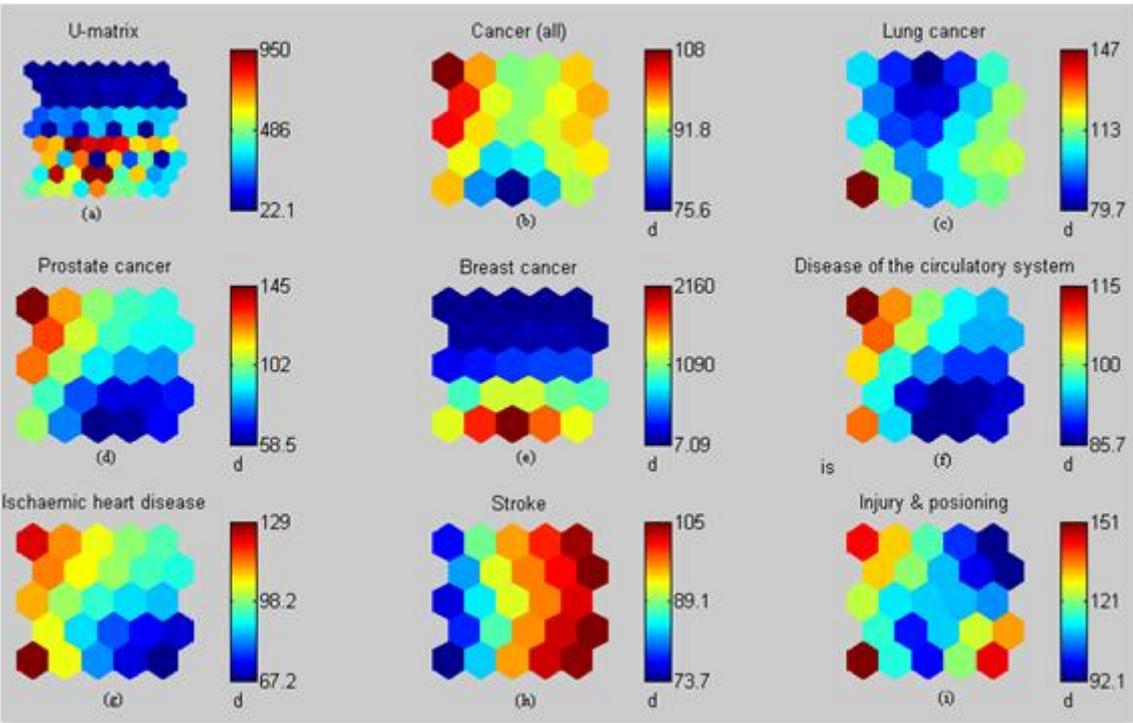


Fig 5.3 SOM training results

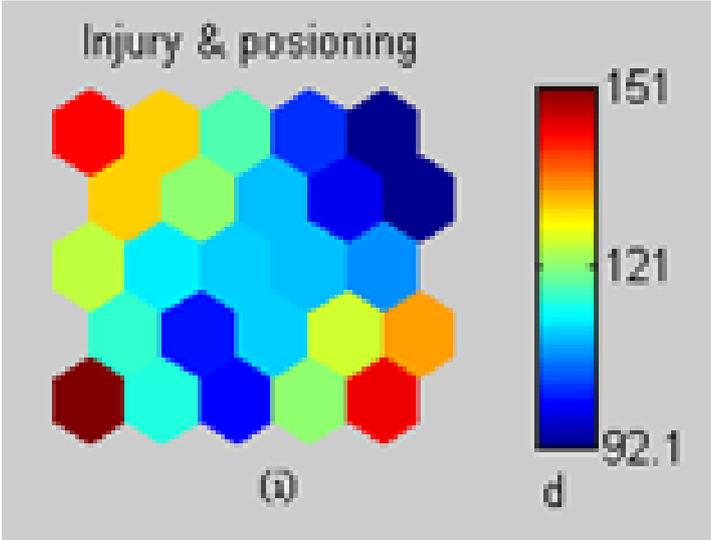


Fig 5.4 SOM injury & poisoning



Fig 5.5 Colour coding for 29 LGAs

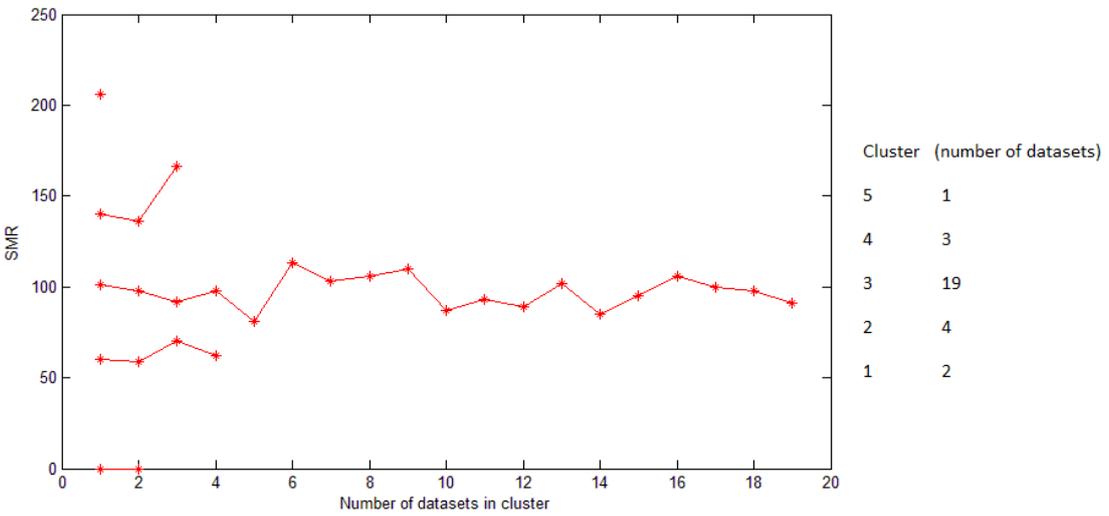
In order to get consistency, the iteration count was set to 100 every time the SOM clustering algorithm was run. The experimental results are shown in Fig 5.6. Fig 5.6(a) shows the SOM clustering analysis output in text format, e.g. (4,1) means LGA1 is grouped in the 4th cluster, (3,2) means LGA 2 is in the 3rd cluster. The first number is the cluster ID, the second number is the LGA code. The first number in the bracket is the cluster index, and the second number is

the input vector. Fig 5.6(b) is the plot of the SOM clustering results, and describes the distance relationship between the centre point and the input vector.

outputs =

(4,1)	(5,11)	(3,21)
(3,2)	(3,12)	(3,22)
(3,3)	(3,13)	(3,23)
(4,4)	(3,14)	(2,24)
(1,5)	(3,15)	(2,25)
(2,6)	(3,16)	(3,26)
(3,7)	(3,17)	(3,27)
(4,8)	(1,18)	(3,28)
(3,9)	(3,19)	(3,29)
(3,10)	(2,20)	

(a) SOM output



(b) Plot of SOM clustering results

Fig 5.6 SOM clustering results

5.3.2 FCM

Experiments were conducted using the FCM clustering algorithm following the same experimental procedure used for the SOM. FCM is capable of producing clustering plotting results at each iteration output. In Fig 5.7(a) *fcn* describes the value of the object function, as described in Chapter 4, during iterations, while in Fig 5.7 (b) *centre* shows the plot result of clustering result. All objects of an individual cluster are joined by a line. There are five lines in the plot. There is a significant difference between the five clusters. In order to compare this result with other clustering results, FCM plots the clustering results on the same scale as other clustering plots as shown in Fig 5.7(d), the black crosses in Fig 5.7(d) are the centres of the clusters.

5.3.3 K-means

The clustering results obtained after applying the k-means algorithm to the epidemiological experimental data are shown in Fig 5.8(a). *sumd* is the initial clustering centre point location and D is the distance from the cluster centre after each iteration. The iteration finishes, when the centre point moves to be the closest location to all points within the cluster. In Fig 5.8(b), C describes the final location of the centre point of the clusters. The plotting results of k -means clustering are shown in Fig 5.8(c). The black crosses are the centre points of the clusters.

```

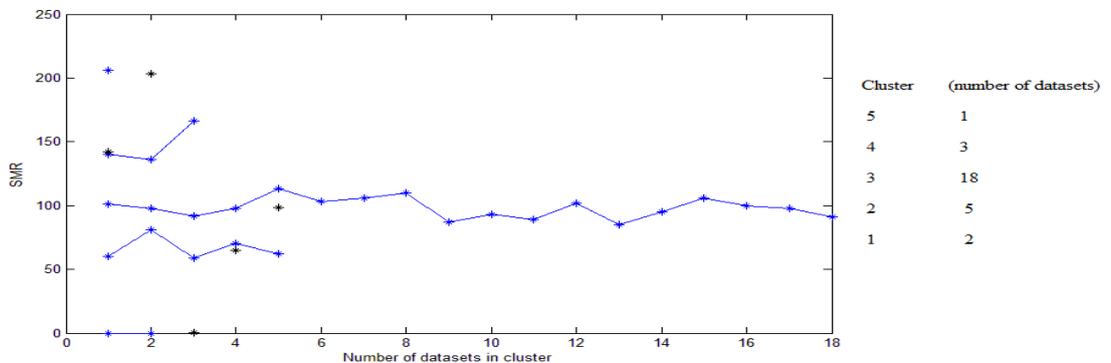
Iteration count = 1, obj. fcn = 11361.214598
Iteration count = 2, obj. fcn = 8310.755057
Iteration count = 3, obj. fcn = 7805.814294
Iteration count = 4, obj. fcn = 7044.308266
Iteration count = 5, obj. fcn = 6341.566656
Iteration count = 6, obj. fcn = 5491.031597
Iteration count = 7, obj. fcn = 5253.673462
Iteration count = 8, obj. fcn = 5085.913201
Iteration count = 9, obj. fcn = 4888.560483
Iteration count = 10, obj. fcn = 4644.033921
Iteration count = 11, obj. fcn = 4242.406557
Iteration count = 12, obj. fcn = 3198.394447
Iteration count = 13, obj. fcn = 2363.331189
Iteration count = 14, obj. fcn = 2141.203002
Iteration count = 15, obj. fcn = 2000.911834
Iteration count = 16, obj. fcn = 1944.889495
Iteration count = 17, obj. fcn = 1914.661431
Iteration count = 18, obj. fcn = 1877.187991
Iteration count = 19, obj. fcn = 1809.467762
Iteration count = 20, obj. fcn = 1689.011140
Iteration count = 22, obj. fcn = 1485.422117
Iteration count = 23, obj. fcn = 1459.090329
Iteration count = 24, obj. fcn = 1451.069659
Iteration count = 25, obj. fcn = 1448.926629
Iteration count = 26, obj. fcn = 1448.374771
Iteration count = 27, obj. fcn = 1448.233397
Iteration count = 28, obj. fcn = 1448.197154
Iteration count = 29, obj. fcn = 1448.187849
Iteration count = 30, obj. fcn = 1448.185456
Iteration count = 31, obj. fcn = 1448.184839
Iteration count = 32, obj. fcn = 1448.184680
Iteration count = 33, obj. fcn = 1448.184639
Iteration count = 34, obj. fcn = 1448.184629
Iteration count = 35, obj. fcn = 1448.184626
    
```

(a) FCM output

<p>Center =</p> <p>141.9209</p> <p>203.4474</p> <p>0.0753</p> <p>64.6548</p> <p>98.1597</p>	<p>U =</p> <table border="0"> <tr> <td>Columns 1 - 4</td> <td>Columns 17 - 20</td> </tr> <tr> <td>0.9836 0.0005 0.9262 0.0815</td> <td>0.6427 0.7061 0.0384 0.0355</td> </tr> <tr> <td>0.0005 0.0002 0.0012 0.0042</td> <td>0.0032 0.0130 0.0028 0.0046</td> </tr> <tr> <td>0.0020 0.0001 0.0075 0.2720</td> <td>0.0267 0.0308 0.7933 0.8684</td> </tr> <tr> <td>0.0104 0.0002 0.0578 0.6271</td> <td>0.3097 0.1074 0.1562 0.0789</td> </tr> <tr> <td>0.0034 0.9990 0.0072 0.0152</td> <td>0.0177 0.1426 0.0092 0.0126</td> </tr> <tr> <td>Columns 5 - 8</td> <td>Columns 21 - 24</td> </tr> <tr> <td>0.0400 0.0000 0.0384 0.8404</td> <td>0.8114 0.0012 0.1223 0.0000</td> </tr> <tr> <td>0.0007 0.0000 0.0028 0.0064</td> <td>0.0024 0.0001 0.0017 0.0000</td> </tr> <tr> <td>0.0109 0.9998 0.7933 0.0191</td> <td>0.0169 0.9949 0.0223 0.9998</td> </tr> <tr> <td>0.9451 0.0001 0.1562 0.0759</td> <td>0.1560 0.0034 0.8454 0.0001</td> </tr> <tr> <td>0.0033 0.0000 0.0092 0.0582</td> <td>0.0135 0.0004 0.0083 0.0000</td> </tr> <tr> <td>Columns 9 - 12</td> <td>Columns 25 - 28</td> </tr> <tr> <td>0.0558 0.0028 0.0028 0.0000</td> <td>0.9853 0.9836 0.0049 0.4455</td> </tr> <tr> <td>0.0038 0.0002 0.0002 1.0000</td> <td>0.0003 0.0005 0.0001 0.0034</td> </tr> <tr> <td>0.6724 0.9875 0.9875 0.0000</td> <td>0.0016 0.0020 0.0019 0.0325</td> </tr> <tr> <td>0.2553 0.0087 0.0087 0.0000</td> <td>0.0109 0.0104 0.9926 0.5008</td> </tr> <tr> <td>0.0127 0.0008 0.0008 0.0000</td> <td>0.0019 0.0034 0.0005 0.0178</td> </tr> <tr> <td>Columns 13 - 16</td> <td>Columns 29</td> </tr> <tr> <td>0.0753 0.0014 0.0154 0.2608</td> <td>0.0734</td> </tr> <tr> <td>0.0120 0.0000 0.0005 0.0028</td> <td>0.0034</td> </tr> <tr> <td>0.7368 0.0008 0.0124 0.0309</td> <td>0.1691</td> </tr> <tr> <td>0.1457 0.9976 0.9698 0.6916</td> <td>0.7416</td> </tr> <tr> <td>0.0302 0.0002 0.0020 0.0139</td> <td>0.0126</td> </tr> </table>	Columns 1 - 4	Columns 17 - 20	0.9836 0.0005 0.9262 0.0815	0.6427 0.7061 0.0384 0.0355	0.0005 0.0002 0.0012 0.0042	0.0032 0.0130 0.0028 0.0046	0.0020 0.0001 0.0075 0.2720	0.0267 0.0308 0.7933 0.8684	0.0104 0.0002 0.0578 0.6271	0.3097 0.1074 0.1562 0.0789	0.0034 0.9990 0.0072 0.0152	0.0177 0.1426 0.0092 0.0126	Columns 5 - 8	Columns 21 - 24	0.0400 0.0000 0.0384 0.8404	0.8114 0.0012 0.1223 0.0000	0.0007 0.0000 0.0028 0.0064	0.0024 0.0001 0.0017 0.0000	0.0109 0.9998 0.7933 0.0191	0.0169 0.9949 0.0223 0.9998	0.9451 0.0001 0.1562 0.0759	0.1560 0.0034 0.8454 0.0001	0.0033 0.0000 0.0092 0.0582	0.0135 0.0004 0.0083 0.0000	Columns 9 - 12	Columns 25 - 28	0.0558 0.0028 0.0028 0.0000	0.9853 0.9836 0.0049 0.4455	0.0038 0.0002 0.0002 1.0000	0.0003 0.0005 0.0001 0.0034	0.6724 0.9875 0.9875 0.0000	0.0016 0.0020 0.0019 0.0325	0.2553 0.0087 0.0087 0.0000	0.0109 0.0104 0.9926 0.5008	0.0127 0.0008 0.0008 0.0000	0.0019 0.0034 0.0005 0.0178	Columns 13 - 16	Columns 29	0.0753 0.0014 0.0154 0.2608	0.0734	0.0120 0.0000 0.0005 0.0028	0.0034	0.7368 0.0008 0.0124 0.0309	0.1691	0.1457 0.9976 0.9698 0.6916	0.7416	0.0302 0.0002 0.0020 0.0139	0.0126
Columns 1 - 4	Columns 17 - 20																																																
0.9836 0.0005 0.9262 0.0815	0.6427 0.7061 0.0384 0.0355																																																
0.0005 0.0002 0.0012 0.0042	0.0032 0.0130 0.0028 0.0046																																																
0.0020 0.0001 0.0075 0.2720	0.0267 0.0308 0.7933 0.8684																																																
0.0104 0.0002 0.0578 0.6271	0.3097 0.1074 0.1562 0.0789																																																
0.0034 0.9990 0.0072 0.0152	0.0177 0.1426 0.0092 0.0126																																																
Columns 5 - 8	Columns 21 - 24																																																
0.0400 0.0000 0.0384 0.8404	0.8114 0.0012 0.1223 0.0000																																																
0.0007 0.0000 0.0028 0.0064	0.0024 0.0001 0.0017 0.0000																																																
0.0109 0.9998 0.7933 0.0191	0.0169 0.9949 0.0223 0.9998																																																
0.9451 0.0001 0.1562 0.0759	0.1560 0.0034 0.8454 0.0001																																																
0.0033 0.0000 0.0092 0.0582	0.0135 0.0004 0.0083 0.0000																																																
Columns 9 - 12	Columns 25 - 28																																																
0.0558 0.0028 0.0028 0.0000	0.9853 0.9836 0.0049 0.4455																																																
0.0038 0.0002 0.0002 1.0000	0.0003 0.0005 0.0001 0.0034																																																
0.6724 0.9875 0.9875 0.0000	0.0016 0.0020 0.0019 0.0325																																																
0.2553 0.0087 0.0087 0.0000	0.0109 0.0104 0.9926 0.5008																																																
0.0127 0.0008 0.0008 0.0000	0.0019 0.0034 0.0005 0.0178																																																
Columns 13 - 16	Columns 29																																																
0.0753 0.0014 0.0154 0.2608	0.0734																																																
0.0120 0.0000 0.0005 0.0028	0.0034																																																
0.7368 0.0008 0.0124 0.0309	0.1691																																																
0.1457 0.9976 0.9698 0.6916	0.7416																																																
0.0302 0.0002 0.0020 0.0139	0.0126																																																

(b) FCM Centres

(c) FCM iteration results



(d) Plot of FCM clustering result

Fig 5.7 FCM clustering results

sumd =	D =	0.5968	1.9600	0.0484	0.1356	0.2588	0.2525	1.2769	0.2401	0.0096	0.0570
		0.1463	1.0201	0.3721	0.0005	0.0141	0.1620	1.0609	0.3481	0.0000	0.0193
0.0747		0.1243	0.9604	0.4096	0.0027	0.0079	0.1871	1.1236	0.3136	0.0008	0.0285
		0.5366	1.8496	0.0676	0.1077	0.2197	0.2233	1.2100	0.2704	0.0046	0.0436
0		0.3938	0	2.6244	1.0646	0.7943	0.0588	0.7569	0.5625	0.0262	0.0005
		0.0008	0.3600	1.0404	0.1865	0.0848	0.0915	0.8649	0.4761	0.0104	0.0015
3.1120		0.0856	0.8464	0.4900	0.0125	0.0008	0.3938	0	2.6244	1.0646	0.7943
		1.0661	2.7556	0.0016	0.3946	0.5910	0.0689	0.7921	0.5329	0.0201	0.0000
0.2556		0.1243	0.9604	0.4096	0.0027	0.0079	0.0014	0.3481	1.0609	0.1952	0.0908
		0.0333	0.6561	0.6561	0.0492	0.0066	0.1541	1.0404	0.3600	0.0001	0.0166
0.1489		2.0521	4.2436	0.1936	1.0572	1.3660	0.0495	0.7225	0.5929	0.0331	0.0017
							0.1040	0.9025	0.4489	0.0067	0.0035

(a) K-means *sumd* and iteration results

C =

62.7500

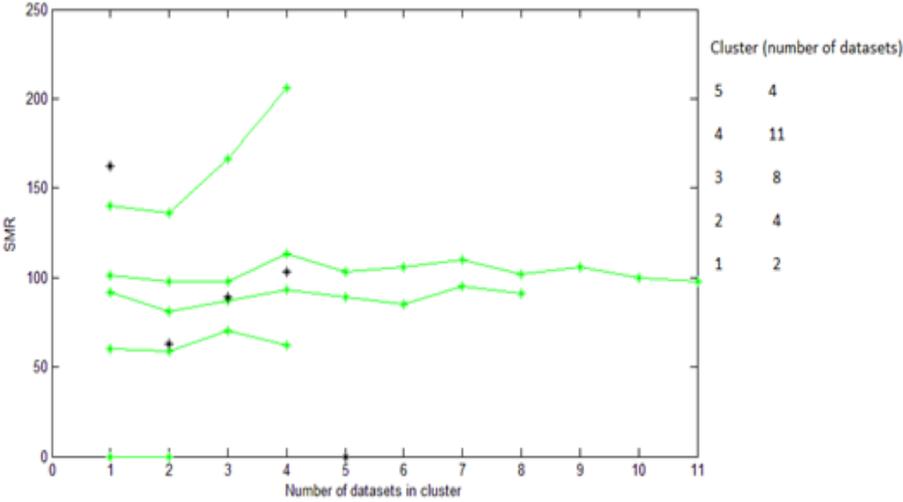
0

162.0000

103.1818

89.1250

(b) K-means centre



(c) Plot of k-means clustering result

Fig 5.8 K-means clustering results

5.4 Experiment Evaluation

The selection of a mathematical clustering algorithm depends both on the information hierarchy and on the aims of the clustering analysis. There is more than one clustering algorithm applicable for clustering analysis. However, clustering selection criteria alone might not be enough for the selection of a single clustering algorithm. In order to choose the most suitable clustering algorithm for DHHS epidemiological data analysis, it was better to test all the preferable algorithms, and use a clustering evaluation algorithm to validate the clustering results. The expression of clustering results had to maintain the features of the original data sufficiently to meet DHHS requirements (Rousseeuw 1990).

For DHHS epidemiological data clustering, three clustering algorithms were chosen: SOMs, FCM and k-means. The flow chart of the epidemiological data clustering process is shown in Fig 3.11. The cluster training process was the core of the clustering analysis. Firstly, the input vector was chosen (Zhang, Shi & Zhang 2009). Secondly, the distances between every input vector and weight vector were calculated (Zhang, Shi & Zhang 2009). After the iteration, the neuron of the training network and its weight vector closest to the input vector were selected. Then the winning neuron and its weight vector were discovered (Zhang, Shi & Zhang 2009). Lastly, the weight vector was updated by moving it closer to the input vector. When all the data had been run through the clustering analysis process, the clustering results were ready for geospatial processing (Zhang, Shi & Zhang 2009).

The experiment data was clustered by each of the SOM, FCM and k-means processes. The clustering analysis examples shown in this chapter describe the clustering results of the SMR for the selected diseases in Tasmania in 2005. During the experiment, sometimes the SOM and FCM could produce different clustering results, and at other times the results could be similar. In order to select the best performing clustering algorithm for epidemiological data clustering, a clustering evaluation algorithm had to be applied. The Davies–Bouldin index is an interval criterion, and was selected for evaluation purposes.

The selection of the evaluation method to be used was based on the interval criteria. Objects within same cluster are close to each other and have more similarities. Objects in different clusters have low similarities. According to DHHS requirements, the closest distance element in clustering represents the highest priority of the clustering results. In order to evaluate the three clustering algorithms, namely SOMs, FCM and k-means, a number of experiments were conducted and the Davies-Bouldin index, used to evaluate the results

Clustering results from the SOM, FCM and k-means are shown in Figs 5.9-5.16 and each figure is produced on two pages. In these figures, r is the value of maximum Davies-Bouldin index for each cluster, and t is the mean value of r . SOM clustering results are shown in red. FCM clustering results are in blue and k-means clustering results are shown in green. The comparison of the Davies-Bouldin validation results is shown in Fig 5.17. The validation output shows that K-Means has the least interval distance, C-Means has the second lowest, and the SOM has the largest interval distance measurement. Therefore K-Means is the best clustering algorithm for the DHHS epidemiological data clustering analysis, and was the chosen algorithm for WebEpi.

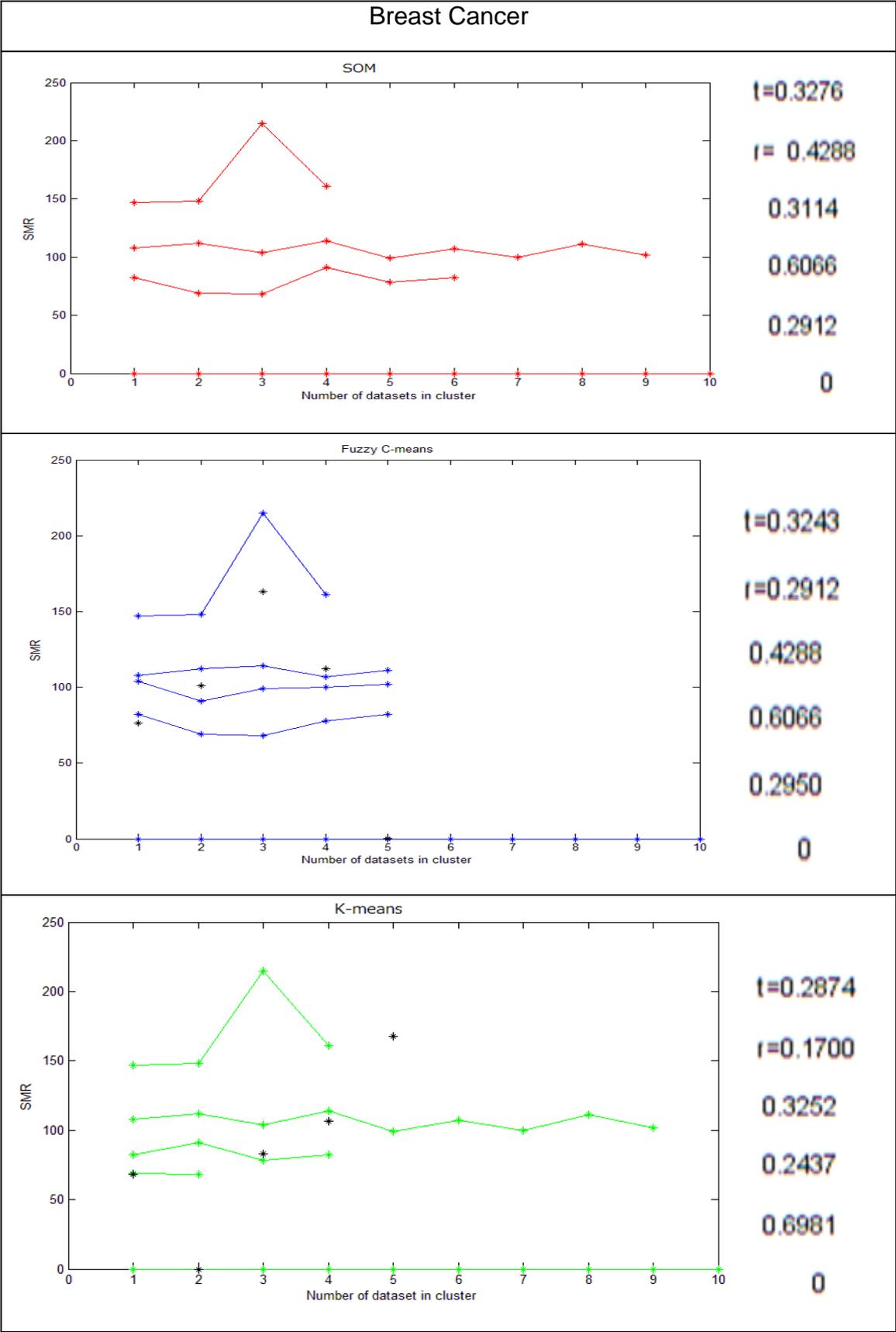


Fig 5.9 Comparison of clustering algorithms for Breast Cancer

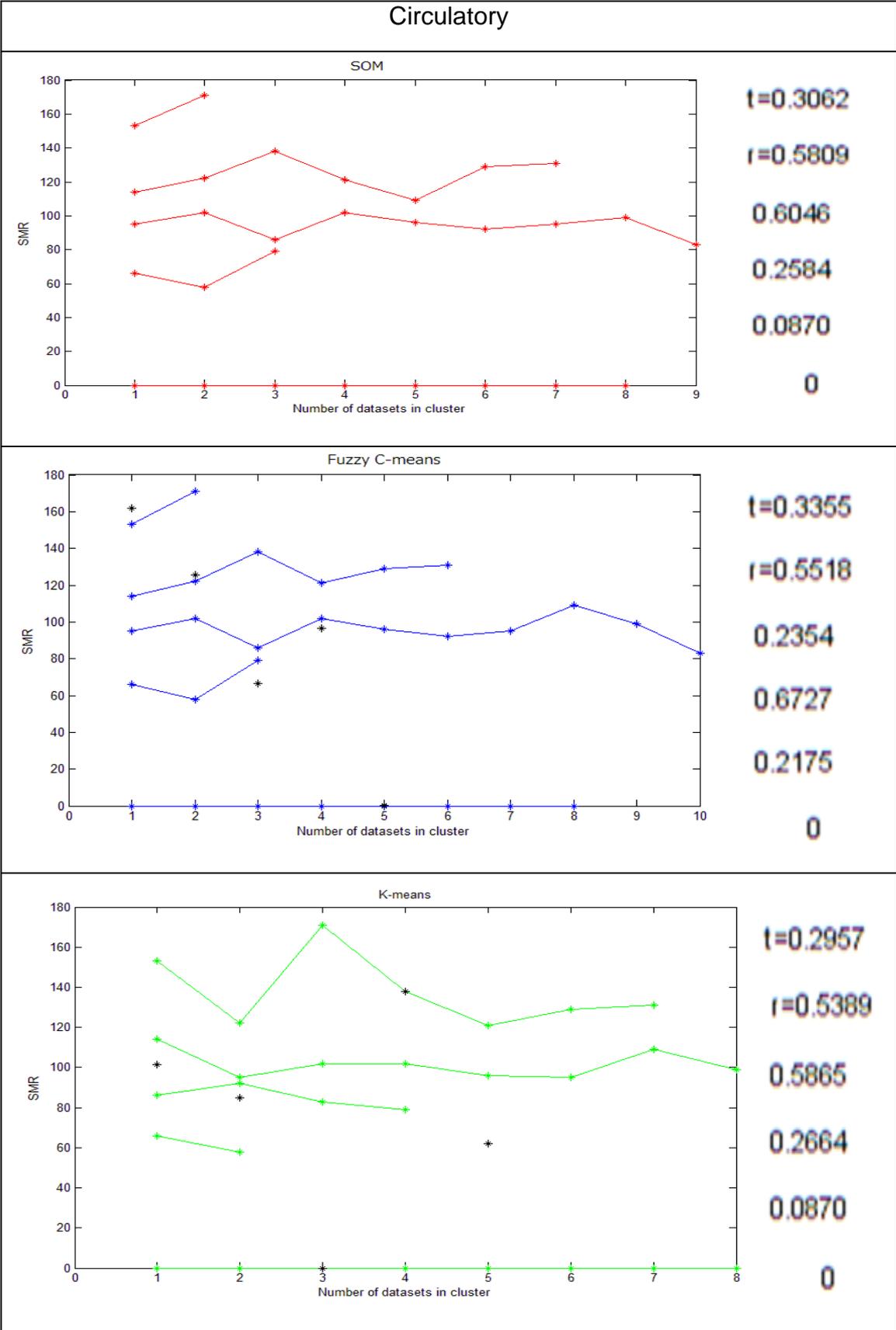


Fig 5.10 Comparison of clustering algorithms for circulatory

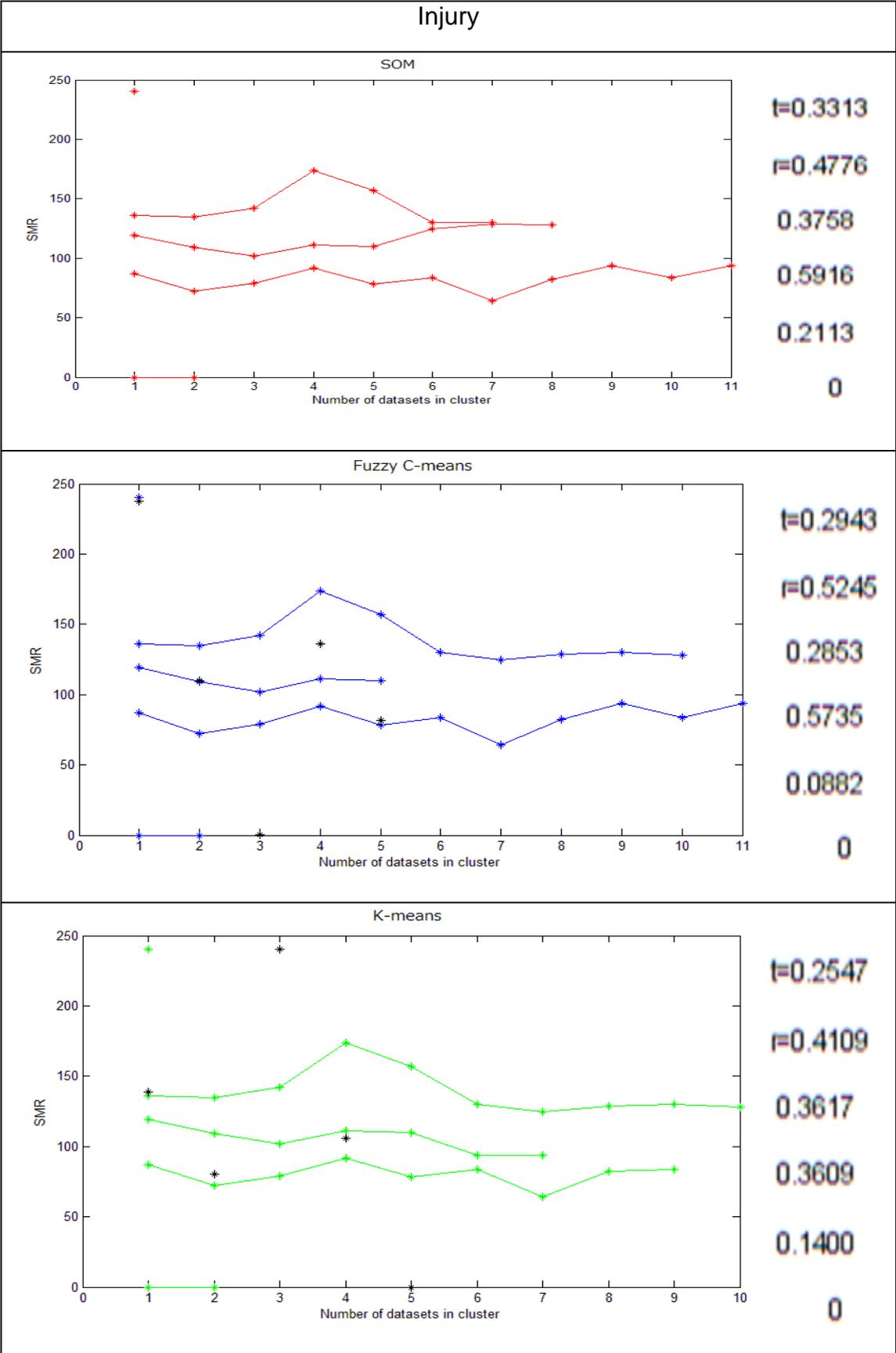


Fig 5.11 Comparison of clustering algorithms for injury

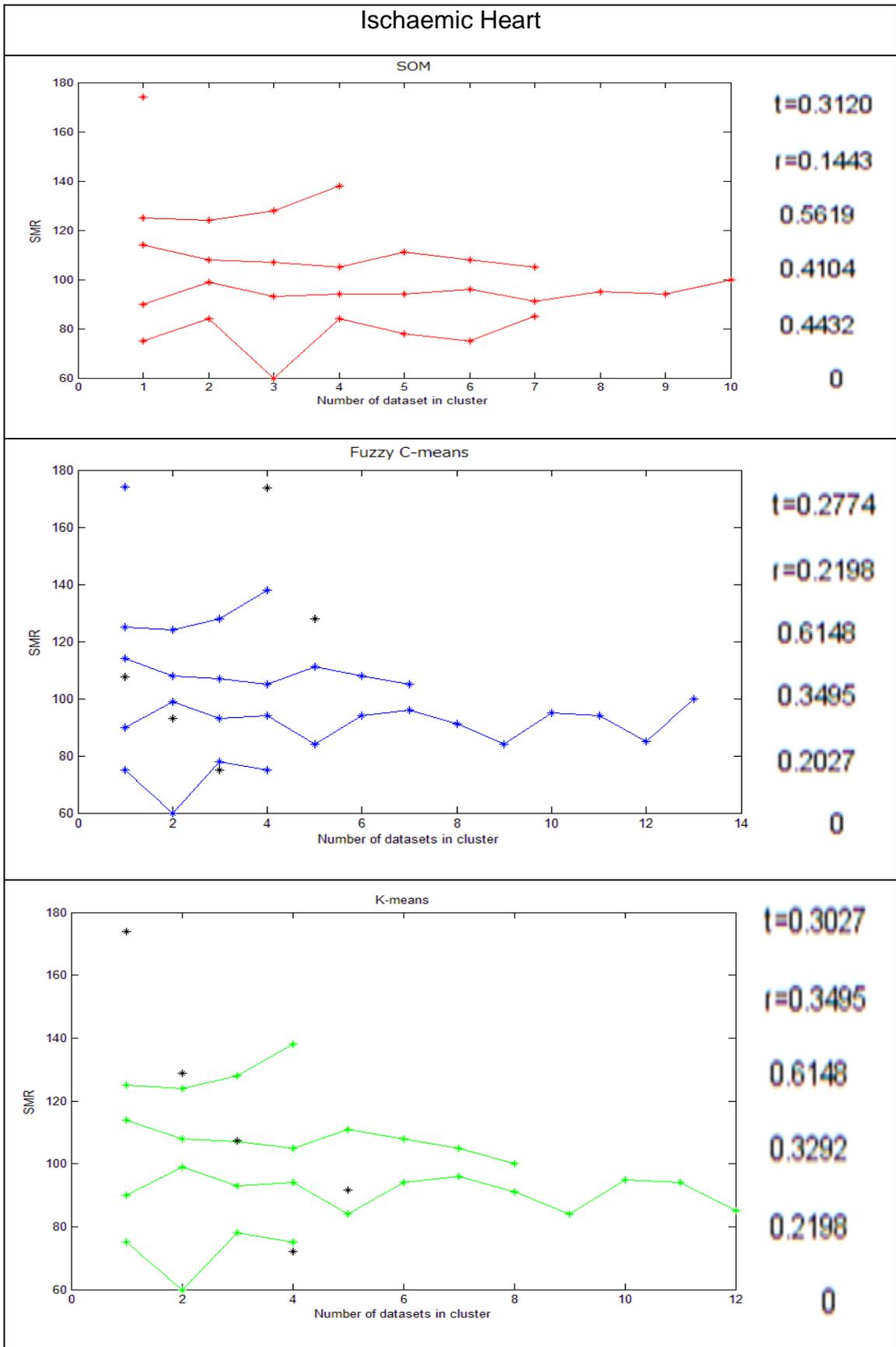


Fig 5.12 Comparison of clustering algorithms for ischaemic heart

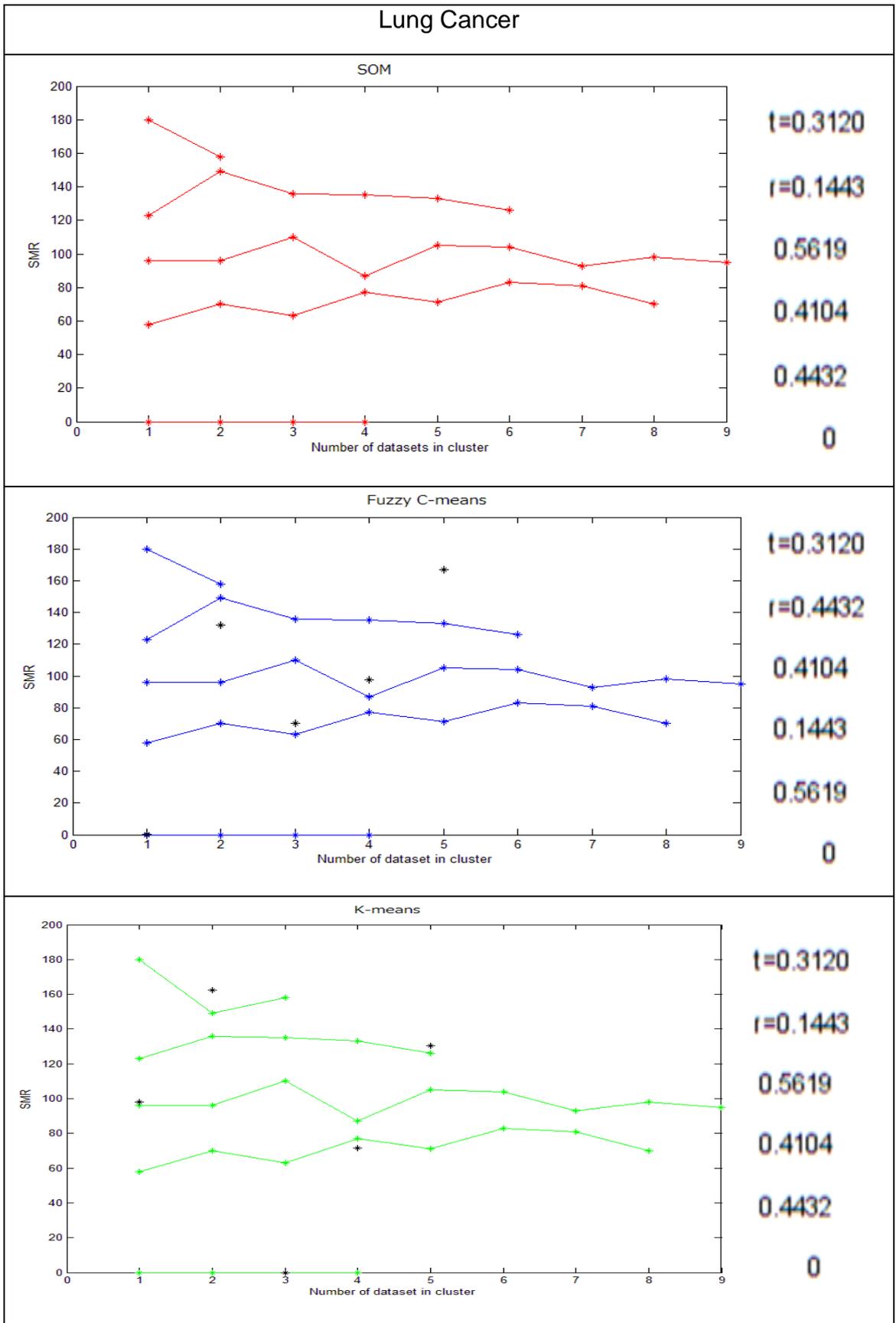


Fig 5.13 Comparison of clustering algorithms for lung cancer

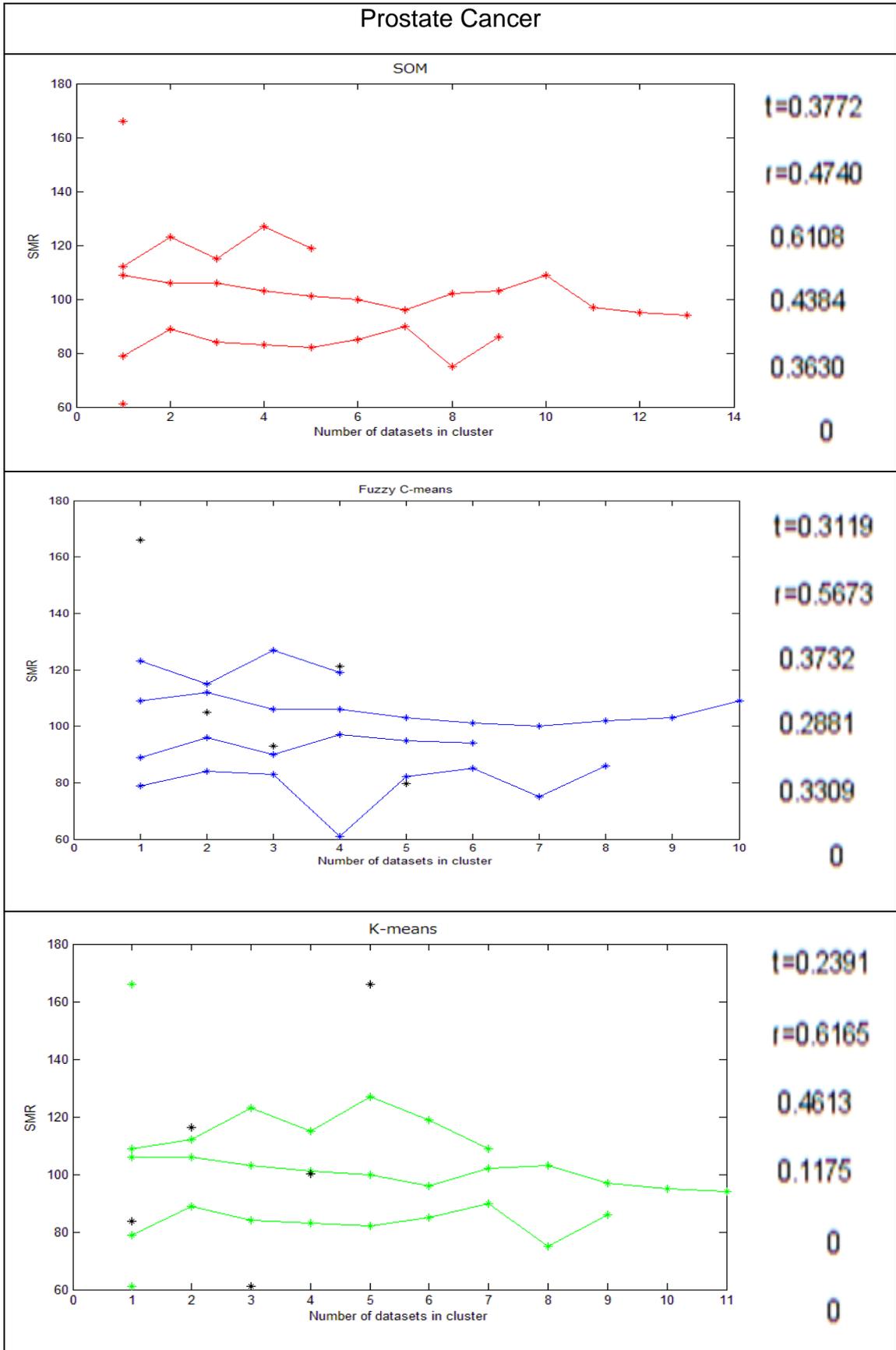


Fig 5.14 Comparison of clustering algorithms for prostate cancer

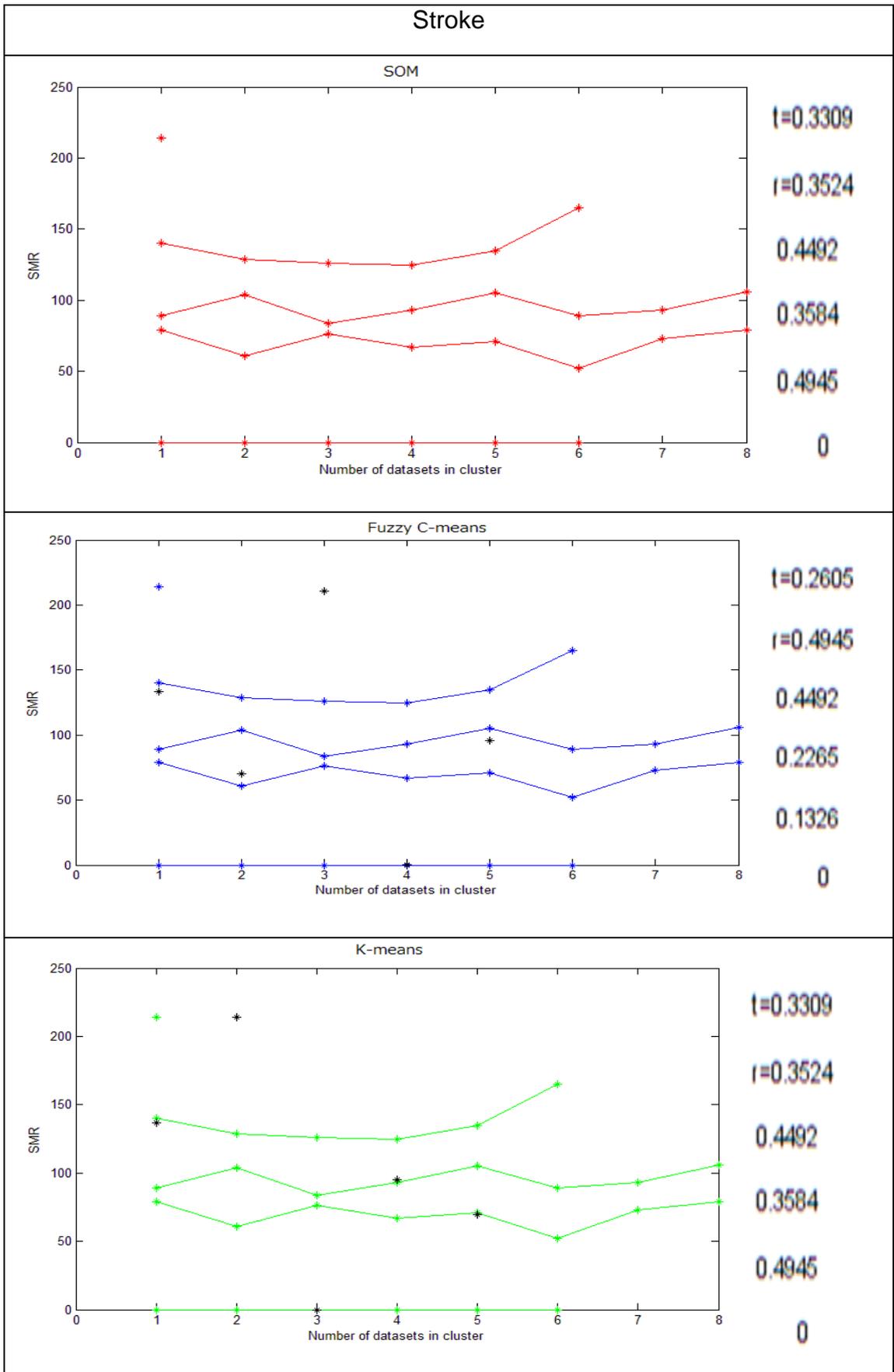


Fig 5.15 Comparison of clustering algorithms for stroke

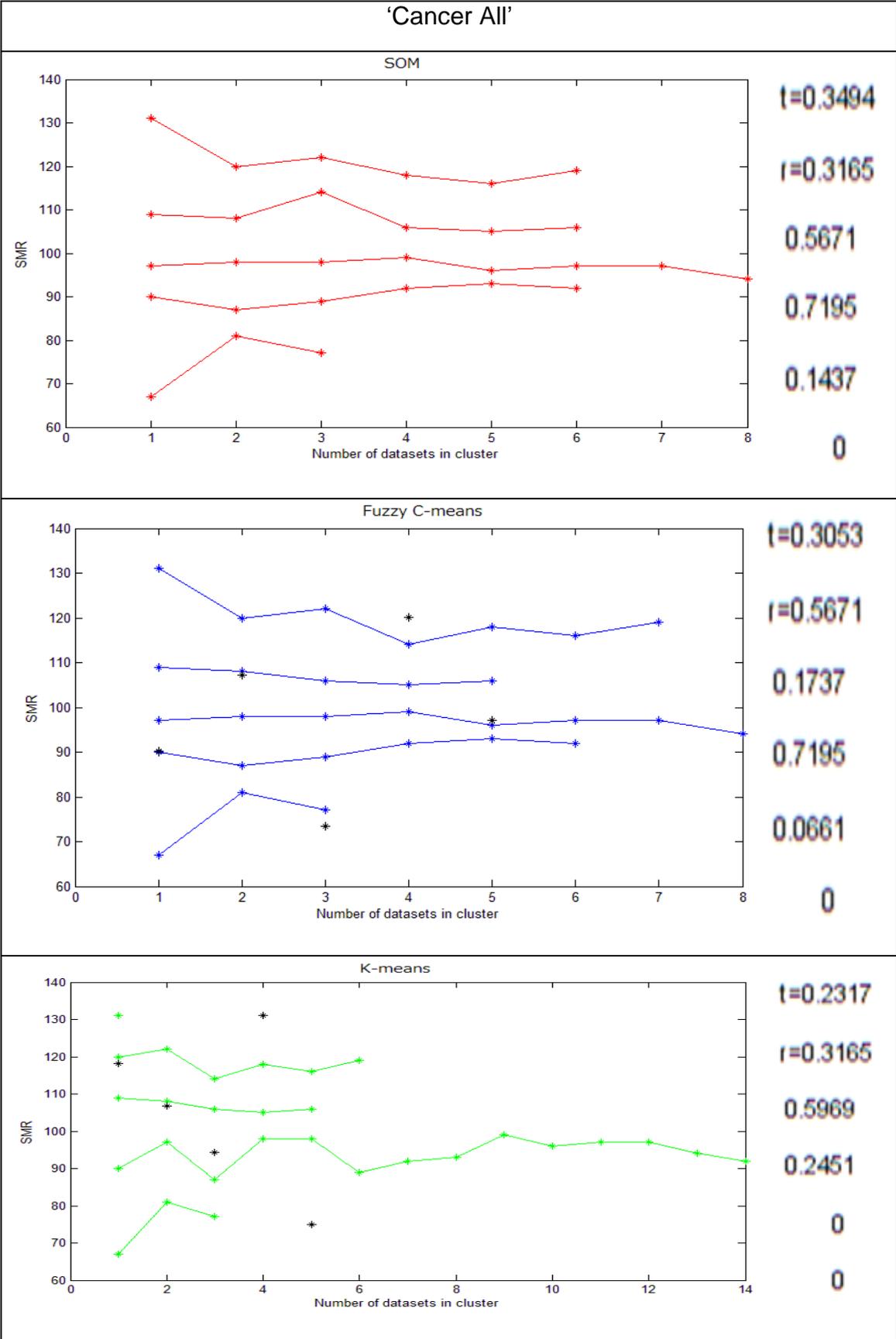


Fig 5.16 Comparison of clustering algorithms for 'cancer all'

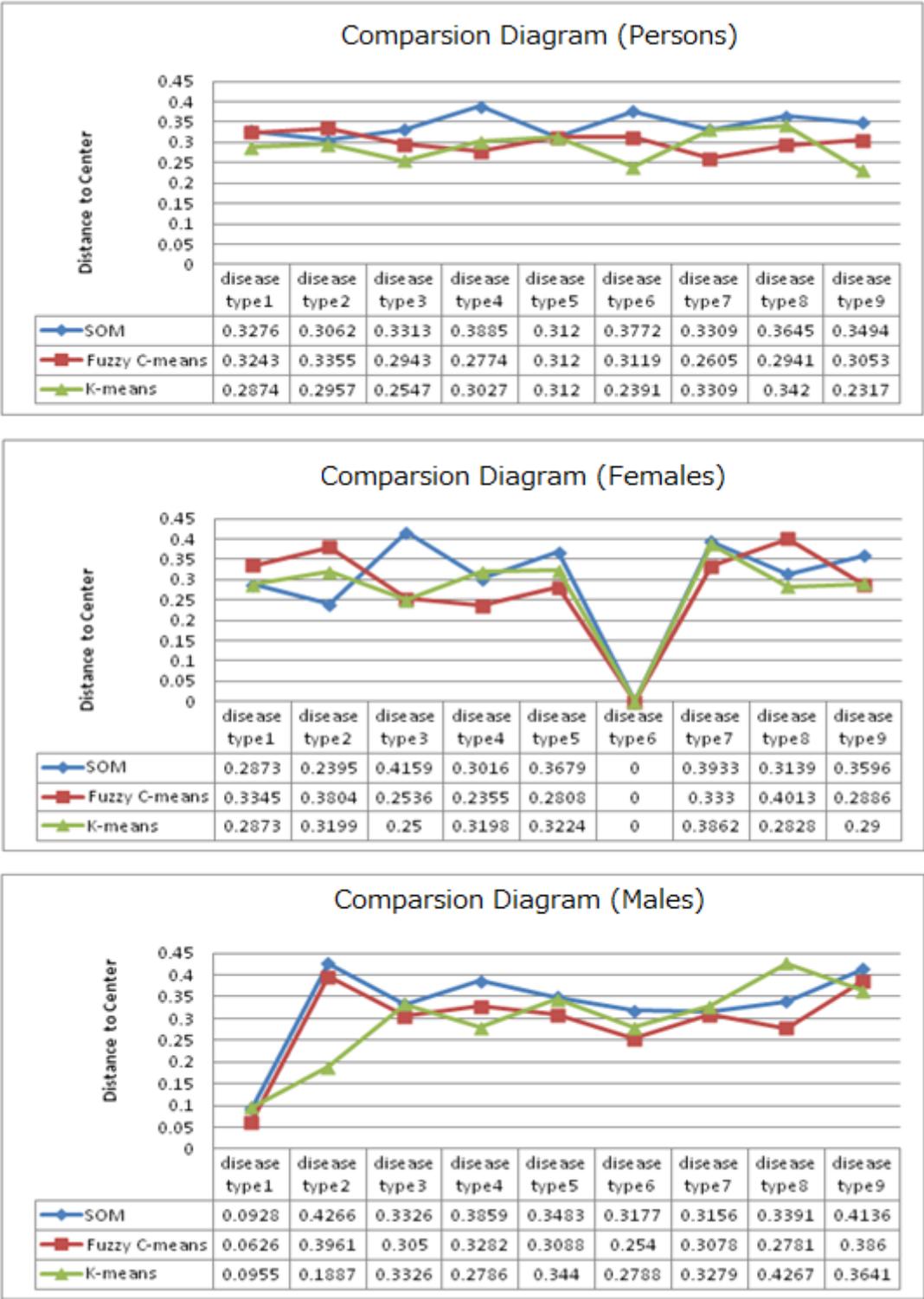


Fig 5.17 Evaluation of the SOM, FCM and k-means

5.5 Epidemiological Data Clustering Automation

The clustering automation of epidemiological data provides efficient data processing and more accurate data analysis so that the relationship between the data can be visualised. Epidemiological data normally forms patterns between epidemiological data attributes and geospatial information. Artificial Neural Networks have been used in some geospatial research areas for optimising the geospatial analysis of large and complex data sets (Zhang, Shi & Zhang 2009).

Data is automatically loaded into the WebEpi program, e.g. *webepi_cancerincidence.m* for cancer incidence data, and clustered by the k-means algorithm, after the clustering iteration process has finished, data is clustered into five groups, and the results are saved in an XML temporary file. Because the data can be organised by different year, gender and disease, the user must specify the year, so that all the data in that particular year will be clustered. There are three gender groups in the epidemiological data set, they are 'Males', 'Females' and 'Persons'. All the gender groups have the same disease category types. However, the clustering has to be conducted for each particular disease on every gender group. Therefore, the automation program not only has to classify the gender groups, but also disease categories. After the appropriate data has been specified, the k-means clustering algorithm will be executed. Previously, DHHS staff had to spend about five working days to finish the clustering process of epidemiological data (Shi et al. 2007). The whole epidemiological data clustering automation program now takes just few seconds to finish all the data clustering analysis, which is a significant improvement over

the manual data clustering analysis previously performed. The main part of the clustering automation code is shown in Fig 5.18 and Fig 5.19.

```
[epi_males,LGA_males]=xlsread(epifile,'males');

[epi_females,LGA_females]=xlsread(epifile,'females');
[epi_persons,LGA_persons]=xlsread(epifile,'persons');

[LGA_num,LGA_GIS]=xlsread('C:\LGA.xls','LGA');

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
malesdir=[xmldir,'males'];
xmlfile1=all_cause(malesdir,epi_males,LGA_males,LGA_GIS);
xmlfile2=Cancer_all(malesdir,epi_males,LGA_males,LGA_GIS);
xmlfile4=Lung_cancer(malesdir,epi_males,LGA_males,LGA_GIS);
xmlfile5=Prostate_cancer(malesdir,epi_males,LGA_males,LGA_GIS);
xmlfile6=Breast_cancer(malesdir,epi_males,LGA_males,LGA_GIS);
xmlfile7=Circulatory_diseases(malesdir,epi_males,LGA_males,LGA_GIS);
xmlfile8=Ischaemic_heart(malesdir,epi_males,LGA_males,LGA_GIS);
xmlfile9=Injury_poisoning(malesdir,epi_males,LGA_males,LGA_GIS);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
femalesdir=[xmldir,'females'];
xmlfile1=all_cause(femalesdir,epi_females,LGA_females,LGA_GIS);
xmlfile2=Cancer_all(femalesdir,epi_females,LGA_females,LGA_GIS);
xmlfile4=Lung_cancer(femalesdir,epi_females,LGA_females,LGA_GIS);
xmlfile5=Prostate_cancer(femalesdir,epi_females,LGA_females,LGA_GIS);
xmlfile6=Breast_cancer(femalesdir,epi_females,LGA_females,LGA_GIS);
xmlfile7=Circulatory_diseases
(femalesdir,epi_females,LGA_females,LGA_GIS);
xmlfile8=Ischaemic_heart(femalesdir,epi_females,LGA_females,LGA_GIS);
xmlfile9=Injury_poisoning(femalesdir,epi_females,LGA_females,LGA_GIS);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

Fig 5.18 WebEpi clustering automation (Part A)

```

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function xmlfile=all_cause(filename, epi_allcause, LGA, LGA_GIS)
xmlname=[filename, '\all_cause'];
allcause=epi_allcause(1:29);
[IDX]=kmeans(allcause, 5);
xmlfile=Epi_xml(xmlname, IDX, LGA, LGA_GIS);
end

function xmlfile=Cancer_all(filename, epi_cancerall, LGA, LGA_GIS)
xmlname=[filename, '\cancer_(all)'];
cancerall=epi_cancerall(30:58);
[IDX]=kmeans(cancerall, 5);
xmlfile=Epi_xml(xmlname, IDX, LGA, LGA_GIS);
end

function xmlfile=Lung_cancer(filename, epi_lungcancer, LGA, LGA_GIS)
xmlname=[filename, '\lung_cancer'];
lungcancer=epi_lungcancer(59:87);
[IDX]=kmeans(lungcancer, 5);
xmlfile=Epi_xml(xmlname, IDX, LGA, LGA_GIS);
end

function xmlfile=Prostate_cancer
(filename, epi_prostatecancer, LGA, LGA_GIS)
xmlname=[filename, '\prostate_cancer'];
prostatecancer=epi_prostatecancer(88:116);
[IDX]=kmeans(prostatecancer, 5);
xmlfile=Epi_xml(xmlname, IDX, LGA, LGA_GIS);
end

```

Fig 5.19 WebEpi clustering automation (Part B)

5.6 Discussion

In this chapter, data clustering algorithms have been investigated to enhance the clustering analysis of epidemiological data. Subsequently, SOMs, FCM and k-means were chosen to investigate their ability to carry out the clustering analysis of DHHS epidemiological data.

The performance and effectiveness of SOMs, FCM and k-means were evaluated using the Davies-Bouldin index, which assigns the best score to the algorithm that produces clusters which containing objects which have high similarity, and objects in different clusters which have high dissimilarity.

According to the evaluation using the Davies-Bouldin index, the k-means algorithm was the best of the three algorithms. The epidemiological data clustering results show that the k-means algorithm is superior to other classifiers, and that it was the most suitable for the DHHS WebEpi epidemiological data analysis.

Therefore, k-means was chosen as the clustering algorithm for the WebEpi System. Automation of the WebEpi clustering analysis of epidemiological data was developed successfully, and reduced the processing time from five working days to five seconds.

Before the development of WebEpi, health researchers used fixed intervals to group the epidemiological data. There are thirty-one types of cancers and diseases and three types of people groups, so health researchers had to manually process ninety-three epidemiological data files and create mappings for those files. In general, a health researcher could process about eighteen epidemiological data results per day. Therefore, to finish all the mapping would take about five working days. WebEpi has reduced this to a matter of seconds.

Chapter 6 Geospatial Processing

6.1 Introduction

This chapter focuses on the geospatial visualisation of epidemiological data. The background of WebGIS is discussed. The components of WebGIS are explained in this chapter. The development of WebEpi Geo-processing, based on WebGIS, is presented. WebEpi Geo-processing for the geospatial visualisation of epidemiological data is demonstrated. There are two major processing components in WebEpi Geo-processing, one is WebEpi Geo-Mashups and the other is the WebEpi geospatial layer. The WebEpi Geo-processing automation program has been designed and implemented. In order to prove the usability of WebEpi, a case study for WebEpi is presented.

6.2 WebGIS

Interactive web mapping services have been introduced as an extension of conventional stand-alone GISs in recent decades (Zhang & Shi 2007). A GIS is a computer system that provides a facility for spatial data input, display, management and analysis. GISs are available for many applications including environmental management, asset management, air quality analysis, land management, business strategy analysis, tax statistics, public service management, social security analysis and property valuation analysis.

The Internet has become one of the most important media for people to exchange information. WebGISs have been gradually introduced to web users. WebGISs combine geographic data, the Global Positioning System (GPS), and data management tools into one resource for users to map information according to their requirements (Zhang, Shi & Zhang 2007).

A WebGIS is a web based application which provides geographic information system functions on the web. It is a solution for offering maps and GIS services to distributed users. In the past, many WebGISs only had the function of mapping (Shi, Zhang & Zhang 2009). Nowadays, WebGISs have an integrated Web Map Service (WMS) (Jain & Wu 2007). Web Feature Service (WFS) and Geo-Mashups. WMS provide standard interfaces to request a map, and WFS provide interfaces to request geospatial features. Geo-Mashups combines non-geospatial objects and geospatial information. WebGISs enable users to control the information layers that appear in the map. The process of visualisation of information layers on the map has been developed, especially the Geo-Mashups process. In general, WMS and WFS have already been set up and can be accessed by the public or authorised users. A Geo-Mashups process has been developed specifically for WebEpi. However, there are some common strategies in the development of WebGISs. Firstly, develop the WebGIS infrastructure; secondly, develop the Geo-Mashups engine and, lastly, design the WebGIS layer file in a geospatial map file format.

6.2.1 WebGIS infrastructure

The WebGIS infrastructure describes the components which are involved in the WebGIS application. In most cases, WMS and WFS are compulsory

components in the WebGIS. WMS provide the map base for the WebGIS and WFS provide geospatial features such as polygon area, point or line. The WebGIS infrastructure illustrates the data flow of map requests and map feature requests. The process of the combination of data visualisation, spatial analysis and mapping process can be demonstrated in the WebGIS infrastructure. Well-designed infrastructures enable the developer to identify some potential challenges in the data geospatial visualisation such as map feature updates and the management of the map layer style. The Geo-Mashups engine is one of the most important components in the WebGIS infrastructure. It is used to combine non-geospatial data and geospatial information, and is crucial in WebGISs.

6.2.2 WebGIS Geo-Mashups

Web mashups are a collection of web objects. Objects can be categorised as data, application logic or user interface types (Jin et al. 2008). Geo-Mashups focus on map-based applications and web service mashups. There are some free Geo-Mashups web services available, such as Google Maps, Yahoo Maps and Bing Maps. These free Geo-Mashups services support Map APIs which are built on AJAX, JavaScript, and Flash. Map APIs allow users to build amazing map applications (Freedman 2007).

There are several mashups models which are widely used for developing Geo-Mashups services, such as the Layer Model and the Component Model. The Layer Model is based on mashups of data flows, and describe data flows from the user interface to a web server. In general, the Geo-Mashups Layer Model can be regarded as three components: Web Services/Web Data Repository, Mashups Engine and User Interface as shown in Fig 6.1. The Mashups Engine

is in the centre of the Layer Model architecture. It connects web data and the user interface. A mashups engine can be developed by using Microsoft .NET, XML and AJAX (Bioernstad & Pautasso 2007). It is focused on the integration of the web data repository and its corresponding geospatial information. It creates visualisation on the client server (Tan et al. 2008). The Web Data Repository describes the Data Repository, Web Server and Map Feature Server (Wood et al. 2007). The data in the Data Repository can be accessed by Structured Query Language (SQL) queries. The Layer Model connects the data and web services. The User Interface component plays a role in request visualisation and clustering analysis.

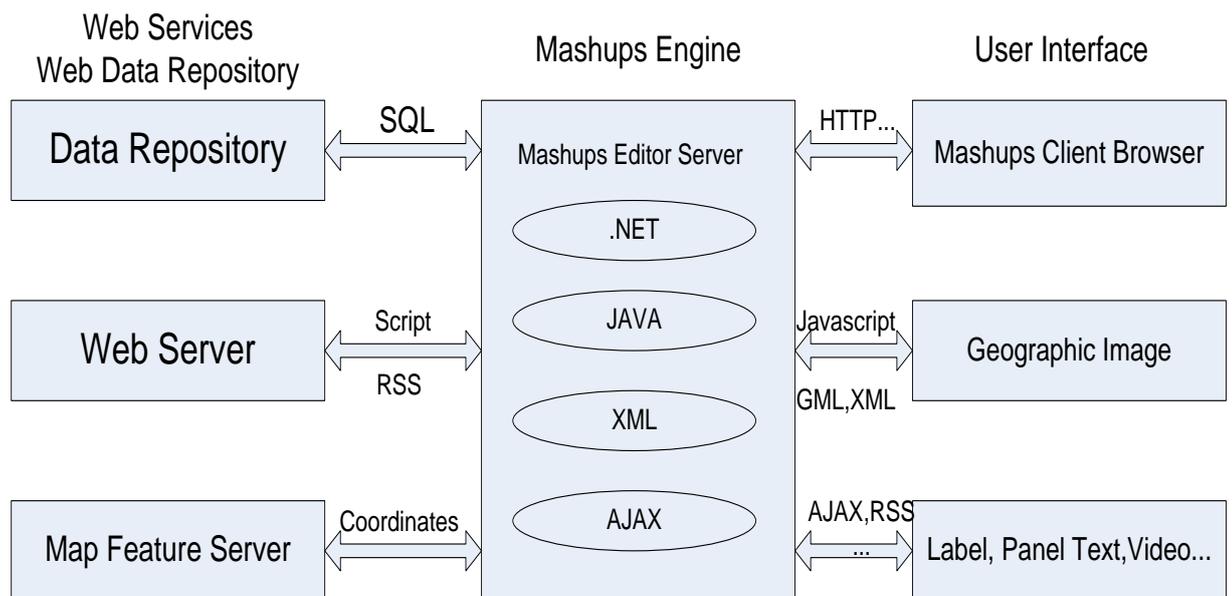


Fig 6.1 Map mashups layer model

The Component Model describes the components which are involved in the mashups process. Most of the free web mashups engines support the Component Model. The Component Model is characterised by the following properties: component type, user interface and extensibility property. The

component type can be source data, an executable web application or a user interface. Yahoo Pipes provide mashups by indentifying data sources and creating executable web applications. The mashups information can be aggregated, which aims to make the combined data more relevant. The user interface is normally supported by Javascript or XML. It provides a better way to navigate through information, and helps the presentation of the most relevant information to particular users. The functionality of mashups components can be extended. The extensibility property refers to the possibility of new function generation. Some components can be modified or specified by APIs. Users can create or modify components by programming. Once the Geo-Mashups model has been developed, the next steps are to select the layer file format and to design the layer file style for the WebGIS.

6.2.3 WebGIS layer file

WebGISs are ultimately facilitated by many web communication standards. Any web developments base on the World Wide Web and XML are compatible with WebGISs (Jain & Wu 2007).

As a common geospatial language, GML, has the capability to describe geospatial information and geospatially related data. GML is a geospatial data format that provides geospatial data management on WebGISs (Jain & Wu 2007). GML is an XML language for encoding geospatial and geospatially related information. All kinds of non-geospatial information can be embedded in GML. The information does not have to be related to location or time (Jain & Wu 2007). GML is based on XML and has similarity in schema with XML. It includes geospatial data and non-geospatial data.

6.3 WebEpi Geo-Processing

GIS files use a standard format for encoding geographic information. They are created mainly by government mapping agencies or GIS software developers. GIS file meta data often includes elevation data in either raster or vector format from the shape layers which are usually expressed as point, line or polygon combined with a coordinate system description (Shi et al. 2007). The DHHS provides map based data as polygons of LGAs. According to DHHS WebEpi requirements, Google Maps was selected as a geospatial visualisation platform for their epidemiological data. Google Maps is utilised as a WMS in WebEpi. WFS are setup in the web server, and the geospatial data uses Tasmania polygon coordinates of LGAs.

Google Maps uses XML-based KML, GML GeoRSS and SVG. They are modelling languages for geospatial visualisation platforms, data sharing and data transmission on the Internet. The WebEpi system maintains a vector format for elevation data. KML file format is chosen as the vector format for the WebEpi system. The KML files specify a set of features such as placemark, image, polygon, 3D model and textual description for display on Google Maps (Shi et al. 2007).

As the LGA geographic data provided by the DHHS is in MapInfo TAB format, it is not compatible with the Google Maps KML format, so conversion is required. Like most GIS packages, the MapInfo TAB format requires several procedures to view a data set within MapInfo Pro. The basic view is the browser view which provides storage of attributes or object data and is represented like a

spreadsheet. The accessible data is in a tabular format and no geographical information is available at this point (Shi et al. 2007).

Tiles2KML Pro is used as a converter for transforming GIS data into KML format. It includes a large number of geospatial objects. They are points, lines or polygons. Tile2KML Pro is also used to convert geospatial data on GPS and SVG geospatial images to KML. It provides a KML/GIS conversion suite which can convert between different geospatial data formats (Shi et al. 2007).

When converting a MapInfo file, geodetic data and projection properties are required. The geodetic data defines the geospatial coordinates, or coordinate degree of the Earth. If the map base is a flat map, the geospatial coordinate is purely a number of type 'float'. If the map is an oval sphere, the coordinate of the map is degree. Referencing geospatial coordinates to the wrong data can result in mapping locations incorrectly in a scale of hundreds of metres. Different countries and agencies use different data for their coordinate systems to identify the geospatial location, to adjust their GPS devices and to set the reference of their geospatial platforms. For example, the Australia government uses the GDA94 coordinate system for Map projections. The Tasmania map projection is WGS 72BE/UTM zone 55S (Shi et al. 2007).

However, there are other challenges in the development of Google Maps based raw epidemiological data visualisation, such as web geospatial mashups techniques and geospatial visualisation style. The WebEpi infrastructure as developed, includes: Pre-processing, Clustering and Geo-processing. Pre-processing and Clustering were covered in Chapters 3 and 4.

6.3.1 WebEpi Geo-processing infrastructure

WebEpi is a web-based system which provides mapping services and epidemiological data analysis. The Geo-processing data flow diagram is shown in Fig 6.2. The Map Server is the WMS, and uses Google Maps. The Web Server includes a WFS which contains Tasmania polygon coordinates of LGAs. The Web Server also serves as data storage for the WebEpi geospatial visualisation files.

After the clustering analysis process of the epidemiological data, which was described in Chapters 4 and 5, the clustering results are saved on the DHHS intranet (Zhang, Shi & Zhang 2009). The clustering results can be retrieved using SQL queries. The query results are passed on to WebEpi for Geo-processing and then their corresponding geospatial information is appended to the epidemiological clustering results. The next step is the visualisation process.

Geospatial visualisation files are usually written in an XML based language such as GML or KML. Map views are provided by different map servers, therefore the mashups data has to be converted to a format which can be visualised on a target map server (Zhang, Shi & Zhang 2009). Google Maps supports KML. KML files can be customised by a WebGIS API. The information about the map can then be presented using different legends, time zones or geographical areas (Zhang, Shi & Zhang 2009).

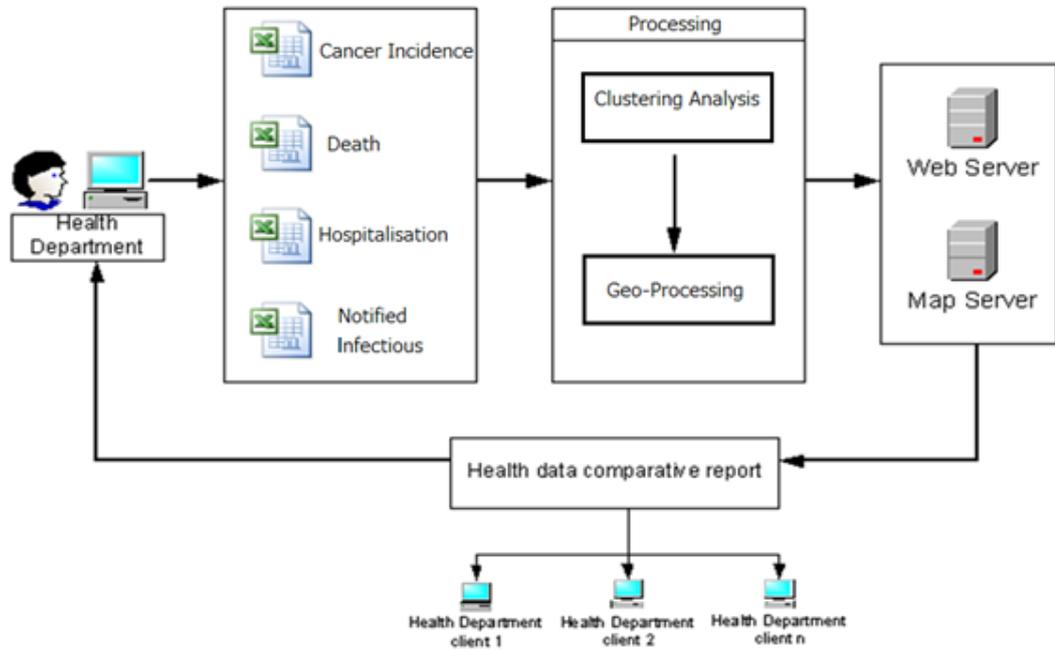


Fig 6.2 WebEpi Geo-processing data flow diagram

6.3.2 WebEpi Geo-Mashups

The WebEpi Geo-Mashups methodology consists of intranet query and a Geo-Mashups engine as illustrated in Fig 6.3. The clustering results of the epidemiological data are located in an intranet server and are extracted using SQL queries. The Geo-mashups engine is built to combine epidemiological clustering results and LGA geospatial data. The Geo-mashups engine contains Mashups browsing, Information classification, Information rating, and Information formatting. Mashups browsing performs a data extraction function that carries out partial selection from the epidemiological data clustering results. Information classification organises the extracted data cluster results and their geospatial information. Then the Information rating process adds rate attributes to the geospatial epidemiological data. The rate is used to indicate the legend of

the map. Information formatting finally converts the mashups data into a standard format which can be read by Google Maps (Zhang, Shi & Zhang 2009).

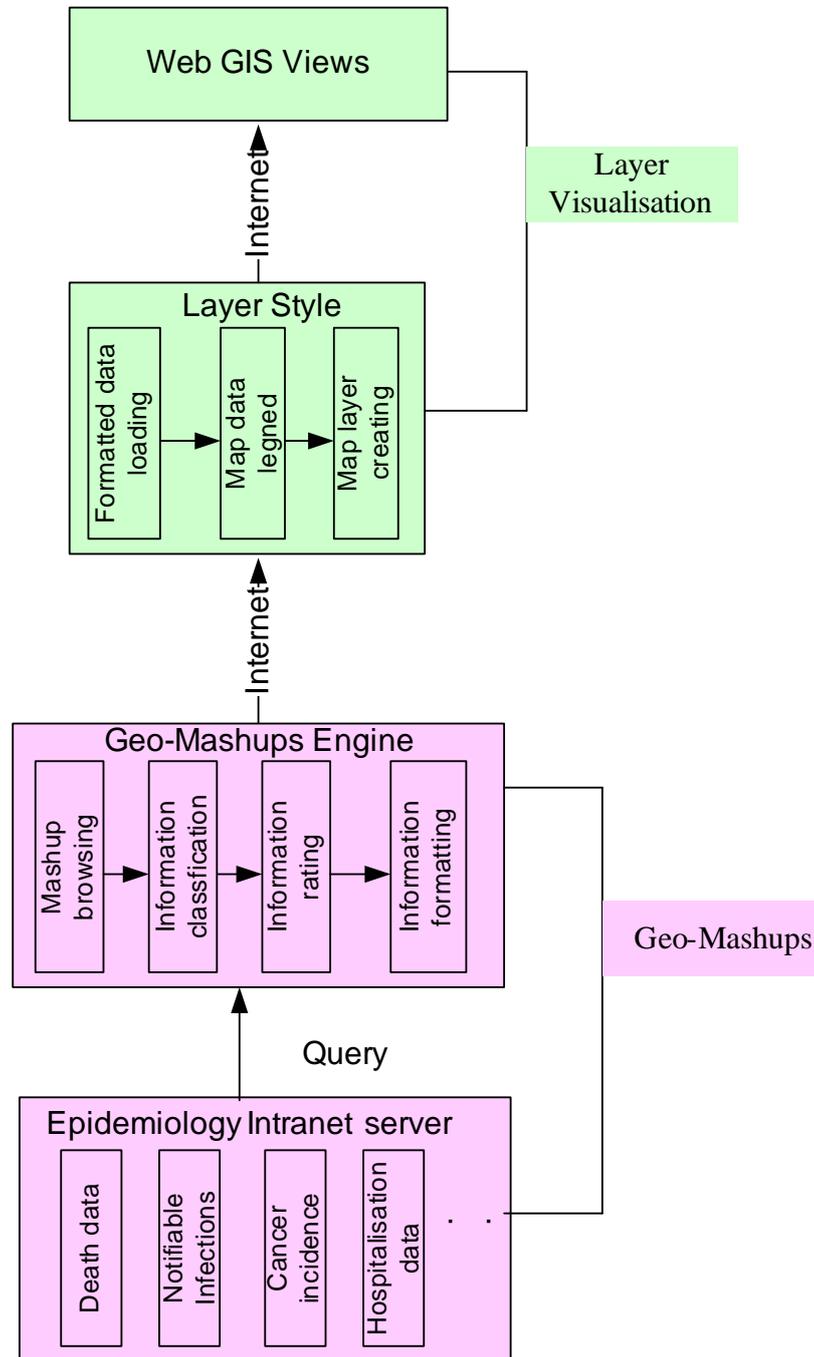


Fig 6.3 WebEpi Geo-processing block diagram

6.3.3 WebEpi geospatial layer

Google Maps uses layers to group related mapping data. Each layer file contains information about one disease in a particular year with a specific gender/group and all twenty-nine Tasmanian LGA areas. Each LGA has more than one polygon area, and all the polygons belonging to the LGA share the same style attribute (Shi et al. 2007). The style attribute consists of two elements: line style and polygon style. The line style describes line width while the polygon style describes polygon colour attribute (Shi et al. 2007). The LGA coordinate information is tagged by `<linearRing><coordinates>` as shown in Fig 6.4. The style file architecture was designed specifically for WebEpi (Shi et al. 2007).

```

<?xml version="1.0" encoding="UTF-8" ?>
- <kml xmlns="http://earth.google.com/kml/2.1">
- <Document>
  <name />
  - <Style id="">
    - <LineStyle>
      <width />
    </LineStyle>
    - <PolyStyle>
      <color />
    </PolyStyle>
  </Style>
  - <Placemark>
    <name />
    <styleUrl />
    - <Polygon>
      - <outerBoundaryIS>
        - <LinearRing>
          <coordinates />
        </LinearRing>
      </outerBoundaryIS>
    </Polygon>
  </Placemark>
</Document>
</kml>

```

Fig 6.4 A LGA definition in KML format

The epidemiological data provided by the DHHS is formatted into epidemiological data structures. The epidemiological data comes with four

epidemiological groups: Cancer Incidence, Death, Hospitalisation and Notified Infectious. The data is collated annually (Shi et al. 2007). One set of data contains three genders/groups: Males, Females and Persons, as explained in Section 3.2.1. Different epidemiological groups have different categories. For example, there are seven categories of Notified Infectious and thirteen categories of Cancer Incidence (Shi et al. 2007). The Death and Notified Infectious epidemiological data structures are described in Figs 6.5 (a) and 6.5 (b). Each category has a figure to a specific condition for all diseases. An entire WebEpi layer keeps records of the collection year, people, disease name, LGA and its SMR. The SMR comes from the population of the LGA and the number of people infected (Shi et al. 2007). The SMR is analysed into five clusters from low to high value. The centres of the clusters are automatically adjusted by distances between data sets. The SMR in each cluster is used to determine the colour coding for the LGA area in the WebEpi geospatial layer file.

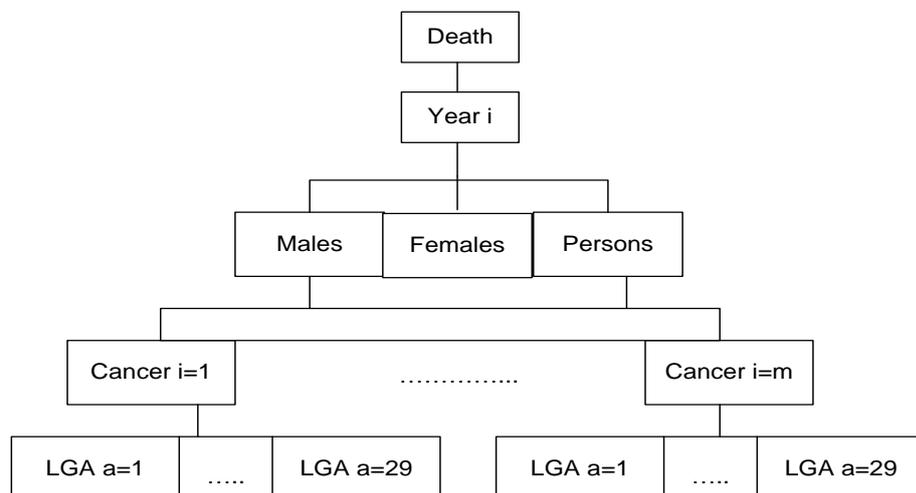


Fig 6.5 (a) Epidemiological data structures

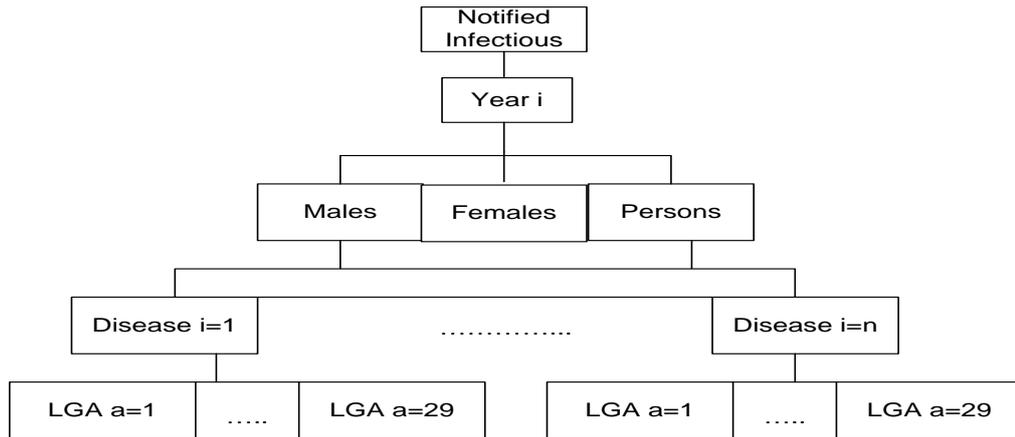


Fig 6.5 (b) Epidemiological data structures

WebEpi is a WebGIS service that locates disease notifications and presents geo-locations of notifiable diseases for the Population Health Epidemiology Unit of the DHHS (Zhang, Shi & Zhang 2009). Each of these categories has specific epidemiology data attributes. Each attribute value is represented by a number as show in Fig 6.6 (Shi, Zhang & Zhang 2009). The data is coded according to State, City, LGA or Post Code. In order to describe the attribute value on the geospatial map, WebEpi has to integrate three elements: epidemiological data attributes, a geospatial image of the LGA and a colour legend. WebEpi describes the attribute value by a colour legend as shown in Fig 6.7. Different clusters are indicated by different colours. Fig 6.7 clearly describes which cluster each particular LGA belongs to. The ratio value is clustered into five groups. The cluster thresholds are automatically adjusted to suite the scale of its SMR (Shi, Zhang & Zhang 2009).

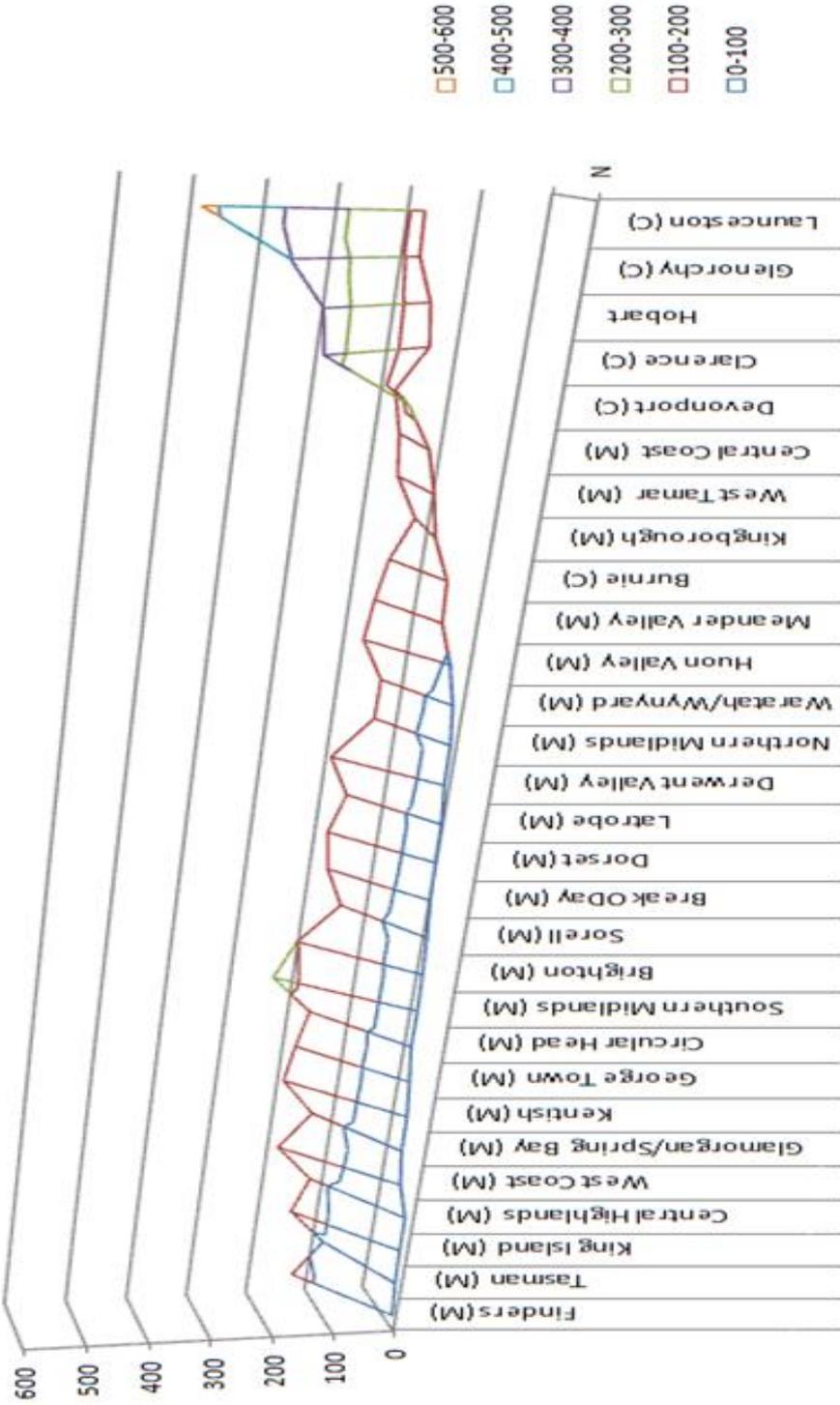


Fig 6.6 Epidemiological data attribute values

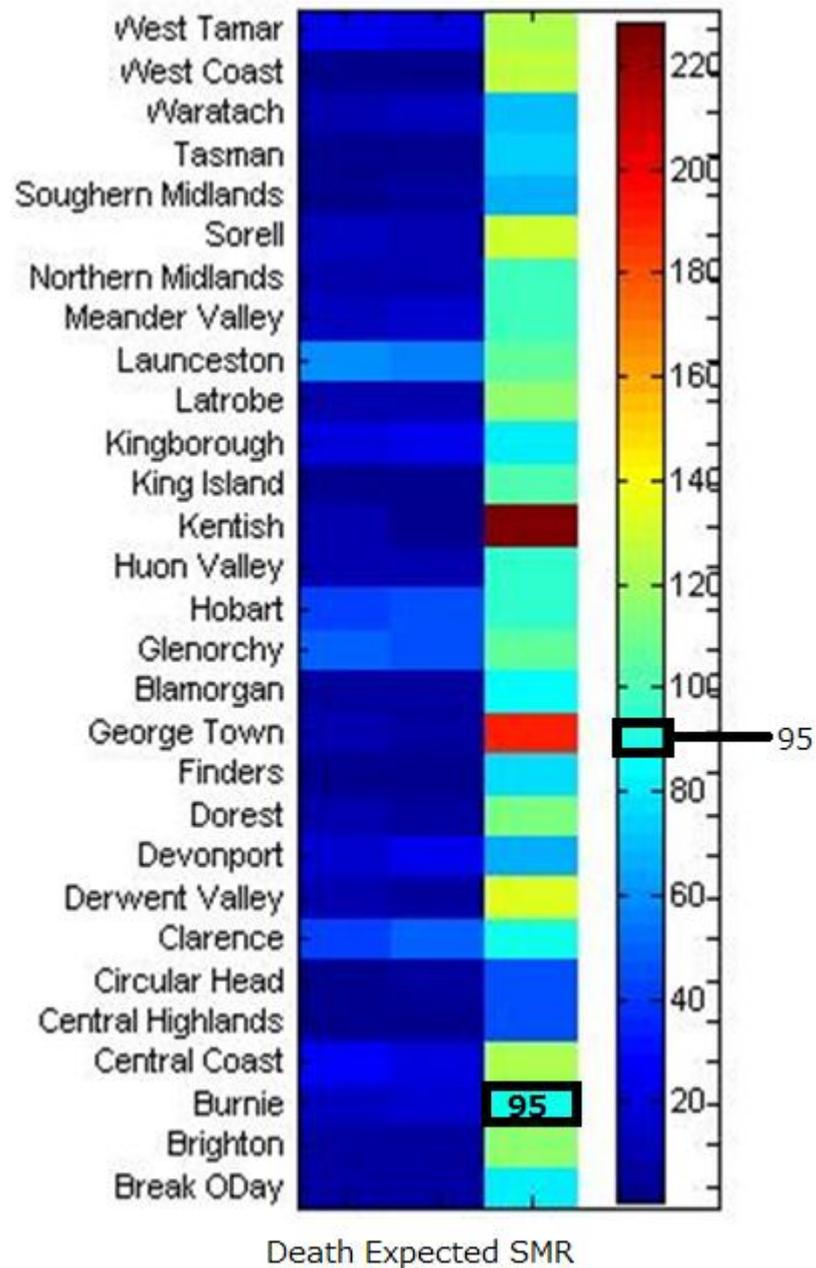


Fig 6.7 Colour legend

The results of mashups are uploaded to the web server for adding a map legend and to create a colour polygons layer. Google Maps uses geospatial layers for geospatially related data visualisation. In this research, each layer file

contains information which includes disease name, gender, LGA name and SMR for the disease. Each LGA has more than one polygon, and all the polygons belonging to the LGA share the same style attribute (Shi et al. 2007). Polygon colour is one of the style attributes which is used for the creation of the map. The polygon colour legend indicates the epidemiological data cluster thresholds. The epidemiological data clusters are presented using different colours. Each colour indicates the thresholds of its cluster. In order to allow users to visualise the road map and satellite map, all the coloured LGA areas are 80% opaque, which is specified by the DHHS, so that the Tasmanian geographic map is visible in the background. For example, in 2005 the value for the Death category, disease Injury & Poisoning SMR for Males in the Burnie area is 95 as shown in Fig 6.7, the colour for this area is cyan.

The Google Maps API provides a JavaScript function which can be used to customise the map overlay. Without the map legend, the map only describes the geospatial information itself. With the legend, the map can also describe other information related to the geospatial location. The customised Google Maps API uploads the legend overlay file to the map server. The entire WebEpi GIS overlay file is stored on the web server. Every map view can be toggled on and toggled off to select different epidemiological data attributes and time frames (Shi, Zhang & Zhang 2009).

6.4 WebEpi Geo-processing Automation

Before the development of WebEpi, DHHS health researchers manually analysed the epidemiological data and visualised the results on commercial

mapping services. The WebEpi mashups automation system has been used on the DHHS (years 2001 to 2005) epidemiological data. In total, there are around 6,300 records in the epidemiological data repository (Zhang, Shi & Zhang 2009). Each epidemiological data record contains its LGA and People Group information. It was necessary to build an automated Geo-processing program for WebEpi, otherwise it would have taken a great deal of time and effort to manually create epidemiological layer files for Google Maps (Shi, Zhang & Zhang 2009).

The automated WebEpi Geo-processing was developed using Microsoft ASP.NET which is a server-side Web application framework designed and developed by Microsoft. It starts with a data query, which enables the querying of the epidemiological clustering results. The data query object diagram is shown in Fig 6.8. The WebEpi Geo-processing sends a DataConn request first and then passes the DataConn request to query an object in the Intranet data repository. After extracting the data from Intranet, the epidemiological clustering results have to be formatted ready for the Geo-Mashups process (Shi, Zhang & Zhang 2009).

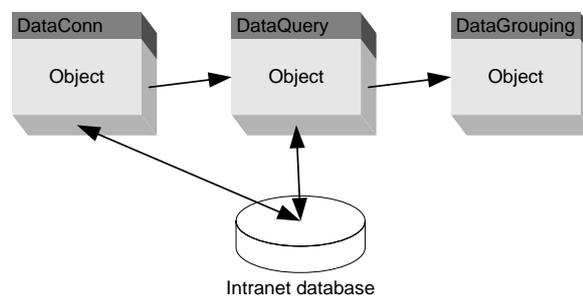


Fig 6.8 Data query

Geo-Mashups is the one of the parts in Geo-processing. Fig 6.9 shows the Unified Modelling Language (UML) diagram for the automated Geo-Mashups process. The Geo-Mashups package has four classes: Mashups class, Classification class, Rating class and Formatting class. The Geo-Mashups package first calls the Mashups class. After getting the source data from the epidemiological database, the Classification class re-structures the data by clustering result. A value indicating the clustering result is then appended to each data set by the Rating class. Then the clustering results of epidemiological data are formatted into KML by the Formatting class (Shi, Zhang & Zhang 2009).

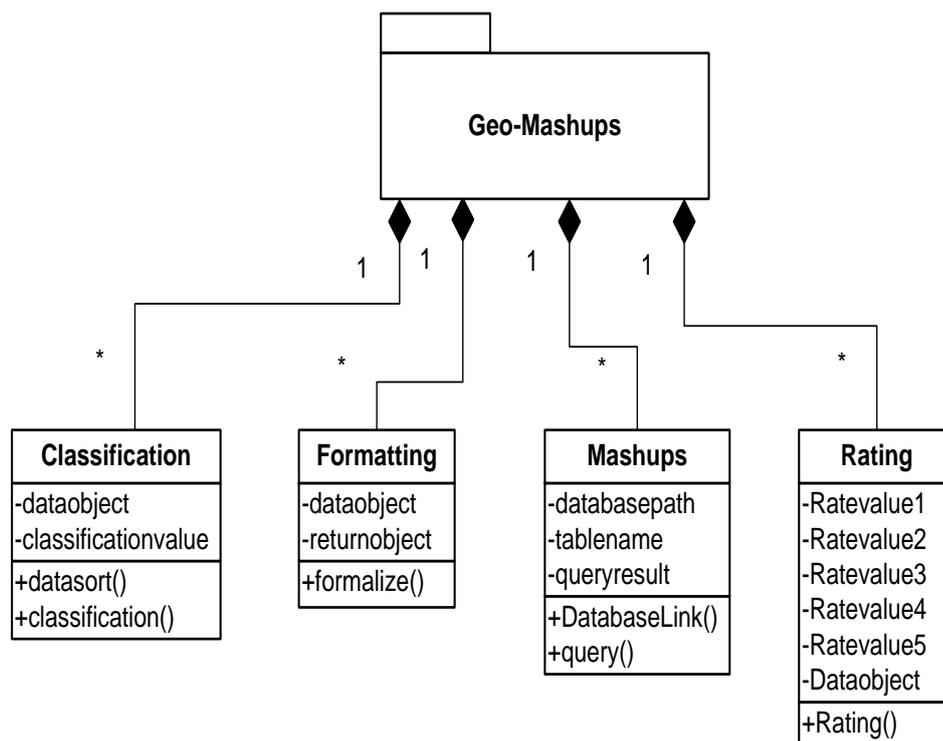


Fig 6.9 Geo-Mashups

The third step is the map feature loading process as shown in Fig 6.10. The map feature file type is KML. The WebGIS API Read Data, loads the map

feature onto the server and Add Legend appends the legend to the KML file. Finally, Layer completes the process and the file schema of epidemiological data is created. The layer file is then ready to be uploaded to the Google Maps server for visualisation (Zhang, Shi & Zhang 2009).

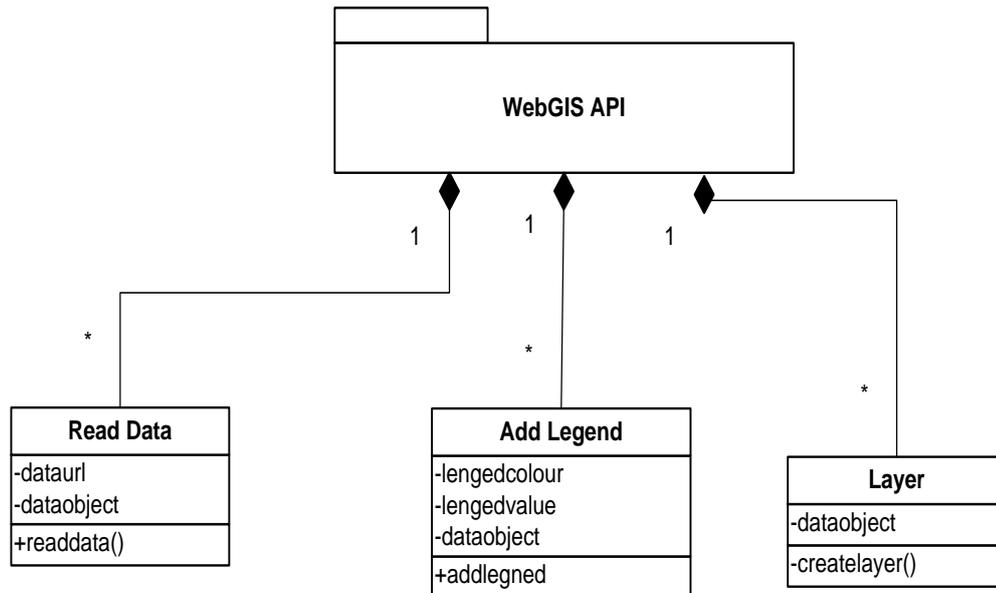


Fig 6.10 Map feature loading

6.5 WebEpi Case Study

Experiments on WebEpi Geo-Mashups automation involved connection with an intranet server process, file format conversion and map file creation (Zhang, Shi & Zhang 2009). The source data stored on the Intranet server was in Excel format as shown in Fig 6.11. It contained geospatial information and some epidemiological attributes. Then the next step of automation was to convert the Excel file to a KML geospatial layer file, which is shown in Fig 6.12. After adding the legend and rating information to the KML file, the map file was created and

ready to be uploaded to the Google Maps server for visualisation as shown in Fig 6.13. The SMR colour legend indicates the SMRs of the LGAs.

	A	B	C	D	E
1	Mortality data				
2					
3	Sex	Disease	LGA	Number	ASR
4	Males	Circulatory diseases	Area	N	ASR
5	Males	Circulatory diseases	Finders (M)	6	125.5
6	Males	Circulatory diseases	Tasman (M)	11	81
7	Males	Circulatory diseases	King Island (M)	13	143.9
8	Males	Circulatory diseases	Central Highlands (M)	14	120.4
9	Males	Circulatory diseases	West Coast (M)	31	182.3
10	Males	Circulatory diseases	Glamorgan/Spring Bay (M)	39	134.3
11	Males	Circulatory diseases	Kentish (M)	40	189.8
12	Males	Circulatory diseases	George Town (M)	46	178.8
13	Males	Circulatory diseases	Circular Head (M)	53	164.3
14	Males	Circulatory diseases	Southern Midlands (M)	53	231.3
15	Males	Circulatory diseases	Brighton (M)	55	201.7
16	Males	Circulatory diseases	Sorell (M)	58	141.1

Fig 6.11 Sample of Epidemiology source data in Excel

```

<?xml version="1.0" encoding="UTF-8"?>
<kml xmlns="http://earth.google.com/kml/2.1">
  <Folder>
    <name>TAS API</name>
    <Document>
      <name>101</name>
      <Style id="transparent101">
        <LineStyle>
          <width>1.5</width>
        </LineStyle>
        <PolyStyle>
          <color>ccffff00</color>
        </PolyStyle>
      </Style>
      <Placemark>
        <description>
          <![CDATA[<b>MUNICIPAL_NO:</b> <i>101</i><br />
<b>CPR_NUMBER:</b> <i>CPR2470</i><br />]]>
        </description>
        <styleUrl>#transparent101</styleUrl>
        <Polygon>
          <outerBoundaryIs>
            <LinearRing>
              <coordinates>
                148.263335098265,-40.90088618234857,-5.576164845377
              </coordinates>
            </LinearRing>
          </outerBoundaryIs>
        </Polygon>
      </Placemark>
    </Document>
  </Folder>
</kml>

```

Fig 6.12 Epidemiology KML file

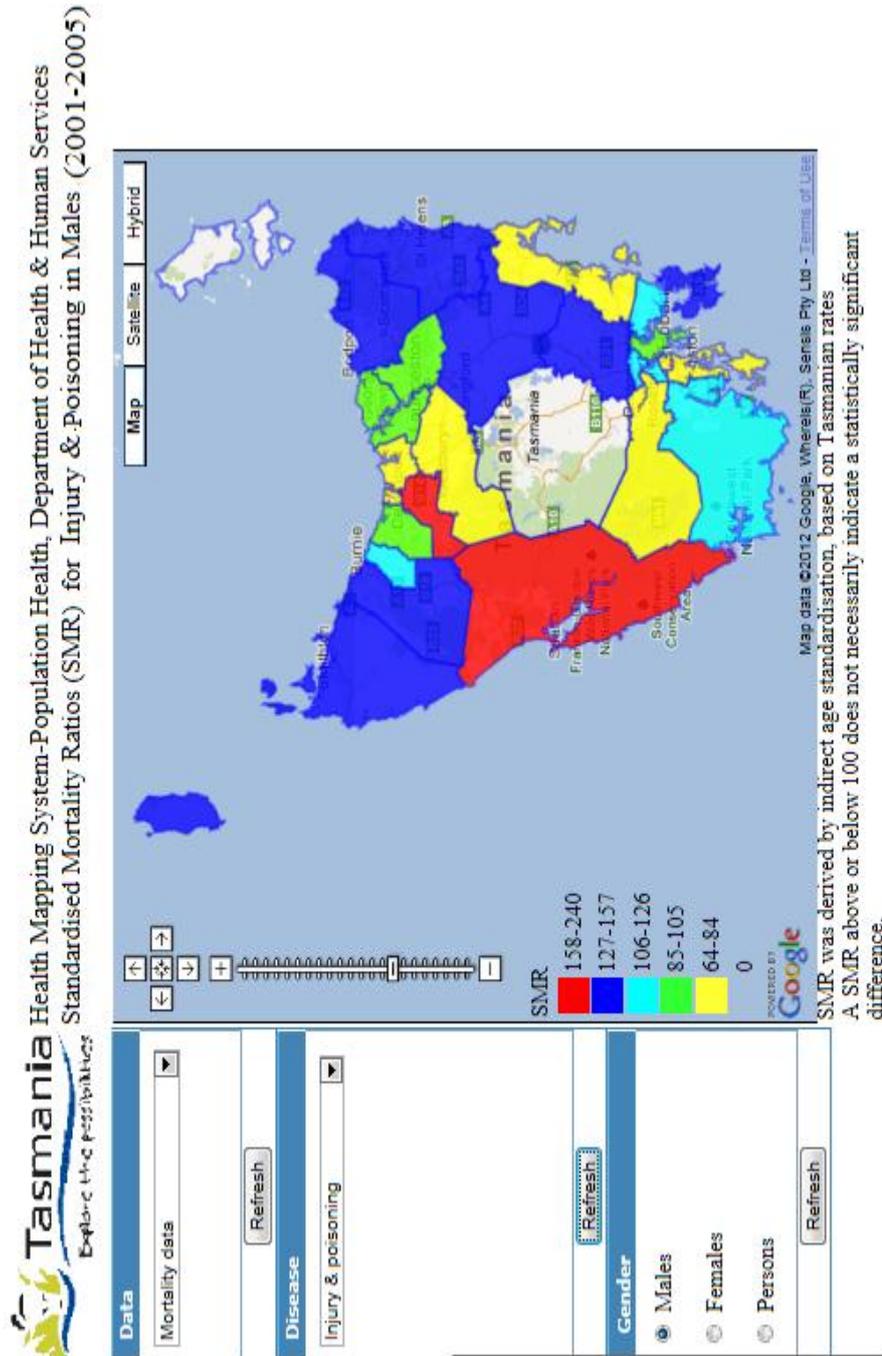


Fig 6.14 Mapping for Males SMR in injury & poisoning

(CD-ROM: Demo\Google Maps\index.html)

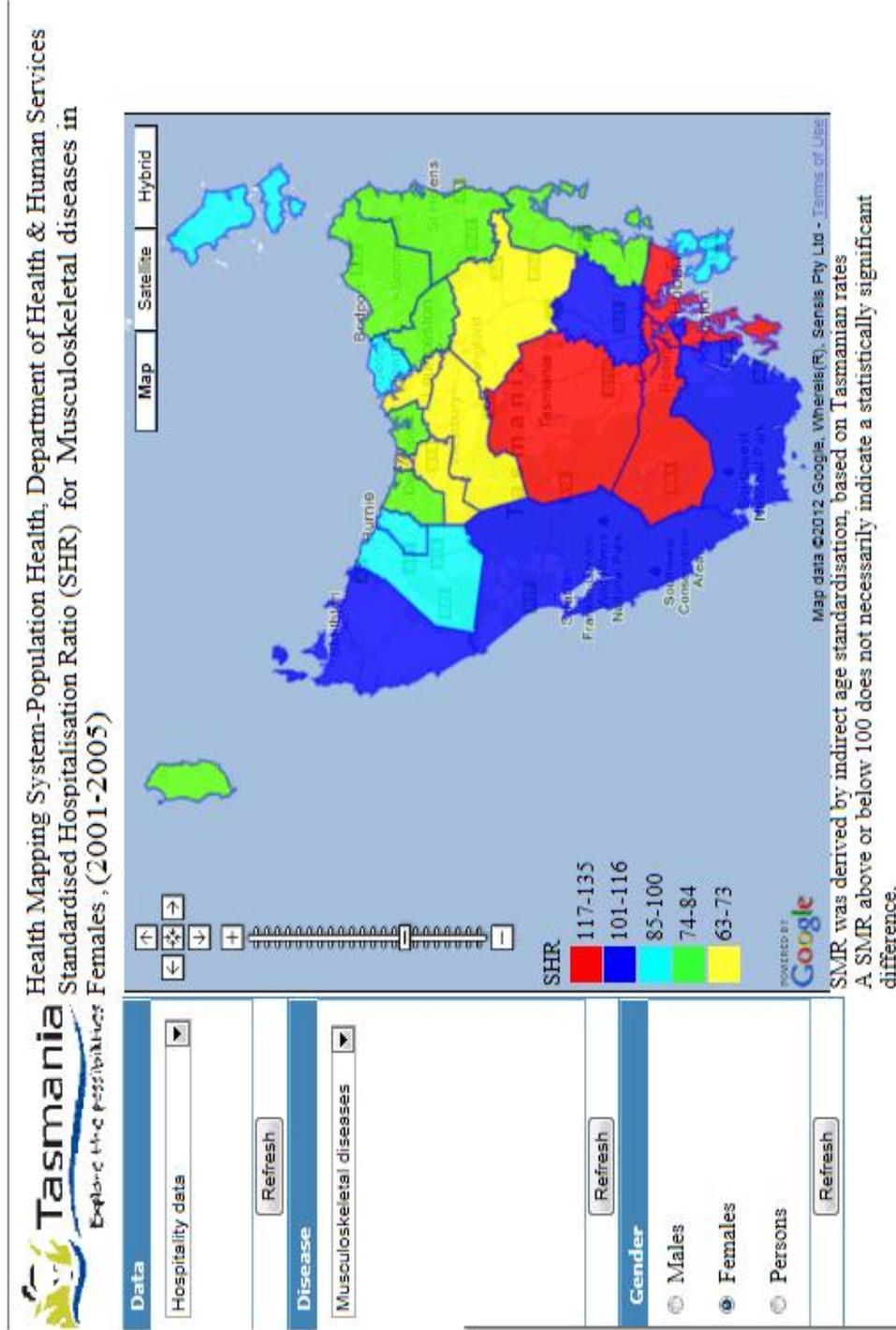


Fig 6.15 Mapping for females hospitalisation data in musculoskeletal disease
 (CD-ROM: Demo\Google Maps\index.html)

6.6 Summary

In this chapter, WebGISs have been reviewed. Based on WebGIS infrastructure theory, the WebEpi Geo-processing was designed and developed, according to DHHS requirements. The design includes the Geo-processing infrastructure, Geo-Mashups and a Geospatial layer. The WebEpi Geo-processing infrastructure has produced a significant contribution in the geospatial visualisation of epidemiological clustering results. In order to improve the work efficiency of WebEpi Geo-processing, an automation program for Geo-processing was built as well. As shown in the case study, the epidemiological clustering results were successfully mapped on Google Maps. An evaluation has been conducted by the DHHS, and WebEpi has received positive feedback. The WebEpi system has been adopted by the DHHS as their epidemiological data reporting system. The DHHS confirmed that WebEpi has produced better results than those produced from the commercial software package MapInfo (Shi et al. 2007). The case study has demonstrated that WebEpi can help the DHHS health researchers to conduct investigations on the relationships between geospatial location and disease distribution.

Chapter 7 Conclusions

The findings of this research include three components. First the best clustering algorithm for epidemiological data has been identified. Secondly, the geospatial visualisation for clustering analysis of epidemiological data has been implemented. Lastly, a WebGIS system, called WebEpi, which provides a precise, effective and intuitive geospatial visualisation of clustering analysis of epidemiological data, has been developed.

7.1 Summary of Contributions

In this research, the epidemiological data was provided by the DHHS. In order to conduct the clustering analysis, the experimental data had to be pre-processed; this was done by analysing the epidemiological data. Then the epidemiological data was restructured and saved as a unified format that was suitable for clustering analysis and geospatial visualisation.

Clustering algorithms were reviewed. SOMs, FCM and k-means were selected for clustering experiments of epidemiological data. A clustering validation algorithm (Davies-Bouldin index) was selected to validate the experiment results. K-means was identified as the best clustering algorithm for DHHS epidemiological data, and was, therefore, chosen for WebEpi.

Because of the WebGIS requirements from the DHHS, Google Maps was selected for geospatial visualisation. In order to combine the clustering analysis of epidemiological data and geospatial data, Geo-processing was developed,

which includes two parts: one is Geo-Mashups; the other is a geospatial layer. Geo-processing delivers the successful mapping of the clustering analysis of epidemiological data on Google Maps.

In order to produce an efficient geospatial visualisation of the clustering analysis of epidemiological data, a fully automated system which combines k-means clustering analysis and Geo-processing, for geospatial visualisation of the clustering analysis of epidemiological data, was required for WebEpi. WebEpi successfully embeds the k-means clustering algorithm and Geo-processing for the DHHS.

This research has made three contributions:

- Choosing the best clustering algorithm for DHHS epidemiological data.
- Implementing automation of the geospatial visualisation for clustering analysis of epidemiological data
- Building the user interactive web-based system, WebEpi.

7.1.1 Clustering analysis of epidemiological data

The data clustering algorithm is the core of this epidemiological analysis. The clustering validation algorithm is the key for the epidemiological clustering algorithm selection. The process of epidemiological clustering consists of two procedures: the clustering analysis and the validation of the clustering results of epidemiological data.

SOMs, FCM and k-means have been applied to the clustering analysis of epidemiological data. However, SOMs, FCM and k-means can produce very

similar clustering results. Therefore the Davies-Bouldin index algorithm was used for clustering validation of the results of epidemiological data. The validation output shows that k-means had the least interval distance and FCM had the second lowest interval. The SOM had the largest interval distance measurement. The one which had the least interval distance won the comparison. Therefore, the k-means clustering algorithm was selected for the DHHS clustering analysis of epidemiological data.

7.1.2 Geospatial visualisation of epidemiological data

A methodology of epidemiological data geospatial visualisation based on Google Maps has been developed; so far no one else has investigated that. The epidemiological data visualisation methodology is one of the most significant contributions in this thesis. Google Maps was selected as a geospatial visualisation platform because Google Maps provides a free geospatial visualisation service and APIs which enable users to customise geospatial mapping with free access. The process includes the development of Geo-Mashups and geospatial layers.

The Geo-Mashups engine combines epidemiological data and geospatial data. In this thesis, the epidemiological data geospatial visualisation process was developed. Geospatial layer files were created for WebEpi visualisation. The geospatial layer file has polygons based on LGAs. The polygon colour of LGAs indicates the SMRs for particular diseases in these LGAs.

7.1.3 WebEpi

WebEpi is a web-based system for geospatial visualisation of the clustering analysis of epidemiological data for the DHHS. Public accessibility to WebEpi allows health researchers and the public to have a better understanding of epidemiological data analysis reports. WebEpi includes loading the epidemiological data Excel file, converting it to XML format, performing a clustering analysis of the XML file using the k-means algorithm, Geo-Mashups of the clustering results, converting the Geo-Mashups results to KML format and forming a visualisation of the KML file on Google Maps using a WebGIS API. The source code of the experiment programs and experiment results can be found in the attached CD-ROM, Appendix D.

7.2 Conclusions

This research not only addresses critical issues associated with improving the quality and accuracy of the clustering analysis of epidemiological data, but also puts emphasis on developing a publicly accessible geospatial visualisation system. It achieves the two major research objectives and makes significant contributions to the clustering analysis of epidemiological data and geospatial visualisation. Massive amounts of epidemiological data have been translated into meaningful epidemiological information, and a WebGIS based geospatial visualisation system provides public access to the clustering analysis of epidemiological data.

The academic contributions of this thesis can be divided into two parts; one is the combination of the clustering analysis and geospatial visualisation for

epidemiological research, the other is the development of the architecture for WebEpi. The automated data processing of raw epidemiological data into a customised geospatial layer is the IT contribution of this project.

WebEpi helps DHHS health researchers avoid having to manually process epidemiological data, thus improving their efficiency in producing epidemiological data geospatial reports. The utilisation of Google Maps has also enabled the DHHS to save a large amount of money on commercial geospatial services (Shi et al. 2007). DHHS health researchers can use WebEpi to conduct public health surveillance and make health service plans in Tasmania. There is no doubt that WebEpi can help DHHS provide an open access reporting system to public.

7.3 Future Work

This research has produced significant results in the geospatial visualisation of k-means clustering results for epidemiological analysis. The WebEpi architecture, which is based on these results, has been successfully designed and developed. It is the first time that geospatial epidemiological data visualisation has been represented on Google Maps and enables public access to this information. WebEpi can be further improved over many areas. More clustering algorithms will be considered in health data analysis. In the future, the WebEpi architecture could be extended and applied to other public health related areas such as the statistics of occupational diseases.

References

- Aksoy, E. 2006, "Clustering With GIS:An Attempt to Classify Turkish District Data", *Shaping the Change XXIII FIG Congress*, Munich, Germany, 8-13 October, 2006, viewed 20 December 2012 < http://www.fig.net/pub/fig2006/papers/ts47/ts47_05_aksoy_0327.pdf > .
- Australian Bureau of Statistics 2010, *Australian Social Trends, Sep 2010*, viewed 12 October 2012 < <http://www.abs.gov.au/AUSSTATS/abs@.nsf/0/97999E77DB6843A1CA2577F80010DD40?opendocument>>
- Australian Government Department of Health and Ageing 2013, *Sources of epidemiological data for use in generating utilisation estimates*, viewed 20 May 2013 <<http://www.pbs.gov.au/info/industry/useful-resources/sources> >
- Balaji, K. & Zacharias, N.J. 2007, *Fuzzy C-Means*, viewed 15 Jan 2013 < http://www.iiitmk.ac.in/wiki/images/d/d2/Fuzzy_CMeans.pdf >
- Basara, H. & Yuan, M. 2008, "Community health assessment using self-organizing maps and geographic information systems", *International Journal of Health Geographics*, vol. 7, no. 1, pp. 67-75
- Bezdek, J.B. 1980, "A convergence theorem for the fuzzy ISODATA clustering algorithms," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-2, pp. 1-8.
- Bezdek, James C. 1981, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Springer

- Bioernstad, B. & Pautasso, C. 2007, "Let it flow: Building Mashups with Data Processing Pipelines", *Proceedings of Mashups'07 International Workshop on Web APIs and Services Mashups (ICSOC'07)*, Vienna, Austria, 17 September 2007, pp 15-28
- Bradley, P.S. & Fayyad, U.M. 1998, "Refining Initial Points for K-Means Clustering", *Proceedings of the Fifteenth International Conference on Machine Learning*, Morgan Kaufmann Publishers Inc, San Francisco, CA, USA, pp. 91-99
- Carvalho, J.P. & Tome, J.A.B. 1999, "Rule based fuzzy cognitive maps and fuzzy cognitive maps-a comparative study", *18th International Conference of the North American on Fuzzy Information Processing Society (NAFIPS)*, New York, USA, 10-12 June 1999, pp.115-119.
- Chang, M., Yu, H. & Heh, J. 1998, "Evolutionary self-organizing map", *Proceedings of the IEEE International Joint Conference on Neural Networks*, AK, 4-8 May 1998, vol. 1, pp. 680-685.
- Chaudhuri, A., De, K. & Chatterjee, D. 2008, "A Study of the Traveling Salesman Problem Using Fuzzy Self Organizing Map", *Proceedings of the IEEE Third international Conference on Industrial and Information Systems (ICIIS)*, Kharagpur, 8-10 Dec 2008, pp.1-5.
- Chi, Y., Han, Y., Rui, L. & Wei, Y. 2010, "Study of Bus Incident Prediction Based on Dynamic Fuzzy-Neural Network", *International Conference on E-Product E-Service and E-Entertainment (ICEEE)*, Henan, China, 7-9 November 2010, pp. 1-6.

- Chu, X., Zhu, Y., Shi, J. & Song, J. 2010, "Method of image segmentation based on Fuzzy c-means Clustering Algorithm and Artificial Fish Swarm Algorithm", *International Conference on Intelligent Computing and Integrated Systems (ICISS)*, Guilin, China, 22-24 October 2010, pp. 254-257.
- Colantonio, A., Moldofsky, B., Escobar, M., Vernich, L., Chipman, M. & McLellan, B. 2011, "Using Geographical Information Systems Mapping to Identify Areas Presenting High Risk for Traumatic Brain Injury", *Emerging Themes in Epidemiology*, vol. 8, no. 1, pp. 7-18.
- Davies, D.L. & Bouldin, D.W. 1979, "A Cluster Separation Measure", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224-227.
- Deng, Q. & Mei, G. 2009, "Combining self-organizing map and k-means clustering for detecting fraudulent financial statements", *IEEE International Conference on Granular Computing (GRC '09)*, Nanchang, China, 17-19 August 2009, pp. 126-131.
- Department of Human Service and Health Tasmania 2003, *Health Indicators Tasmania 2003*, viewed 20 May 2013 <http://www.dhhs.tas.gov.au/data/assets/pdf_file/0004/60097/HealthIndicatorsTasmania2003.pdf>
- Dong, X. & Li, Y. 2009, "Standardization of SVG in Implementing WebGIS", *International Conference on Environmental Science and Information Application Technology (ESIAT)*, Wuhan, China, 4-5 July 2009, pp. 534-537.

- Du, Y., Yu, C. & Liu, J. 2009, "A Study of GIS Development Based on KML and Google Earth", *Fifth International Joint Conference on INC, IMS and IDC (NCM '09)*, Seoul, South Korea, 25-27 August 2009, pp. 1581-1585.
- Dunn, J.C. 1973, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", *Cybernetics and Systems*, vol.3, no. 3, pp.32-57.
- Dykes, J., MacEachren, A. & Kraak, M. 2005, "Chapter 33 Evaluating self-organizing maps for geovisualization" in *Exploring Geovisualization*, ed. MJ. Kraak, Elsevier Science, Amsterdam, pp. 629-630.
- Esri 2012, *Understanding our world*, viewed 21 January 2013<
<http://blogs.esri.com/esri/esri-insider/2011/07/18/understanding-our-world/> >.
- Esri Australia 2012, *GIS in the Web Area*, viewed 10 January 2013<
<http://esriaustralia.com.au/> >.
- Everitt, B.S., Landau, S. & Leese, M. 2001, *Cluster Analysis*, 4th edition, Oxford University Press Inc., USA.
- Freedman, C. 2007, *Yahoo! Maps Mashups*, 1st edition, John Wiley & Sons, Inc.
- Fu, C., Wang, Y., Xu, Y. & Li, Q. 2010, "The logistics network system based on the Google Maps API", *International Conference on Logistics Systems and Intelligent Management*, Harbin, China, 9-10 Jan 2010, pp. 1486-1489.
- Fu, P. & Sun, J. 2010, *Web GIS: Principles and Applications*, Esri Press.

- Goujon-Bellec, S., Demoury, C., Guyot-Goubin, A., Hemon, D. & Clavel, J. 2011, "Detection of clusters of a rare disease over a large territory: performance of cluster detection methods", *International Journal of Health Geographics*, vol. 10, no. 1, pp. 53-64.
- Gu, J., Zhou, J. & Chen, X. 2009, "An Enhancement of k-means Clustering Algorithm", *International Conference on Business Intelligence and Financial Engineering (BIFE '09)*, Beijing, China, 24-26 July 2009, pp. 237-240.
- Han, L., Wang, N., Wang, C. & Chi, Y. 2010, "The Research on the WebGIS Application Based on the J2EE Framework and ArcGIS Server", *International Conference on Intelligent Computation Technology and Automation (ICICTA)*, Changsha, China, 11-12 May 2010, pp. 942-945.
- Hong, Y. & Kwong, S. 2009, "Learning Assignment Order of Instances for the Constrained K-Means Clustering Algorithm", *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 39, no. 2, pp. 568-574.
- Hoppner, F., Klawonn, F. & Kruse, R. 1999, *Fuzzy Cluster Analysis*, Wiley Press, New York.
- Huang, Z., Xie, Y., Liu, D. & Hou, L. 2009, "Using Fuzzy C-means Cluster for Histogram-Based Color Image Segmentation", *International Conference on Information Technology and Computer Science (ITCS)*, Kiev, Ukraine, 25-26 July 2009, pp. 597-600

- Hung, C. & Huang, J. 2011, "Mining rules from one-dimensional self-organizing maps", *International Symposium on Innovations in Intelligent Systems and Applications (INISTA)*, Istanbul, Turkey, 15-18 June 2011, pp. 292-295.
- Huo, X.J., Moon, K.S., Lee, S.H., Seung, T.Y. & Kwon, K.R. 2011, "Protecting GIS vector map using the k-means clustering algorithm and odd-even coding", *17th Korea-Japan Joint Workshop on Frontiers of Computer Vision (FCV)*, Ulsan, South Korea, 9-11 February 2011, pp. 1-5.
- Jain, L. & Wu, X. 2007, *The Geospatial Web*, 1st edition, Springer, London.
- Jin, S., Li, Y., Lu, G., Luo, J., Chen, W. & Zheng, X. 2011, "SOM-based hand gesture recognition for virtual interactions", *IEEE International Symposium on VR Innovation (ISVRI)*, Singapore, 19-20 March 2011, pp. 317-322.
- Jin, Y., Benatallah, B., Casati, F. & Daniel, F. 2008, "Understanding Mashup Development", *IEEE Internet Computing*, vol. 12, no. 5, pp. 44-52.
- Khan, S.S. & Ahmad, A. 2004, "Cluster center initialization algorithm for K-means clustering", *Pattern Recognition Letters*, vol. 25, no. 11, pp. 1293-1302.
- Kohonen, T. 1982. "Self-Organized Formation of Topologically Correct Feature Maps". *Biological Cybernetics* 43 (1): 59–69.
- Kohonen, T. 1990, "The self-organizing map", *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464-1480.
- Kohonen, T. 1995 *Self-Organizing Maps*, 1st ed. Berlin Heidelberg, Springer
- Kohonen, T. 1997, *Self-Organizing Map*, 2nd edition Springer, Berlin.

- Kohonen, T. 2001, *Self-Organizing Maps*, 3rd ed. Information Sciences. Berlin, Heidelberg, Springer
- Lai, Y., Orlandic, R., Wai, G.Y. & Kulkarni, S. 2007, "Scalable Clustering for Large High-Dimensional Data Based on Data Summarization", *IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, Honolulu, USA, 1 March – 5 April 2007, pp. 456-461.
- Lebel, A., Pampalon, R. & Villeneuve, P. 2007, "A multi-perspective approach for defining neighbourhood units in the context of a study on health inequalities in the Quebec City region", *International Journal of Health Geographics*, vol. 6, no. 1, pp. 27-31.
- Li, X. 2010, "Research on text clustering algorithm based on improved K-means", *International Conference on Computer Design and Applications (ICCCA)*, Qinhuangdao, China, 25-27 June 2010, pp.V4-573 - V4-576.
- Liu, T., Dai, G., Zhang, L. & Wang, Z. 2009, "Design of K-Means Clustering Algorithm Based on Distance Concentration", *Second International Symposium on Electronic Commerce and Security (ISECS '09)*, Nanchang, China, 22-24 May 2009, pp. 256-259.
- Logeswari, T. & Karnan, M. 2010, "An Enhanced Implementation of Brain Tumor Detection Using Segmentation Based on Soft Computing", *International Conference on Signal Acquisition and Processing (ICSAP '10)*, Bangalore, India, 9-10 February, pp. 243-247.
- MacQueen, J.B. 1967, "Some Methods for Classification and Analysis of MultiVariate Observations", *Proceedings of the fifth Berkeley Symposium*

- on *Mathematical Statistics and Probability*, eds. L.M.L. Cam & J. Neyman, University of California Press, vol. , pp. 281-297.
- Markovic, M. & Kloos, C.D. 2009, "Representing time and location using web mashups", *9th International Conference on Telecommunication in Modern Satellite, Cable, and Broadcasting Services (TELSIKS)*, Nis, Serbia, 7-9 October 2009, pp. 322-325.
- MathWorks 2010, *Self Organizing-Feature Maps*, viewed 15 February 2013 < <http://www.mathworks.com.au/help/nnet/ref/selforgmap.html> >.
- MathWorks, *Matlab*, viewed 5 January 2013 < <http://www.mathworks.com.au/products/matlab/> >
- Negnevitsky, M. 2002, "Artificial Neural Networks" in *Artificial Intelligence*, 1st edition, Addison Wesley, pp. 207-209.
- Pathak, E., Reader, S., Tanner, J. & Casper, M. 2011, "Spatial clustering of non-transported cardiac decedents: the results of a point pattern analysis and an inquiry into social environmental correlates", *International Journal of Health Geographics*, vol. 10, no. 1, pp. 46-56.
- Qiang, X., Cheng, G. & Li, Z. 2010, "A survey of some classic self-organizing maps with incremental learning", *2nd International Conference on Signal Processing Systems (ICSPS)*, Dalian, China, 5-7 July 2010, pp. V1-804 – V1-809.
- Qu, X., Sun, M., Xu, C., Li, J., Liu, K., Xia, J., Huang, Q., Yang, C., Bambacus, M., Xu, Y. & Fay, D. 2011, "A spatial web service client based on Microsoft

- Bing Maps", *19th International Conference on Geoinformatics*, Shanghai, China, 24-26 June 2011, pp. 1-5.
- Rong, L. & Fan, J. 2009, "A Fuzzy C-means Type Clustering Algorithm on Triangular Fuzzy Numbers", *Sixth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD '09)*, Tianjin, China, 14-16 August 2009, pp. 12-16.
- Rousseeuw, K. 1990, *Finding Groups in Data (An Introduction to Cluster Analysis)*, John Wiley & Sons, Inc.
- Santhalakshmi, S. & Bharathi, G. 2011, "Local and spatial information based fuzzy C-Means clustering for color image segmentation", *3rd International Conference on Electronics Computer Technology (ICECT)*, Kanyakumari, India, 8-10 April 2011, pp. 396-400.
- Sathiracheewin, S. & Surapatana, V. 2011, "Daily typical load clustering of residential customers", *8th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, Khon Kaen, Thailand, 17-19 May 2011, pp. 797-800.
- Shi, H., Zhang, J. & Zhang, Y. 2009, "New WebEpi Technologies for Epidemiology Data Geo-visualization Mashups", *Proceedings of the International Conference on Modeling, Simulation and Visualization Methods (MSV'09)*, Las Vegas, USA, 13 – 16 July 2009, pp. 36-41
- Shi, H., Zhang, Y., Zhang, J., Wan, P. & Shaw, K. 2007, "Development of Web-Based Epidemiological Reporting System for Tasmania Utilizing a Google

Maps Add-On", *Digital Image Computing: Techniques and Applications*, Adelaide, Australia, 3-5 December 2007, pp. 118-123.

SOM toolbox 2000, viewed 28 March 2013
 <<http://www.mathworks.com.au/matlabcentral/linkexchange/links/949-som-toolbox>>

Spence, R. 2007, *Information Visualization: Design for Interaction*, Pearson/Prentice Hall.

Srinivasa, K.G., Singh, A., Thomas, A.O., Venugopal, K.R. & Patnaik, L.M. 2005, "Generic Feature Extraction for Classification using Fuzzy C - Means Clustering", *Proceedings of the 3rd International Conference on Intelligent Sensing and Information Processing (ICISIP '05)*, Bangalore, India, 14-17 December 2005, pp. 33-38.

Su, L. 2011, "A method for building thematic map of GIS based on Google Maps API", *19th International Conference on Geoinformatics*, Shanghai, China, 24-26 June 2011, pp. 1-4.

Tan, X., Zhou, M., Zuo, X. & Cui., Y. 2008, "Integration WebGIS with AJAX and XML Based on Google Maps", *First International Conference on Intelligent Networks and Intelligent Systems (ICINIS '08)*, Wuhan, China, 1-3 November 2008, pp. 376-379.

Tryon, R. C. 1939, *Cluster Analysis*. New York: McGraw-Hill.

Vijayabhanu, R. & Radha, V. 2010, "Recognition and elimination of missing values and outliers from an anaerobic wastewater treatment system using K-Means cluster", *3rd International Conference on Advanced Computer*

- Theory and Engineering (ICACTE)*, Chengdu, China, 20-22 August 2010, pp. V4-186 – V4-190.
- Wang, H., Qi, J., Zheng, W. & Wang, M. 2009, "Balance K-Means Algorithm", *International Conference on Computational Intelligence and Software Engineering (CiSE)*, Wuhan, China, 11-13 December 2009, pp. 1-3.
- Wang, R., Zhao, S., Xin, Q. & Liu, A. 2011, "Data interoperability analysis of MIF in ArcGIS environment", *19th International Conference on Geoinformatics*, Shanghai China, 24-26 Jun 2011, pp. 1-4.
- Wang, W., Zhang, Y., Li, Y. & Zhang, X. 2006, "The Global Fuzzy C-Means Clustering Algorithm", *The Sixth World Congress on Intelligent Control and Automation (WCICA)*, Dalian, China, pp. 3604-3607.
- Wang, Z. 2010, "Comparison of Four Kinds of Fuzzy C-Means Clustering Methods", *Third International Symposium on Information Processing (ISIP)*, Qingdao China, 15-17 October 2010, pp. 563-566.
- Wang, S., Chung, K., Deng, Z. & Hu, D. 2007, "Robust fuzzy clustering neural network based on ε -insensitive loss function", *Applied Soft Computing*, vol. 7, no. 2, pp. 577-584.
- Wilson, H.G., Boots, B. & Millward, A.A. 2002, "A comparison of hierarchical and partitional clustering techniques for multispectral image classification", *IEEE International Geoscience and Remote Sensing Symposium (IGARSS '02)*, 24-28 June 2002, vol.3, pp. 1624-1626.

- Windham, M.P. 1982, "Cluster Validity for the Fuzzy c-Means Clustering Algorithm", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-4, no. 4, pp. 357-363.
- Wood, J., Dykes, J., Slingsby, A. & Clarke, K. 2007, "Interactive Visual Exploration of a Large Spatio-temporal Dataset: Reflections on a Geovisualization Mashup", *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1176-1183.
- Wu, X. & Yao, C. 2010, "Application of improved K-means clustering algorithm in transit data collection", *3rd International Conference on Biomedical Engineering and Informatics (BMEI)*, Yantai, China, 16-18 October 2010, pp. 3028-3030.
- Yan, D., Jiang, S., Zhang, L. & Li, Y. 2010, "Study of WebGIS architecture based on GML and SVG", *2nd International Conference on Information Science and Engineering (ICISE)*, Hangzhou, China, 4-6 December 2010, pp. 4056-4061.
- Yang, L. & Deng, M. 2010, "Based on k-Means and Fuzzy k-Means Algorithm Classification of Precipitation", *International Symposium on Computational Intelligence and Design (ISCID)*, Hangzhou, China, 29-31 October 2010, pp. 218-221.
- Yeh, Y.C. & Lin, H.J. 2010, "Cardiac arrhythmia diagnosis Method using Fuzzy C-Means algorithm on ECG signals", *International Symposium on Computer Communication Control and Automation (3CA)*, Tianan, Taiwan, 5-7 May 2010, pp. 272-275.

- Yin, K. & Gang, L. 2010, "Fault Pattern Recognition of Thermodynamic System Based on SOM", *International Conference on Electrical and Control Engineering (ICECE)*, Wuhan, China, 25-27 June 2010, pp. 3742-3745.
- Yu, G., Soh, L.K. & Bond, A. 2005, "K-means clustering with multiresolution peak detection", *IEEE International Conference on Electro Information Technology*, Lincoln, NE, USA, 22-25 May 2005, pp. 1-6.
- Zhang, J. & Shi, H. 2007, "Geospatial Visualization using Google Maps: A Case Study on Conference Presenters", *Proceedings of the Second International Multi-Symposiums on Computer and Computational Sciences* IEEE Computer Society, Washington, DC, USA, 13-15 August 2007, pp. 472-476.
- Zhang, J., Shi, H. & Zhang, Y. 2009, "Self-Organizing Map Methodology and Google Maps Services for Geographical Epidemiology Mapping", *Digital Image Computing: Techniques and Applications (DICTA '09)*, Melbourne, Australia, 1-3 December 2009, pp. 229-235.
- Zhang, J., Shi, H. & Zhang, Y. 2007, "Web Mapping For Location Based Decision Making", *International Conference on Communications System, Networks and Application (CSNA 2007)*, Beijing, China, 08-10 October 2007, pp. 220-224.
- Zhang, J., Shi, H. & Zhang, Y. 2009, "Geo-Mashups Automation for Web-based Epidemiological Reporting System", *Proceedings of the 2009 International Conference on Modeling, Simulation & Visualization Methods (MSV'09)*, Las Vegas Nevada, USA, July 13-16 2009, CSREA Press , pp. 56-61.

- Zhou, H. & Liu, Y. 2006, "3D Modelling from Multi-view Registered Range Images Using K-means Clustering", *IEEE International Conference on Industrial Technology (ICIT 2006)*, Mumbai, India, 15-17 December 2006, pp. 722-727.
- Zhu, G. & Zhu, X. 2010a, "The Growing Self-organizing Map for Clustering Algorithms in Programming Codes", *International Conference on Artificial Intelligence and Computational Intelligence (AICI)*, Sanya, China, 23-24 October 2010, pp. 178-182.
- Zhu, R., Liu, Y., Jiang, H. & Yin, Z. 2011, "Visualization of weather-induced disaster warning information system using Google Earth API based on Mashup", *International Conference on Multimedia Technology (ICMT)*, Hangzhou, China, 26-28 July 2011, pp. 3789-3793.
- Zhu, X. & Zhu, G. 2010b, "Self-Organizing Map for Clustering Algorithms in Programming Codes", *Third International Conference on Business Intelligence and Financial Engineering (BIFE)*, Hongkong, China, 13-15 August 2010, pp. 24-27.

Appendices

The attached CD-ROM contains epidemiological experimental data and some program source code. There are three folders: **Demo**, **WebEpi** and **Clustering**.

The **Demo** folder includes WebEpi demonstration files on Google Maps and Google Earth. The **WebEpi** folder includes WebEpi clustering automation programs in MATLAB and the WebGIS API for Google Maps. The **Clustering** folder includes WebEpi clustering experimental results and clustering plotting programs.

A. Demonstration Files

In the **Demo** folder, two sets of demonstration files are stored in the Google Maps and Google Earth folders.

If Google Maps does not change its API configuration requirement, the live demo of WebEpi on Google Maps can be found at:

<http://www.cyberdesign.com.au/webepi/>

A.1 Google Maps visualisation

Inside the Google Maps folder there is webpage called: *index.html*. Before running the webpage please make sure to enable the Internet Explorer *Active X* control and *initialize and script ActiveX controls not marked as safe for scripting* as shown in Fig A.1 and A.2. Then double click on the *index.html* web page in folder the Google Maps, as shown in Fig A.3. The visualisation result will appear as shown in Fig A.4

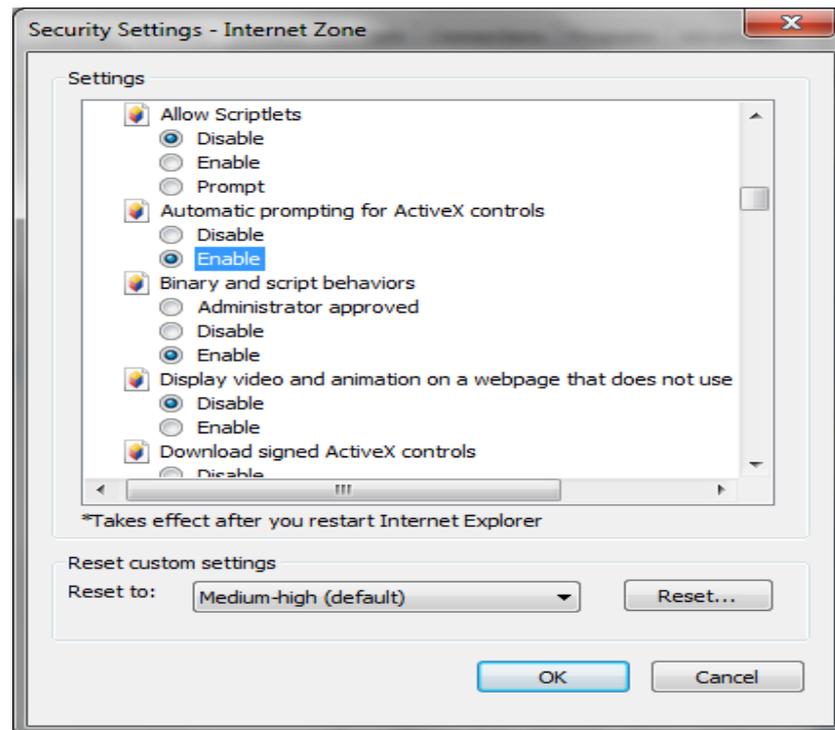


Fig A.1 Security settings(1)

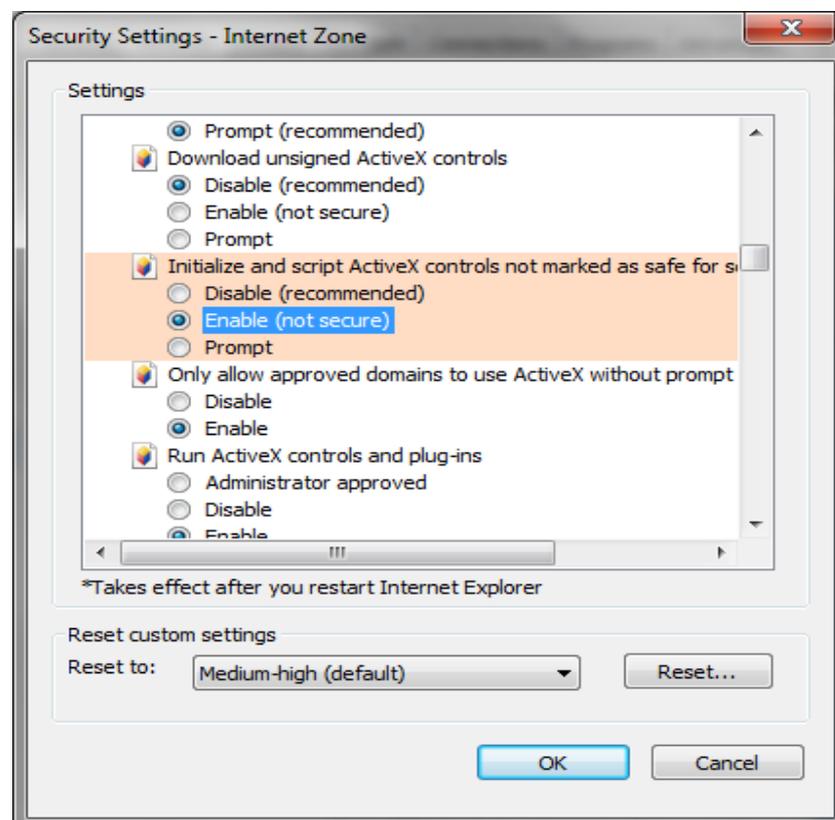


Fig A.2 Security settings(2)

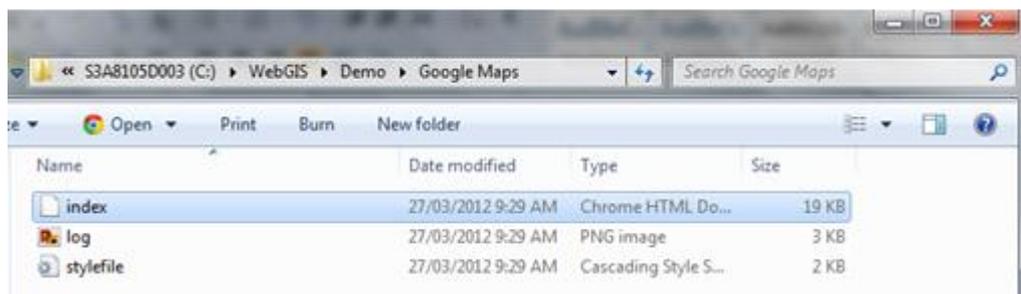


Fig A.3 File location

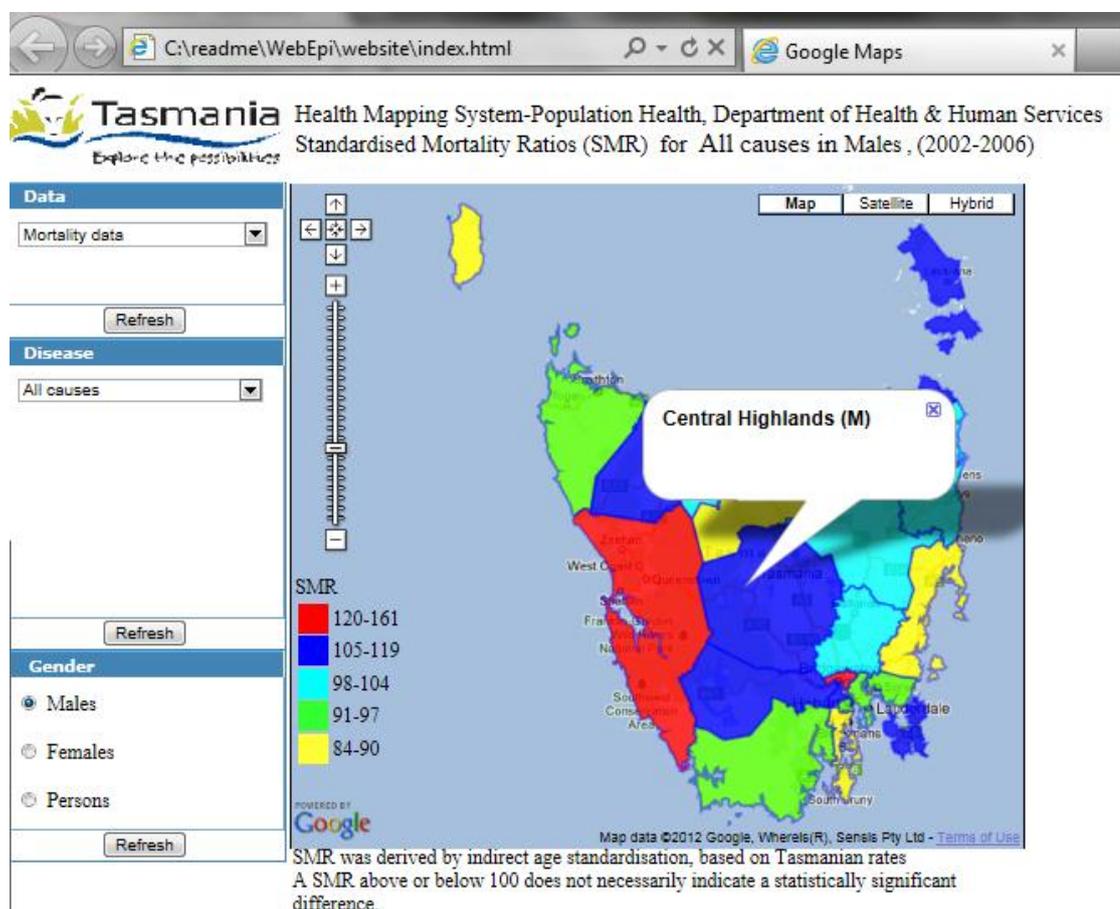


Fig A.4 WebEpi mapping

A.2 Google Earth visualisation

Before running the demonstration of WebEpi geospatial layer files on Google Earth, Google Earth has to be installed by running GoogleEarthSetup.exe, in

the Google Earth folder as shown in Fig A. 5. After the installation, double click on any KML file in the Google Earth subfolder, the WebEpi geospatial layers can be visualised on Google Earth and the results will appear as shown in Fig A.6

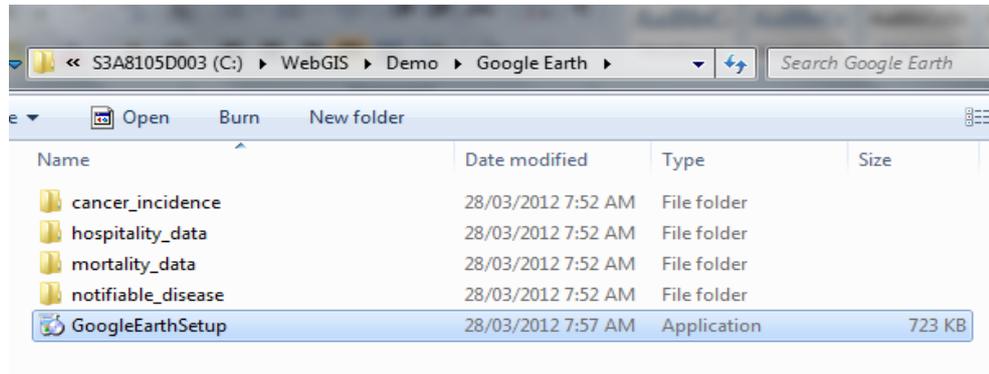


Fig A.5 GoogleEarth installation

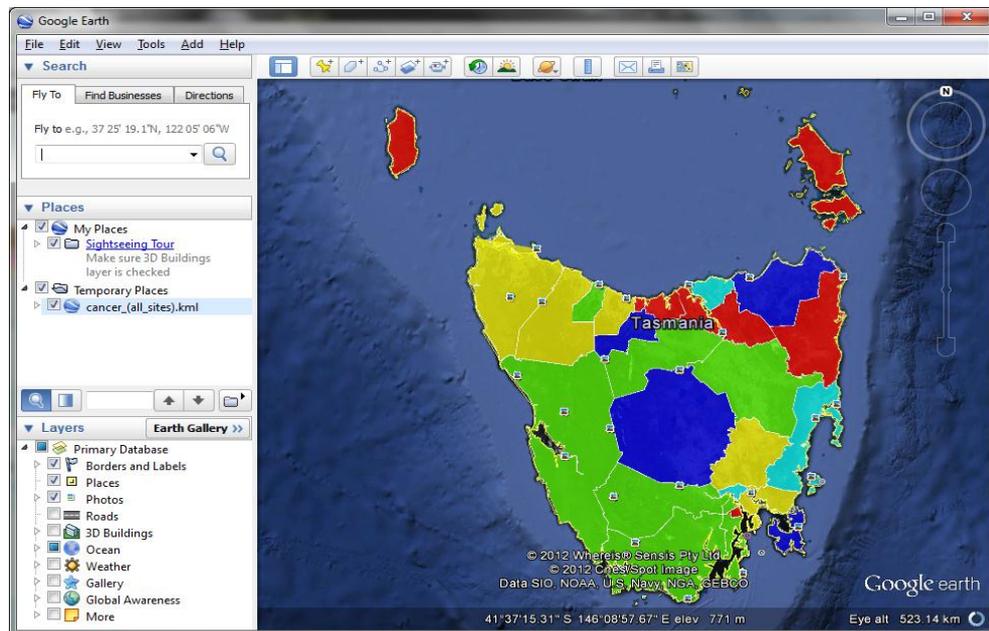


Fig A.6 GoogleEarth mapping

B. WebEpi Guideline

In the CD-ROM, WebEpi source code for clustering analysis and geospatial visualisation are saved in the WebEpi folder. There are four steps in running the WebEpi MATLAB program.

Step 1: copy CD-ROM *WebGIS* folder to C: drive root directory.

Step 2: copy the MATLAB library. Before starting the experiment of epidemiological data, the tools library for MATLAB has to be setup. The library folder is located at:

C:\WebGIS\WebEpi\Matlab\lib folder

as shown in Fig B.1

The *nnet* folder stores the SOM clustering algorithm, the *fuzzy* folder stores the FCM clustering algorithm and the *stats* folder stores the K-means clustering algorithm as shown in Fig B.2. Copy these folders to the MATLAB tools directory.

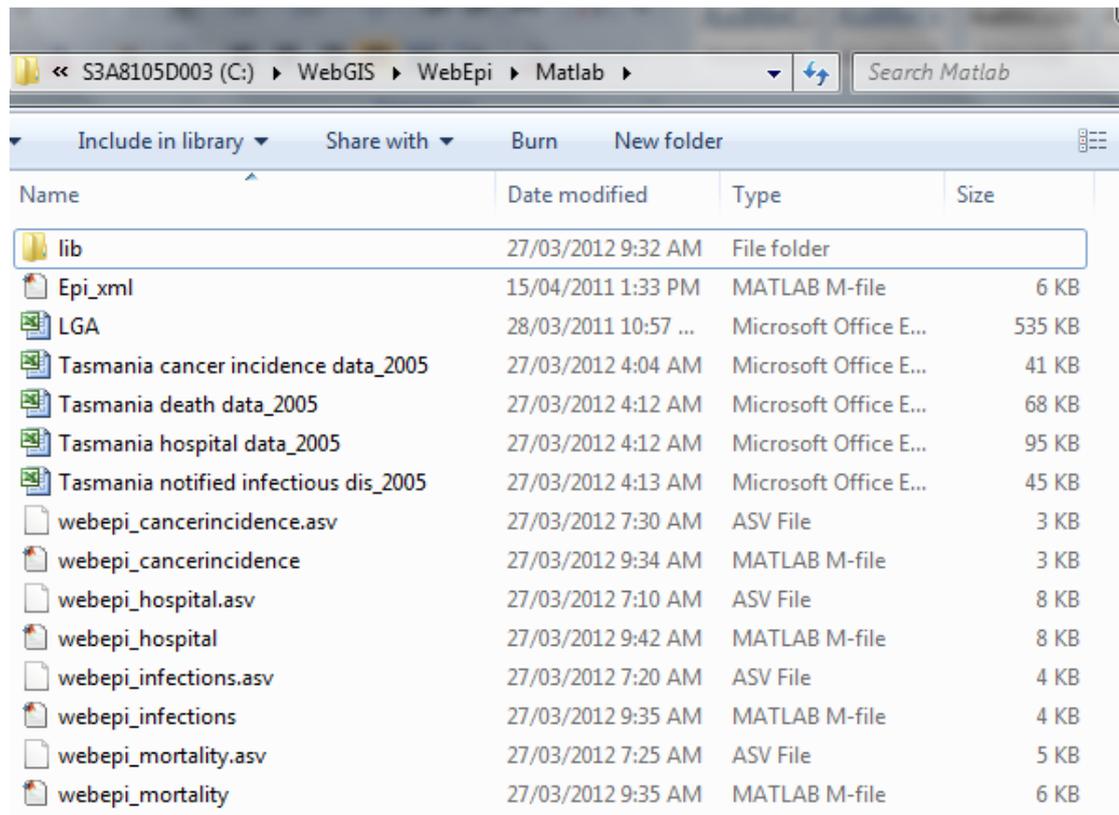


Fig B.1 WebEpi file location

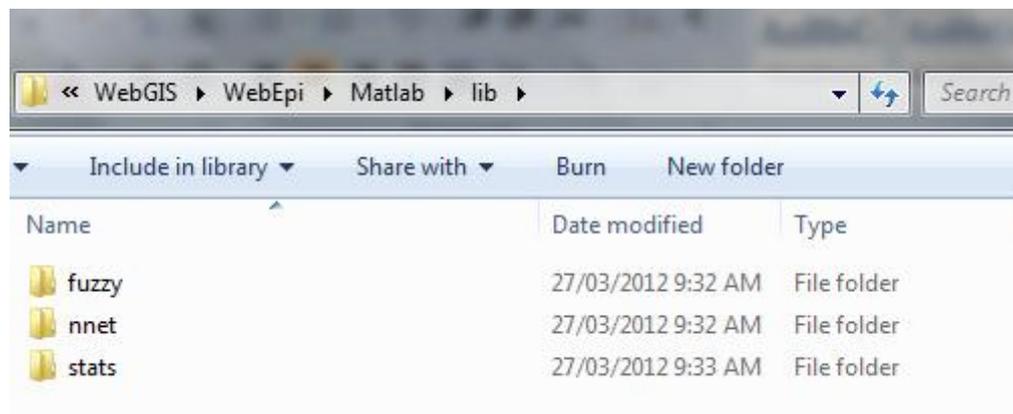


Fig B.2 WebEpi clustering location

Step 3: set the current directory for the MATLAB environment. These functionalities are coded in MATLAB. Set the current directory to C:\WebGIS\WebEpi\Matlab folder in MATLAB as shown in Fig B.3.

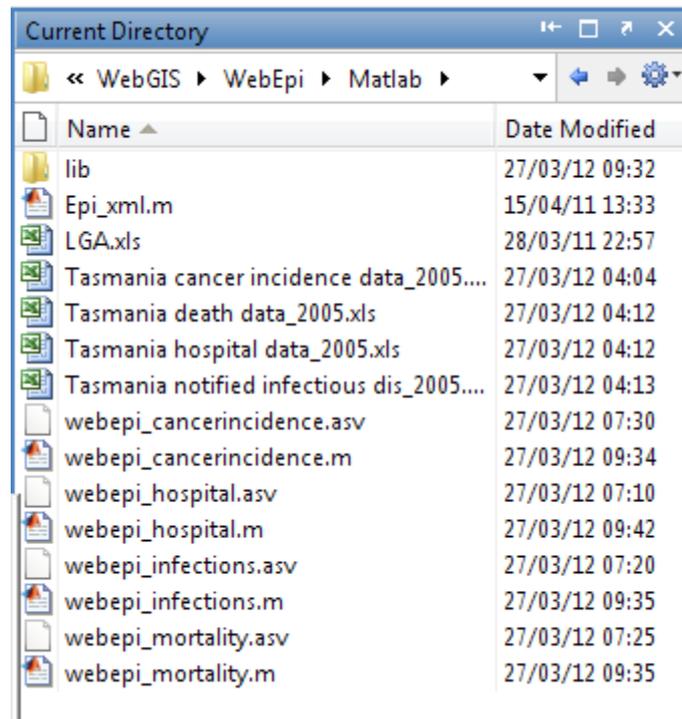


Fig B.3 WebEpi MATLAB

The MATLAB file *Epi_xml.m* executes the Geo-Mashups and geospatial layer customisation. It is a combination of clustering results and LGA geospatial coordinates. The geospatial layers are also created in this process. MATLAB file: *webepi_cancerincidence.m* executes the K-means clustering analysis and Geo-Mashups of category Cancer Incidence data; *webepi_hospital.m* executes the K-means clustering analysis and Geo-Mashups of category Hospitalisation data. *webepi_infections.m* conducts the K-means clustering analysis and Geo-Mashups of category Epi Notified Infectious data. *webepi_mortality.m* conducts the K-means clustering analysis and Geo-Mashups of category Death data.

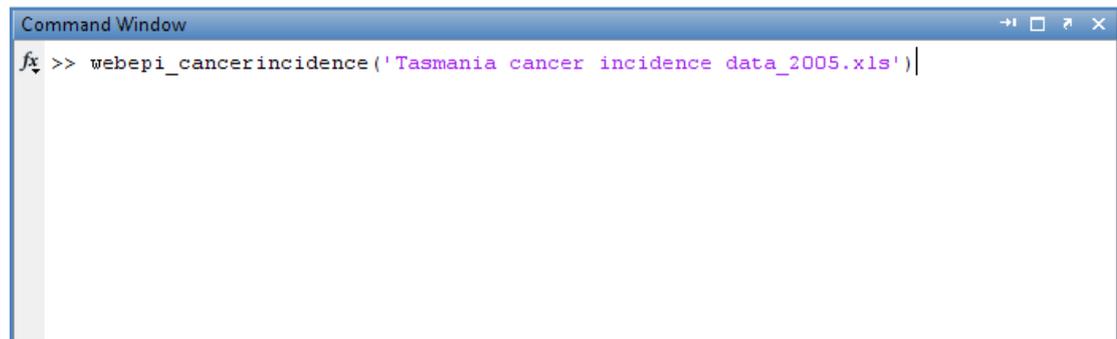
Step 4: input the WebEpi MATLAB program command. In the MATLAB command window type in the flowing commands as shown in Fig B.4 and the WebEpi clustering and Geo-Mashups are conducted.

```
webepi_cancerincidence('Tasmania cancer incidence data_2005.xls')
```

webepi_hospital('Tasmania hospital data_2005.xls')

webepi_infections('Tasmania notified infectious dis_2005.xls')

webepi_mortality('Tasmania death data_2005.xls')



```
Command Window
fx >> webepi_cancerincidence('Tasmania cancer incidence data_2005.xls')|
```

Fig B.4 MATLAB code(1)

The other two clustering algorithms, i.e. SOM, and Fuzzy c-means, are also available by changing one line of source code:

For example in Fig B.5, the highlighted code: *[IDX]=kmeans(cancerall,5);*

can be changed to

[IDX] = newsom(cancerall,[1 5]); % for SOM

[IDX] = fcm(cancerall,5); % for Fuzzy C-Means

```

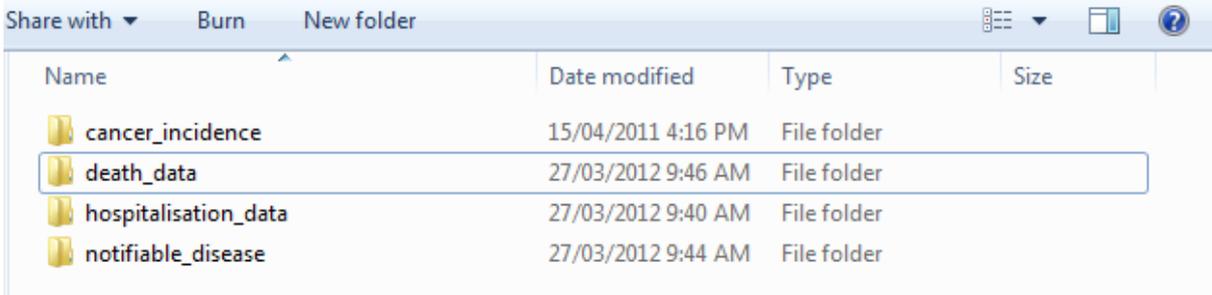
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% the following code is for epidemiologicla data clustering analysis and
% converting clustering analysis to KML file
% SOM: [IDX]=newsom(cancerall,[1 5]);
% Fuzzy C-means:[IDX]=fcm(cancerall,5);
% K-means:[IDX]=kmeans(cancerall,5);

function xmlfile=Cancer_all(filename,epi_cancerall,LGA,LGA_GIS)
xmlname=[filename,'\cancer_(all_sites)'];
cancerall=epi_cancerall(1:29);
[IDX]=kmeans(cancerall,5);
xmlfile=Epi_xml(xmlname,IDX,LGA,LGA_GIS);
end

```

Fig B.5 MATLAB code (2)

Once all the functions have been executed, the WebEpi geospatial layer KML files will have been created in the directory: C:\Webepi\ which is shown in Fig B.6



Name	Date modified	Type	Size
cancer_incidence	15/04/2011 4:16 PM	File folder	
death_data	27/03/2012 9:46 AM	File folder	
hospitalisation_data	27/03/2012 9:40 AM	File folder	
notifiable_disease	27/03/2012 9:44 AM	File folder	

Fig B.6 Mapping file location

The layer files are organised according to disease categories, gender, disease types, as shown in Fig B.7.

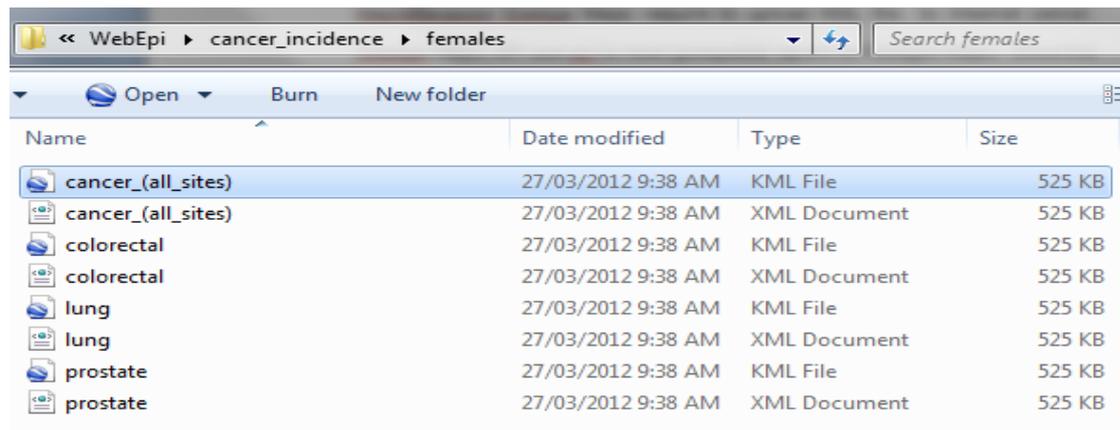


Fig B.7 KML file

The KML file can be directly opened by Google Earth not Google Maps because Google Maps requires uploading the KML file to the Internet server. The Google Maps API uses a URL to load geospatial layers on Google Maps. Once Google Maps is installed just double click on the KML file as shown in Fig B.7. For the process of browsing KML files see section A.2

WebEpi Google Maps visualisation can be seen by browsing website index.html which is located in WebGIS\Webepi\WebGISAPI. The WebEpi KML uploaded to the Internet server was collected in 2006, and is different from the experimental data. The WebEpi website can be uploaded to the Internet server by configuring the Google Maps API key in the source code of index.html.

```
<script
src="http://maps.google.com/maps?file=api&v=2&
key=AlzaSyBI91vulxVFncpsRHCj8bGG4tsTbgw6XOE "
type="text/javascript">
</script>
```

The KML server route directory can also be updated by the source code.

```
var kmlfileroot="http://www.cyberdesign.com.au/WebEpi/";
```

The process of browsing index.html is shown in Section A.1

C. Clustering Algorithms

Inside the **Clustering** folder, there are plotting programs stored for three clustering algorithms. MATLAB program *create_c_net.m* is used to plot the SOM clustering results. *cmeans.m* is for plotting the FCM clustering results and *kmeans_epi.m* allows to plot the k-means clustering results.

D. CD-ROM