

**THE APPLICATION OF USING HIERARCHICAL CLUSTERING
TO ADMISSION SCORES**

RAJJAKRIJ WASUBHADDARADILOK

**A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE
(TECHNOLOGY OF INFORMATION SYSTEM MANAGEMENT)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY**

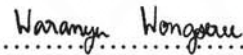
2012

COPYRIGHT OF MAHIDOL UNIVERSITY

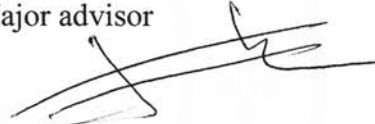
Thesis
entitled
**THE APPLICATION OF USING HIERARCHICAL CLUSTERING
TO ADMISSION SCORES**



Mr. Rajjakrij Wasubhaddaradilok
Candidate



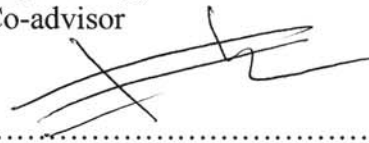
Lect. Waranyu Wongseree,
Ph.D (Electrical Engineering)
Major advisor

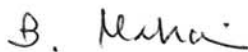


Lect. Supaporn Kiattisin,
Ph.D. (Electrical and Computer
Engineering)
Co-advisor



Asst. Prof. Bunlue Emaruechi,
Ph.D. (Environmental Systems
Engineering)
Co-advisor





Prof. Banchong Mahaisavariya,
M.D., Dip Thai Board of Orthopedics
Dean
Faculty of Graduate Studies
Mahidol University

Lect. Supaporn Kiattisin,
Ph.D. (Electrical and Computer
Engineering)
Program Director
Master of Science Program in
Technology of Information System
Management
Faculty of Engineering
Mahidol University

Thesis
entitled
**THE APPLICATION OF USING HIERARCHICAL CLUSTERING
TO ADMISSION SCORES**

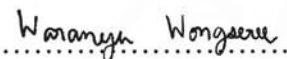
was submitted to the Faculty of Graduate Studies, Mahidol University
for the degree of Master of Science
(Technology of Information System Management)

on
November 26, 2012

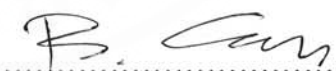

.....
Mr. Rajjakrij Wasubhaddaradilok
Candidate


.....
Assist. Prof. Adisorn Leelasantitham, Ph.D.,
(Electrical Engineering)
Chair

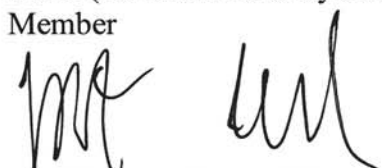

.....
Theera Piroonratana,
Ph.D. (Electrical engineering)
Member


.....
Lect. Waranyu Wongseree,
Ph.D. (Electrical Engineering)
Member


.....
Lect. Supaporn Kiattisin,
Ph.D. (Electrical and Computer
Engineering)
Member


.....
Asst. Prof. Bunlue Emaruechi,
Ph.D. (Environmental Systems Engineering)
Member


.....
Prof. Banchong Mahaisavariya,
M.D., Dip. Thai Board of Orthopedics
Dean
Faculty of Graduate Studies
Mahidol University


.....
Lect. Worawit Israngkul
M.S. (Technical Management)
Dean
Faculty of Engineering
Mahidol University

ACKNOWLEDGEMENTS

The success of this thesis would not have been succeeded without the support, hard work and efforts of a number of people. First, I would like to express my deepest sincere gratitude to my major advisor, Dr. Waranyu Wongseree for his supervision and valuable guidance to this research.

I also would like to express my gratitude to Dr. Supaporn Kiattisin and Assist. Prof. Bunlue Emaruechi, my co-advisor for abundant helps with various useful suggestions and methodologies. I sincerely thank to Dr.Theera Piroonratana, the external examiner of the thesis defense and Asst.Prof.Adisorn Leelasantitham, the chairman of thesis defense, for their kindness in significant suggestions and generous comments.

My appreciation to Khun Cholticha Boondireke, Computer Scientist, Division of Information Technology (MUIT), Office of the President, Mahidol University for provided the reports of University Admission Score for the tests of GAT, O-NET, and PAT on the years of 2009, 2010, 2011 and 2012.

Besides, heartiest thank to my friends and staffs in Master of Science Program in Technology of Information System Management, Mahidol University.

Finally, this thesis would not have been completed without great understanding and support of my family, especially my deepest gratitude parents for entirely care, love and encouragement through my life. The usefulness on this thesis, I dedicate to my father, my mother and all of my teachers.

Rajjakrij Wasubhaddaradilok

THE APPLICATION OF USING HIERARCHICAL CLUSTERING TO ADMISSION SCORES

RAJJAKRIJ WASUBHADDARADILOK 5136028 EGTI/M

M.SC. (TECHNOLOGY OF INFORMATION SYSTEM MANAGEMENT)

THESIS ADVISORY COMMITTEE: WARANYU WONGSEREE, Ph.D.
(ELECTRICAL ENGINEERING), SUPAPORN KIATTISIN, Ph.D. (ELECTRICAL
AND COMPUTER ENGINEERING), BUNLUE EMARUECHI, Ph.D.
(ENVIRONMENTAL SYSTEMS ENGINEERING)

ABSTRACT

The Central University Admissions System of Thailand has required students to take O-NET and GAT tests for further education in university. Analytical statistics used were factor analysis and cluster analysis. Factor analysis of the weight value of 16 exam subjects and 36 departments of Mahidol University was initially utilized. This research found that the variable of weight value could be decreased by factor analysis into 4 factors which were: core subjects, learning subjects, professional subjects and advanced English, and advanced Mathematics. Then that result could be arranged into 6 faculties of Mahidol University: Administration, Engineering, Science, Health Science & Medicine, Sport Science & Medical Technology and Fine Arts. Additional analysis was done by factor analysis and cluster analysis of the O-NET and GAT data. This experiment used data from O-NET and GAT tests from the academic year 2009, 2010, 2011, and 2012. Clustering analysis by method of K- mean clustering could define clustering into 4 groups. Research found variables of exam subjects in all 4 clusters in which similar contents were arranged into the same cluster.

KEY WORDS: UNIVERSITY ADMISSSION SCORE/ FACTOR ANALYSIS/
CLUSTER ANALYSIS

99 pages

การประยุกต์ใช้คลัสเตอร์แบบลำดับชั้นกับคะแนนเอ็นทรานส์

THE APPLICATION OF USING HIERARCHICAL CLUSTERING TO ADMISSION SCORES

รัชกฤต วสุภัทรคติล 5136028 EGTI/M

ว.ทม. (เทคโนโลยีการจัดการระบบสารสนเทศ)

คณะกรรมการที่ปรึกษาสารนิพนธ์: วรรณู วงษ์เสรี, Ph.D. (ELECTRICAL ENGINEERING), สุภากรณ กียรติสิน, Ph.D. (ELECTRICAL AND COMPUTER ENGINEERING), บัณฑิต เอเมะรุจิ, Ph.D. (ENVIRONMENTAL SYSTEMS ENGINEERING)

บทคัดย่อ

การสอบเข้าศึกษาต่อในระดับอุดมศึกษาผ่านระบบส่วนกลาง กำหนดให้สอบด้วยวิชาพื้นฐาน(O-NET) และ GAT โดยนักเรียนจะต้องทำการสอบคัดเลือกและเลือกคณะจัดอันดับ จะทำการวิเคราะห์ปัจจัยค่าน้ำหนักของวิชาที่ใช้สอบในการสอบคัดเลือกเข้าศึกษาต่อ ในระดับอุดมศึกษา และจัดกลุ่มคณะและสาขาวิชาของมหาวิทยาลัยมหิดลจำนวน 36 สาขาวิชา โดยใช้ข้อมูลค่าน้ำหนักของวิชาที่ใช้สอบคัดเลือกจำนวน 16 วิชา สถิติที่ใช้ในการวิเคราะห์ คือ การวิเคราะห์ปัจจัยและการวิเคราะห์กลุ่ม ผลการวิจัย พบว่า การวิเคราะห์ปัจจัยจะลดจำนวนตัวแปรค่าน้ำหนักเหลือเพียง 4 ปัจจัย คือ กลุ่มวิชาพื้นฐานแกนหลัก กลุ่มสาระการเรียนรู้ กลุ่มวิชาเฉพาะทางและวิชาภาษาอังกฤษขั้นสูง และ กลุ่มวิชาคณิตศาสตร์ขั้นสูง ผลลัพธ์ที่ได้นำมาจัดกลุ่มสาขาวิชาของมหาวิทยาลัยมหิดลได้จำนวน 6 กลุ่ม คือ กลุ่มสาขาวิทยาลัยการจัดการ สาขาวิศวกรรมศาสตร์ สาขาวิทยาศาสตร์สาขาวิทยาศาสตร์การแพทย์และแพทยศาสตร์ สาขาวิทยาศาสตร์การกีฬาและเทคโนโลยีการศึกษา การแพทย์ และ สาขาศิลปศาสตร์ เมื่อทำการวิเคราะห์ข้อมูลในส่วนแรกเสร็จจะนำข้อมูลคะแนน O-NET และ GAT ซึ่งเป็นคะแนนในการสอบเข้ามหาวิทยาลัยจริงของนักศึกษามาทำการวิเคราะห์เพิ่มเติมโดยใช้การวิเคราะห์ปัจจัยและการวิเคราะห์กลุ่ม พบว่า สามารถสกัดปัจจัยคะแนน O-NET และ GAT ของนักศึกษาในปีการศึกษา 2552, 2553, 2554 และ 2555 ได้ 2 ปัจจัย ส่วนการวิเคราะห์จัดกลุ่มใช้วิธีการ K-mean clustering ให้จำนวนกลุ่มในการจัดกลุ่มเท่ากับ 4 กลุ่ม พบว่า ในแต่ละกลุ่มจะมีตัวแปรวิชาที่ใช้สอบที่มีลักษณะคล้ายคลึงกันถูกจัดอยู่ในกลุ่มเดียวกัน

CONTENTS

	Page
ACKNOWLEDGEMENTS	iii
ABSTRACT (ENGLISH)	iv
ABSTRACT (THAI)	v
LIST OF TABLES	viii
LIST OF FIGURES	x
LIST OF EQUATIONS	xi
CHAPTER I INTRODUCTION	1
1.1 Background and statement of problems	1
1.2 Objectives of study	8
1.3 Scopes of study	8
1.4 Hypothesis	8
1.5 Thesis organization	8
CHAPTER II LITERATURES REVIEWS	10
2.1 Basic knowledge of admission examination	10
2.2 Profile of O-NET, GAT, PAT and admission	13
2.3 The 3 rd GAT/PAT testing on October of academic year 2010	15
2.4 Contents of 3 rd GAT/ PAT on academic year 2010	16
2.5 Related literatures	18
CHAPTER III MATERIALS AND METHODS	21
3.1 Related theories	21
3.2 Research methodology	48
3.3 Materials and research tools	50
3.4 Schedule of research	50

CONTENTS (cont.)

	Page
CHAPTER IV RESULT	51
4.1 Result of weight value of Mahidol University	51
4.2 Result of Admission score (New Data Sets) of Mahidol University	58
CHAPTER V DISCUSSION	75
5.1 Data Collection	75
5.2 Factor Analysis	75
5.3 Cluster Analysis	76
5.4 Tools and limitations in the research	77
CHAPTER VI CONCLUSION AND RECOMMENDATIONS	78
6.1 Conclusion	78
6.2 Recommendation for factor analysis	79
6.3 Recommendation for cluster analysis	79
6.4 Future work	79
REFERENCES	80
APPENDICES	82
Appendix A	83
Appendix B	88
Appendix C	97
BIOGRAPHY	99

LIST OF TABLES

Table	Page
3.1 Comparison between Rotated and Unrotated Factor Loadings	28
3.2 Basic K-means Algorithm	41
3.3 Table of notation	43
3.4 K-means	44
3.5 Schedule of research	50
4.1 Descriptive Statistics from factor analysis	51
4.2 Rotated Component Matrix from factor analysis	52
4.3 Means and Standard deviation of 16 variables	53
4.4 Total Variance Explained (Initial Eigen values)	54
4.5 Total Variance Explained (Rotation Sums of Squared Loadings)	54
4.6 Rotated Component Matrix	55
4.7 Component Score Coefficient Matrix	55
4.8 The 1 st factor: Core subjects	56
4.9 The 2 nd factor: Learning subjects (GPA)	56
4.10 The 3 rd factor: Professional subjects and Advanced English	56
4.11 The 4 th factor: Advanced Mathematics	56
4.12 Descriptive Statistics for factor analysis (academic year 2010)	58
4.13 Correlation Matrix for factor analysis (academic year 2010)	59
4.14 Total Variance Explained for factor analysis (academic year 2010)	59
4.15 Rotated Component Matrix for factor analysis (academic year 2010)	60
4.16 Descriptive Statistics for factor analysis (academic year 2011)	61
4.17 Correlation Matrix for factor analysis (academic year 2011)	62
4.18 Total Variance Explained for factor analysis (academic year 2011)	62
4.19 Rotated Component Matrix for factor analysis (academic year 2011)	63
4.20 Descriptive Statistics for factor analysis (academic year 2012)	64
4.21 Correlation Matrix for factor analysis (academic year 2012)	65

LIST OF TABLES (cont.)

Table	Page
4.22 Total Variance Explained for factor analysis (academic year 2012)	66
4.23 Rotated Component Matrix for factor analysis (academic year 2012)	67
4.24 Final Cluster Centers for cluster analysis (academic year 2010)	68
4.25 Distances between Final Cluster Centers for cluster analysis (year 2010)	68
4.26 ANOVA for cluster analysis (academic year 2010)	69
4.27 Final Cluster Centers by K-means algorithm (academic year 2011)	70
4.28 Distances between Final Cluster Centers by K-means algorithm (year 2011)	71
4.29 ANOVA by K-means algorithm (academic year 2011)	71
4.30 Final Cluster Centers by K-means algorithm (academic year 2012)	72
4.31 ANOVA by K-means algorithm (academic year 2012)	73

LIST OF FIGURES

Figure	Page
3.1 Stages 1-3 in the Factor Analysis Decision Diagram	23
3.2 Stages 4-7 in the Factor Analysis Decision Diagram	24
3.3 Orthogonal Factor Rotation	26
3.4 Oblique Factor Rotation	27
3.5 Choosing the number of Factor axes	31
3.6 Example of VARIMAX rotation	32
3.7 Clustering Methodology	34
3.8 Taxonomy of clustering approaches	36
3.9 Example of a hierarchical clustering in a dendrogram	37
3.10 Example of a hierarchical clustering in Venn diagram	37
3.11 Using the K-means algorithm to find 3 clusters in sample data	42
3.12 Three optional and non-optimal clusters	45
3.13 Poor starting centroids for K-means	46
3.14 Two pairs of clusters with a pair of initial centroids within each pair of clusters	46
3.15 Two pairs of clusters with more or fewer than two initial centroids within a pair of clusters	47
3.16 Research methodology	49
4.1 Scree Plot by factor analysis	52
4.2 Relation of Eigen value for each factor and Component Number	54
4.3 Hierarchical Cluster Analysis	57
4.4 Scree Plot for factor analysis (academic year 2010)	60
4.5 Scree Plot for factor analysis (academic year 2011)	63
4.6 Scree Plot for factor analysis (academic year 2012)	66

LIST OF EQUATIONS

Equation	Page
3.1 Factor Analysis	22
3.2 Factor Analysis	22
3.3 Principal Component Analysis	29
3.4 Principal Component Analysis	29
3.5 Euclidean distance	29
3.6 Inverse of Variances	29
3.7 Covariance $\text{cov}(X, Y)$	30
3.8 Covariance $\text{cov}(X, Y)$	30
3.9 Covariance Matrix	30
3.10 Covariance Matrix	31
3.11 Cluster proximity	40
3.12 Sum of the Squared Error (SSE)	43
3.13 Centroid of cluster C_i	44
3.14 Euclidean distance	48

CHAPTER I

INTRODUCTION

1.1 Background and statement of problems

It would describe about knowledgeable history of university entrance exam on the system of Central University Admissions System (CUAS).

1.1.1 The history of university entrance exam on the system of Central University Admissions System (CUAS) before academic year 1961 – academic year 1999.

- **Before academic year 1961** Individual universities processed their own system independently.

- **Academic year 1961** Kasetsart University and University of Medical Science organized jointly. Office of the Education Council was coordinator.

- **Academic year 1962** Five of currently established universities as Chulalongkorn University, Thammasat University, University of Medical Science, Kasetsart University and Silpakorn University organized jointly.

This examination was solved the problem of qualified students waived the right to study in university and also limited the high number of alternates who pass several exams. Furthermore, exam applicants had to test several exams from several universities then it increased much more expenses and high number of applicants.

- **Academic year 1966** The Cabinet approved to the proposal of Chulalongkorn University that all universities had processed the university examination independently. But several problems occurred. Then universities had to postpone the new semester because of the problem of attending several interviews.

- **Academic year 1967** The Cabinet approved the proposal of The Education Council that all universities went back to use the centralized exam procedure.

- **Academic year 1973** Ministry of University Affairs transferred all responsibilities from The Education Council since that time until now.

The university examination at that time allowed examination applicants to apply for six faculties or universities and at the same time applied for examination subjects also. After processed this Centralized Entrance Examination for a while then problems were occurred as below:

1. The originally examination negatively affected to the educational system of upper secondary school student because several of students ignored subjects which not be tested in the university examination. Their goal of studying was passed the university examination or passed the required examination subjects. Then the result was all good students tried to complete the Open Examination to meet all requirements of graduation on upper secondary school student without attending in school class and aiming to pass entrance exam. Parents or students liked the reason for saving costs of education and time. Then the result was premature students were occurred when studying in university and also the poor problem of educational system in upper secondary school student.

2. The originally examination performed students who less academic proficiency to faculty because examination were tested only on compulsory subjects.

3. The numbers of examination applicants were increasing gradually.

4. The examination was tremendous tense to applicants and their parents because they required choosing faculty/university that they wanted to study in at the same time of submitting the applicant.

5. Time limitation is occurred because the examination was processed only at the end of semester.

1.1.2 The new system of university examination (1999-2005)

Ministry of University Affairs appointed committee to consider and rectify the system of university examination (Document No.1). The committee was consisted of Representative of Department of Curriculum and Instruction, Representative of General Education, Ministry of Education, Representative of the Education Commission and various experts from universities. They jointed the committee and

informed the new system of university examination for Ministry of University Affairs and then approved since academic year 1999.

The objectives of rectify the system of examination were 2 objectives as:

1. Universities/institutes were able to get literacy students for meeting requirement of faculty.
2. To promote more philosophy and objective in the class procedure of upper secondary school student.

In order that, the new system of examination would consider applicants from following factors:

1. The sum of Grade Point Average from the entire course of Secondary School Grades 4-6 or equivalent was weighed for 10%.
2. Scores of examination on Compulsory subjects and Professional subjects (or called that Knowledge Test) was weighed for 90%.

The examination on Professional subjects and Compulsory subjects were tested 2 times annually and then the higher score would take to calculate for ranking the exam result.

Prominent point of the new examination

1. The sum of Grade Point Average from the entire course of upper secondary school student was utilized.
2. The exams were tested 2 times annually then applicants had opportunity to improve themselves. Score test could be kept for 2 years. Higher score could be used for submit the application.
3. Exam result were reported before making decision on faculty alternative then students could choose more appropriated faculty in accordance with their academic capability and preference.

Weakness point of the new examination

1. Examination applicants had more burdens and tenses from two times examinations.
2. Schools had to speed up their final examination before the examination of October to prepare the most omniscience students then it would be affected to normal educational system in schools.

1.1.3 The examination of university admission from Centralized University Admissions (Admissions)

1.1.3.1 The examination by Admissions system on phrase 1

From document at Tor Por Or.44/147 on the dated 19 April 2001 from the Meeting of Director-General of Thailand (Tor Por Or.) offered Ministry of University Affairs considering to improve the system of examination of university admissions in Centralized University Admissions System. It would be launched since the academic year 2004. The main idea was more concentrated on the sum of Grade Point Average from the entire course of upper secondary school student and considered capability of candidates from score of examination on Compulsory subjects and/or Standard examination which be tested from the National Institute of Educational Testing Service which would be established soon. For the GPA must be testified from the Office of the Basic Education Commission or the Office of the Permanent Secretary for Ministry of Education or any of original affiliations to prevent problem of grade point average miscalculation. University/institute might specify others particularities or specify the Special subject which each university or group of universities or the National Institute of Educational Testing Service were arranged the examination or Centralized Admissions System (Ministry of University Affairs was coordinated and university/institute was arranged the examination process). Now the efficiency of this system was acceptable and also be the core system organization of Centralized Admissions System. That improvement worked in harmony with the Educational reform of the National Education Act 1999.

On the Meeting of Director-General of Thailand (Tor Por Or.) which coordinated with Office of the Higher Education Commission was holding the principles to specify for Centralized Admissions System. It had the principles as below:

1. The new system had to change from the system of examination to access to university (Entrance Examination) to be the system of admission (Admissions) which considered from sum of Grade Point Average from the entire course of upper secondary school student. This system must be faired, revealed and testified.

2. The consideration on the sum of grade point average for admission would consider from methods and ranges of times which be specified in the course of Basic Education year 2001 on the Educational Reform Process.

The initially improvement of examination was launched in orderly. The final conclusion was on 30 April 2005 (With all details of Document No.2). The Meeting of Director-General of Thailand launched the system of examination of university admissions on the academic year 2006. It consisted of following components:

1. GPAX	10%
2. GPA (Learning subject)	20%
3. O-NET	35-70%
4. A-NET and/Professional Subject (Less than 3 subjects)	0-35%

(The reason of improvement on the examination system on academic year 2006 because it was the first year which students graduated from the course of the Basic Education Program year 2001 and access to university on the academic year 2006.)

Prominent points on the system of Admissions:

1. Use of school grades such as GPAX, GPA, Learning Subject and O-NET in Admissions system, then student must take more concentrate in school class. It was in accordance with the Admissions which want to take school grades to be one of factors in examination system. It would promote and enhance class procedure of upper secondary school student to meet the philosophy and objectives of education course. That objective would success when efficiency of class procedures in school was acceptable from general public. Then universities were expected that the educational reform would also improve the Basic Education and then increasing in school grades was more acceptable.

2. Additional examination had to less than 3 subjects. It prevented unnecessary tutorial for student. Faculty in university specified the additional examination only for Compulsory subjects which were necessary needed for the examination.

Weak points on the system of Admissions:

1. School standardization was prolonged disputed.
2. On the case that the calculated percentage of Professional subjects or A-NET was less than the calculated percentage of grade point average might be effected to Admissions. Anyway, it was no evidence to prove at that time.
3. The test of A-NET / Professional subjects was annually. Then students who missed the test had to wait till next year for applying in required faculty in university.

1.1.3.2 The examination by Admissions system on phrase 2 since year 2010

The examination by Admissions which launched since academic year 2006 was criticized from several related persons, it shown of oversized school grade was occurred. From the fact that either the examination by Entrance or by Admissions was still criticized by related persons with agree and disagree comments. But Admissions system was aimed for the complete procedure that university were capable to specify criteria of admission considerations and could announce that criteria to public. Students or applicants who apply for certain university must present standardized scores from the National Institute of Educational Testing Service. Then all students would be under the same standardization exam procedure and try to fewer burdens to student as well. Students had to present score to the Centralized examination committee which be standard organized, acceptable and fair under principle of good governance.

The improvement of examination on academic year 2010 which came from the Meeting of Director-General of Thailand who assigned the Associate group of Admissions and Assessment for operating all procedures. The approval principles from the Meeting of Director-General of Thailand were taking school grade from upper secondary school student and also the Aptitude Test to be the components of admissions. The school scores consisted of GPAX and O-NET. For Aptitude Test was replaced for A-NET and/ Professional subjects because Aptitude Test was the test for general not for main substance then it would be tested several times a year.

Components of Admissions for academic year 2010, under the approval from the Meeting of Director-General of Thailand were as below:

1. GPAX	20%
2. O-NET (8 Subject area)	30%
3. GAT (General Aptitude Test)	10-50%
4. PAT (Professional Aptitude Test)	0-40%

The examination of university admissions from Centralized University Admissions System (Admissions) had 2 main objectives were: 1) University/institute successfully get the omniscient and skillful students to their faculty 2) It can improve learning procedures of Secondary School students Grades 6 and fulfill philosophy and objective of Educational Course. The components of university admission relevant to 1) The sum of Grade Point Average from the entire course of upper secondary school student or equivalent (GPAX) which be calculated percentage for 10% 2) Ordinary National Educational Test (O-NET) which be calculated percentage for 35-70% 3) Grade Point Average from the entire course of Secondary School Grades 6 on Group of Learning Subjects will be calculated percentage for 20% 4) Advanced National Educational Test (A-NET) will be calculated percentage for 0-35% 5) The test of interview and physical examination which be not be weighted according to determination of Centralized University Admissions. From this procedure then faculties/departments of every university will specify different examined subjects. The subjects were depended on the vocational areas [1]. Every department would examine the fundamental subjects and then examine the special subjects for each department. Therefore, these procedures would be affected to determination of university admission.

From the university admissions found some problems were: qualified students could not pass the initially educational proficiency of faculty thus that students could not further education in these departments/faculties because they could not pass the minimum scores of departments. Problems above came from The Centralized University Admission System differently defined the weight values of subjects for every departments. Then this research had analyzed the scores of university admission with the techniques of Hierarchical Data Grouping of all departments in Mahidol University. It would help supporting for faculty and

department to determination for upper secondary school student and further more it also help the Freshy students who passed university admission but could not further education in that faculty or department as well as faculty and department could get more optimized student at the same time.

1.2 Objectives of study

1.2.1 Analyzing the O-NET, GAT scores for each faculty in Mahidol University.

1.2.2 Using Factor Analysis for analyzing admission score to reduce factor.

1.2.3 Using Cluster Analysis for grouping admission score's data to get additional knowledge for helping students to choose faculty in undergraduate education.

1.3 Scopes of study

1.3.1 Application for the O-NET, GAT scores data of 36 faculties of Mahidol University as base for analyze.

1.3.2 Use the O-NET, GAT scores of Mahidol University's students for academic year 2010, 2011 and 2012.

1.4 Hypothesis

- Science proficiency students might be capable to study in related science subjects such as Health Sciences and Medicine, Natural Resources Science, Sport Science and Engineering.

- Science proficiency students might be capable to study in Arts also.

- Scores of O-NET, GAT would effect to Factor analysis and Faculty alternative.

1.5 Thesis organization

The thesis is outlined as follows:

Chapter I: Introduction, this chapter explains about the background and statement of problems, the objectives of study, scopes of study, and hypothesis – expected results.

Chapter II: Literature review, this chapter reviews related researches such as clustering.

Chapter III: Materials and methods, this chapter presents research methodology that proposes factor analysis, cluster analysis, clustering tool, research tool and research schedule.

Chapter IV: Results, this chapter presents result of factor analysis, result of cluster analysis.

Chapter V: Discussion, this chapter discusses the result of factor analysis, result of cluster analysis and other parts of research.

Chapter VI: Conclusion and recommendation, this chapter presents some conclusions as well as some further developments are suggested.

CHAPTER II

LITERATURE REVIEW

2.1 Basic knowledge of admission examination

2.1.1 The composition factor of examination for university admissions on academic year 2011

The composition factors of examination for university admissions on academic year 2011 are consisted of:

1. Sum of Grade Point Average from the entire course of upper Secondary School or equivalent (**GPAX**) is counted for **20%**.
2. Ordinary National Educational Test (**O-NET**) which has comprised of 8 major subject areas from the entire course of upper Secondary School (6 semesters) is counted for **30%**.
3. General Aptitude Test (**GAT**) is counted for **10-50%**.
4. Professional and Academic Aptitude Test (**PAT**) is counted for **0-40%**.

2.1.2 Explanation for composition factors

1st Composition factor: **GPAX** (Sum of Grade Point Average) or equivalent (6 semesters)

“GPAX” is grade points which accumulate all studying subjects and all academic years. It indicated genuine learning outcome of studying and assured the graduation from upper secondary school. It enhanced the human development procedure of educational planning. Though, critics complained about under standardization and tension of GPAX because all students tried to focus on GPAX all the time. But GPAX is still the best index currently. From initially studying and beware of related institutions such as Office of the National Education Commission or some university indicated that GPAX reports from school were reliable and no any mistakes. For the tension of students about try to making the highest points of GPAX,

then may cause some tension but on the other hand it enhances students to work hard and pay more attention to their studies. Anyway, attitude adjustment and family support should be relieved that tension. By the way, there have no details for proving the completeness of GPAX recently, then the Meeting of Director-General of Thailand (Tor Por Or.) defines the weight value of GPAX just only 20%.

2nd Composition/ factor: **O-NET** (Ordinary National Education Test)

From the Basic Educational Core Curriculum 2008 defined the National Quality Evaluation. Educational institutions must provide the National Quality Evaluation (National Educational Testing O-NET) for students who studied on the last semester of Grade 6, Grade 9 and Grade 12. The tests are including for 8 major subject areas such as Thai language, Mathematics, Science, Social Science, Religion and Culture, Foreign Languages, Health and Physical Education, Art and Career and Technology. These three tests would be arranged by National Institute of Educational Testing Service (Public organization) which be established for educational testing service. It was essential and useful for ensuring of education quality and also as index for improving study in schools. From the principal that GPA has got from school must not be significantly different from O-NET. In case of it was be so different then school must perform improvement on learning procedures and evaluations to meet the standard definitions by the Ministry of Education. Besides from that objective, O-NET will be useful for university admission as well.

O-NET for Grade 6, Grade 9 and Grade 12 for the academic year 2010 was the first operation. The National Institute of Educational Testing Service (NIETS) arranged O-NET for Grade 6, Grade 9 and Grade 12 on a half day test. Then NIETS analyzed various statically values and report to schools for indicating weakness and strong points. The shorter test of O-NET for Grade 6 and Grade 9 can save time, labor and budget but for Grade 12 was still use the longer test of O-NET.

3rd Composition/ factor: **GAT** (General Aptitude test or General Potential Test)

It's the test for student potential and capacity to success for further studying in university. GAT consists of 2 parts as below:

1. Reading, writing, critical thinking skills and problem solving skills 50%

2. Ability to communication in English 50%

4th Composition/ Factor: **PAT** (Professional and Academic Aptitude Test)

PAT is the professional and academic aptness, skillful, ability and knowledge test for students to achieve in their career and also capacity to succeed in studying on subjects. PAT includes of 7 areas and skills assessed as below:

PAT 1: Mathematic areas and skills assessed

It consists of 2 parts:

1) Knowledge in Mathematic, Algebra, Geometry, Calculus, Statistic etc.

2) This areas was achieving on learning Mathematic in university such as Mathematic thinking, Mathematic solving, Mathematic reading and Mathematic understanding also with Mathematic resolution as well.

PAT 2: Science areas and skills assessed

It consists of 2 parts:

1) Knowledge to achieve in studying science in the Faculty of Science and related field such as Earth Science, Chemistry, Biology, Physics and ICT.

2) Skill to achieve in studying science in university as thinking like scientist's style, problem solving by scientific process.

PAT 3: Engineering areas and skill assessed

It consists of 2 parts:

1) Basic knowledge for engineer studying is Mathematics, Science and Technology.

2) Skills to achieve in studying engineer in university as thinking engineering's style etc.

PAT 4: Architect areas and skills assessed

It consists of 2 parts:

1) Basic knowledge for Architect studying is Mathematics, Science and Technology.

2) Skills to achieve in studying Architect in university as imaginations on 3 dimensions pictures and designs.

PAT 5: Educational Profession areas and skill assessed It

consists of 2 parts:

1) Basic knowledge to study in Faculty of Education as Humanities, Health Education, Art, Environment etc.

2) Skills to achieve studying in Faculty of Education or proficiency of teacher as knowledge searching skills, intercommunication skills etc.

PAT 6: Art areas and skills assessed

It consists of 2 parts:

1) Knowledge on theory of Fine Art, Dancing Arts, Music and others basic knowledge to achieve studying in Faculty of Liberal Arts or related.

2) Skills in studying on Arts as creative thinking etc.

PAT 7: Foreign Languages areas and skills assessed

It consists of 2 parts:

1) Knowledge in grammar, linguistic, literature etc.

2) Skills in listening, speaking, reading, writing, conclusion, shorten, describe, synthesis, analysis etc.

The examination for university admissions would be significantly considered on the entire school report of upper secondary school and tried to avoiding of additional exam. If additional exam was occurred, then it must be tested not more than 3 subjects.

2.2 Profile of O-NET, GAT, PAT and admission

O-NET questions for upper secondary school student in academic year 2010 (On February of academic year 2011)

1. Thai languages (Subject Code 01):

- Reading
- Writing
- Listening , observation and speaking
- Principles on language application
- Literature and Literary outputs

2. Social Science, Religion and Culture (Subject Code 02):

- Religion, morality and morality

- Civil responsibility, culture and life in society
- Economics
- History
- Geography

3. Foreign languages (English) (Subject Code 03):

- Language for communication
- Language and cultural
- Language and other subject groups relationship
- Language, community and work relationship

4. Mathematics (Subject Code 04):

- Number and operation
- Measurement
- Geometry
- Algebra
- Data analysis and probability
- Skill/Procedure of Mathematics

5. Science (Subject Code 05):

- Living beings and life existence processes
- Life and environment
- Matter and properties of matter
- Force and mobility
- Energy
- Evaluation of earth
- Astronomy and space
- Nature of science and technology

6. Mixed 3 subjects group in one paper (Subject Code 06):

- **Health Education and Physical Education**
 - Human growth and development
 - Life and family
 - Movements, physical exercises, game, Thai and international sports

- Building up health capacity and sickness prevention
- Life and safety
- **Art**
 - Visual art
 - Music
 - Thai traditional dance
 - Innovative style
- **Career and Technology**
 - Life living and family
 - Career
 - Design and technology
 - Information system
 - Technology for work and career
 - Innovative style

2.3 The 3rd GAT/PAT testing on October of academic year 2010

GAT (General Aptitude Test) or General Aptness

Section 1: Reading, Writing, Critical thinking skills and problem solving skills is counted for 50%.

Section 2: Ability to communication in English is counted for 50%.

PAT (Professional and Academic Aptitude Test) or Aptness on professional and academic

PAT 1 Mathematics

PAT 2 Science

PAT 3 Engineering

PAT 4 Architecture

PAT 5 Educational profession

PAT 6 Fine and Applied Arts

PAT 7 Languages

PAT 7.1 French

PAT 7.2 German

PAT 7.3 Japanese

PAT 7.4 Chinese

PAT 7.5 Arabic

PAT 7.6 Bali

2.4 Contents of 3rd GAT/ PAT on academic year 2010 (tested on October 2010)

GAT (General Aptitude Test)

Section 1: Reading & analysis, writing & analysis, critical thinking & analysis and problem solving

Section 2: Ability to communication in English:

1. Speaking and conversation
2. Vocabulary
3. Structure and writing
4. Reading comprehensive

PAT 1: Mathematics

1. Logic
2. Set
3. Real number system
4. Relation and Function
5. Trigonometry function
6. Geometry Analytic
7. Exponential & Logarithm Function
8. Matrix
9. Vector
10. Complex number
11. Series
12. Fundamental of Calculus
13. Permutation and Combination
14. Probability
15. Fundamental of Data Analysis
16. Numeral Distribution
17. Real-time Function among data

PAT 2: Sciences

1. Chemistry
2. Environmental Biology
3. Physics

4. Earth

PAT 3: Engineering

1. Mechanic
2. Electric
3. Chemistry
4. Energy, thermal and fluid
5. Mathematics and applied engineering statistic

PAT 4: Architecture

1. Logic and problem solving
2. Science
3. General knowledge

PAT 5: Educational Profession

General knowledge on content of teaching

PAT 6: Fine and applied Arts

1. General knowledge on arts and cultural
2. Fine art and design
3. Thai-international music
4. Thai dancing art-performance art

PAT 7.1 French

1. Basic vocabulary
2. Grammar
3. Idiom
4. French cultural
5. Pronunciation

PAT 7.2 German

1. Basic vocabulary
2. Grammar
3. Cultural life

PAT 7.3 Japanese

1. Basic vocabulary
2. Fundamental of “ kunji”
3. Basic grammar

4. Basic idiom in daily life
5. Japanese education

PAT 7.4 Chinese

1. Vocabulary
2. Grammar and structure
3. Idiom
4. General knowledge about China

PAT 7.5 Arabic

1. Grammar
2. Cultural of Arabic-Thai.
3. Vocabulary
4. Comprehension

PAT 7.6 Bali

1. Basic vocabulary
2. Grammar and structure
3. Comprehension

2.5 Related literatures

Supansa [13] presented the clustering of Secondary School in Khon Kaen province and invented model for clustering schools by using the basic data and exterior standard evaluation of 100 Secondary Schools from Educational Service Area of Khon Kaen province. The research found that this clustering could formed characteristics of Secondary Schools into 4 groups as Group of education for level of district, Group of education for level of provincial, Group of education for level of out of district and Group of required characteristic of student and then the result of the Discriminant analysis shown that the predictions accuracies was 91.9%.

Narongsak [5] presented the applying of related regulation for guiding student to further education in university. From grade point average of the core 7 subjects as Mathematics, Chemistry, Biology, Physics, Thai Language, Social and English Language which related to education data analyzing by FP-Growth Algorithm shown that the accuracy of means was 89.87%.

Kijtipong [4] presented the Analysis and Inspection on attacking data for WAN system by using **Cluster Analysis** and **Discriminant Analysis** for inspection data and found that the accuracy of grouping by system was 85.87%.

P.L. Hsu, et. Al. [6] presented the techniques of further education in bachelor degree by applying procedure of relation on association rule with procedure of genetics and found that the efficiency of forecasting is better and utilized fewer time of processing than SGA.

Krisana Waiyamai, et. al. [7] presented the searching for role of relationship, data classification and data forecasting to search for the interesting topic from students' data. This searching could help student to determinate on preference of faculty alternative and also on the prediction of learning outcomes for the next semester. The research used the Tree Diagram to be the center of model for data classification.

Wei Zong [8] presented **Principal Component Analysis (PCA)** was used to evaluate the curriculum designing and **Hierarchical Cluster Analysis. (HCA)** were used to analysis the teaching effect by utilizing grades of 20 students who graduated in 2007. PCA showed that every curriculums were necessary for cumulative. HCA showed that by this curriculum system if any students studied hard, they could get a good grade accompanying with integrated development ability. So PCA and HCA could be useful tools for teaching of good engineering specialty.

Yang Yang, et. al. [9] presented multivariate statistical techniques such as **Principal Component Analysis (PCA)** or **Cluster Analysis (CA)**. They were applied for evaluation on variations and interpretation of large complex for water quality on data set of the lakes in Wuhan. It generated in 2009, monitoring of 21 parameters at 70 different lakes (1470 observations) located at the 7 core districts of Wuhan, Hubei Province, China. Results revealed that Hierarchical cluster analysis grouped 70 lakes into 3 clusters as Less Polluted (LP), Medium Polluted (MP) and Highly Polluted (HP) lakes which based on the similarity of water quality characteristics. Thus, this study illustrated the usefulness of multivariate statistical techniques for analysis and interpretation of complex data sets, and in water quality assessment. It identified on pollution sources/factors and understanding temporal/spatial variations in water quality for effective lakes water quality management. This study suggested that PCA

and CA techniques were useful tools for identification of important surface water quality monitoring stations and parameters.

Jun Qu, et. al. [10] presented the selection on appropriate number of clusters and distinguishable partially overlapping on irregular data were two important problems in clustering. **Hierarchical Clustering** provided a good solution to those problems. Similarity measure was the key of controlling the iterative process of hierarchical clustering. The research gave a definition of overlap similarity measure and proposed a hierarchical clustering algorithm without prior specified number of clusters. The appropriated value could be decided in the iterative process. The algorithm stopped clustering accordance to the overlap similarity between clusters. **Cluster Analysis** was a useful approach to unsupervised image segmentation. After discussing some related topics, the research applied it to synthetic and real image segmentation for evaluating performance of the clustering algorithm and compared with others algorithms. Moreover the research estimated parameters of algorithm in image segmentation. The results showed that this approach could be effectively applied to image segmentation.

Jinliang Zhang [11] presented the parameters of flow zone indicator, porosity, permeability, clay content and poro-throat radius (R35). Reservoir characteristics were selected as evaluation parameters. **Clustering Analysis** Theory defined reservoir flow units in Jiangsu oil field into 4 types according to evaluation criterion. The result discussed on main features of each flow unit combination with sediment logical and performance data. Finally, the research summarized that type 2 and 3 were the main zones of remaining oil distribution by analyzing the relation of flow unit and remaining oil.

CHAPTER III

MATERIALS AND METHODS

3.1 Related theories

3.1.1 Factor analysis

3.1.1.1. Definition

Factor Analysis is the one technique of Multivariate Analysis. This technique will assemble relevant variables into the same cluster. The relation of factors may on the positive direction (same direction) or the negative direction (opposite direction). These methods are based on linear algebra, and also on a tool which is very useful for clustering and pattern recognition. [2]

3.1.1.2 Objectives

1. Decreasing numbers of variables by assembling several variables into the same cluster.
2. Accurately testifying such as on a certain research, researcher has to define the weight value to variable. Sometime researcher made a measurement error and then they can utilize technique of Factor Analysis for more accuracy on testing results.

3.1.1.3 Benefit

1. Decreasing numbers of variables by assembling several variables into the same factor. Thus the new created factor is the new variable. So we find the value of that new created factor. It calls Factor Score.
2. Solving the problem of relation on independent variable in technique of Regression Analysis (Multicollinearity).
3. Enhancing structure of studied variables relation because technique of Factor Analysis will find the correlation coefficient for a couple of variables. Then more related variable was assembled to the same factor.

4. Explaining the meaning of each factor. It was come from the meanings of variables on that factor. Then we can take that meanings for others planning.

3.1.1.4 Criterion

Factor analysis technique is the technique of changing originally relative variables into the new irrelative factor. Thus that factor is the Linear combination of original variable. Then try to take all details of original variable into factor. Then estimated equation of Factor j is

$$F_j = W_{j1}X_1 + W_{j2}X_2 + \dots + W_{jp}X_p + e \quad (3.1)$$

When X_j is Factor J, W_j is coefficient of Factor J. Relation OF Variable X_i is Linear combination of factors as below:

$$\begin{aligned} Z_1 &= L_{11}F_1 + L_{12}F_2 + \dots + L_{1M}F_M + e_1 \\ Z_2 &= L_{21}F_1 + L_{22}F_2 + \dots + L_{2M}F_M + e_2 \end{aligned} \quad (3.2)$$

$$Z_p = L_{p1}F_1 + L_{p2}F_2 + \dots + L_{pm}F_m + e_p$$

When Z_i is standardized variable X_j , $j=1,2,3,\dots,p$, p is number of variable, M is the number of factor when $m < p$, $F_1, \dots, F_m$ is the Common factor, E is the Unique factor, L_{ij} is the coefficient or the Factor loading.

3.1.1.5 Step of analysis

Step 1: Testify for relation of variables.

Step 2: Factor Extraction: Purpose of the step is for calculating number of factor that can substitute for all of variables.

Step 3: Factor Rotation: Purpose of Factor rotation is increasing or decreasing value of Factor loading of variables. Until it shown that what variable should or should not stay in certain factor.

Step 4: Calculation of factor score [2].

3.1.1.6 Decision Process

Stage 1

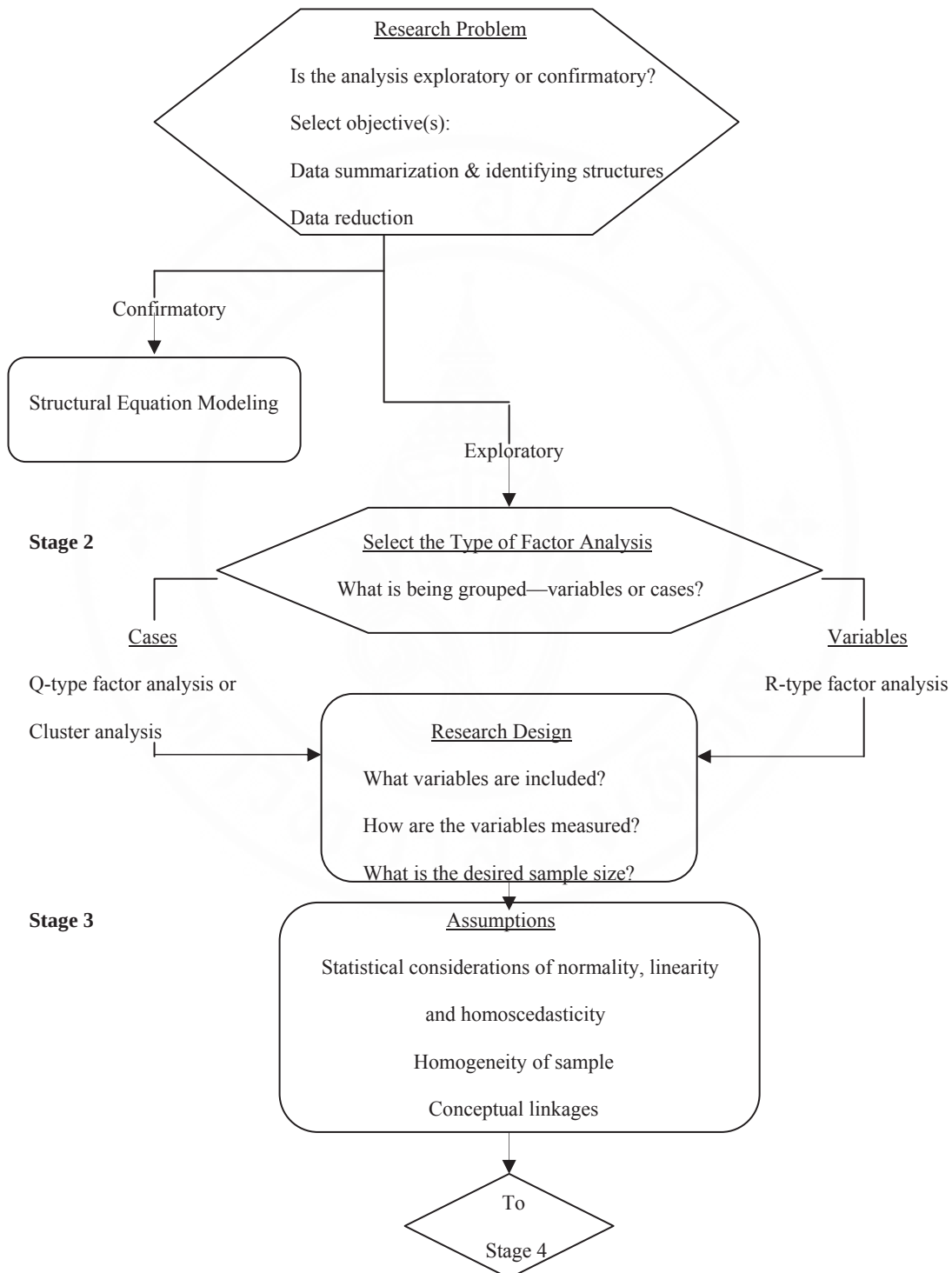


Figure 3.1 Stages 1-3 in the Factor Analysis Decision Diagram [22]

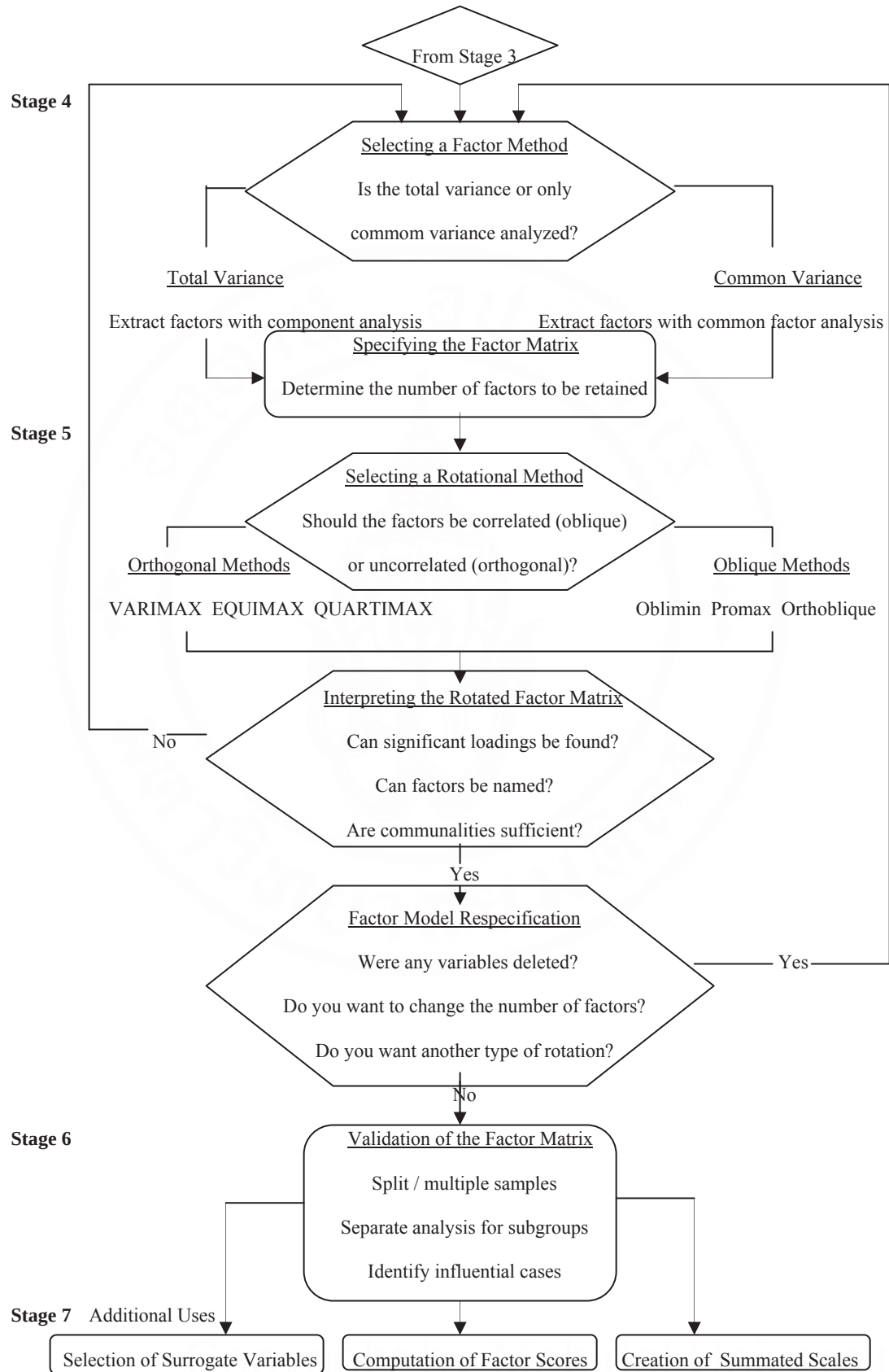


Figure 3.2 Stages 4-7 in the Factor Analysis Decision Diagram [22]

Criteria for the Number of Factors to Extract

In deciding of how many factors to extract, researcher must combine a conceptual foundation with some empirical evidence. Generally, it begins with some predetermined criteria, such as the general number of factors plus some general thresholds of practical relevance. These criteria are combined with empirical measures of the factor structure.

The Three Processes of Factor Interpretation

Factor interpretation is circular in nature. The research first evaluates initial results then makes a number of judgments in viewing, refining and returning to evaluative step. Thus the research should not be surprised to engage in several iterations until a final solution is achieved.

Estimate the Factor Matrix

The initial unrotated Factor Matrix is computed. It contains the Factor Loading for each variable on each factor. Factor loading are the correlation of each variable and factor. Loadings indicate degree of correspondence between variable and factor. The higher loadings make variable representative of factor. Factor Loadings are the means of interpreting the role of each variable plays in defining each factor.

Factor Rotation

Factor Rotation should simplify factor structure. Researcher uses the rotational method to achieve simpler and theoretically meaningful factor solutions. In most cases the Factor Rotation improves the interpretation by reducing some of the questions which often occur with initial unrotated factor solutions.

Factor Interpretation and Respecification

In the final process, researcher evaluates the rotated Factor Loadings to determine the role of variable and contribution.

On the evaluative process, the need may occur to specify the factor model. Respecification of factor model involves to the returning to the extraction stage, extracting factors and beginning the process of interpretation again.

Rotation of Factors

The simplest case of rotation is Orthogonal Factor Rotation which the axes are maintained at 90 degrees. It's possible to rotate the axes and not retain the 90-

degree angle between the reference axes. These factor rotations are demonstrated in Figure 3.3 and figure 3.4 below:

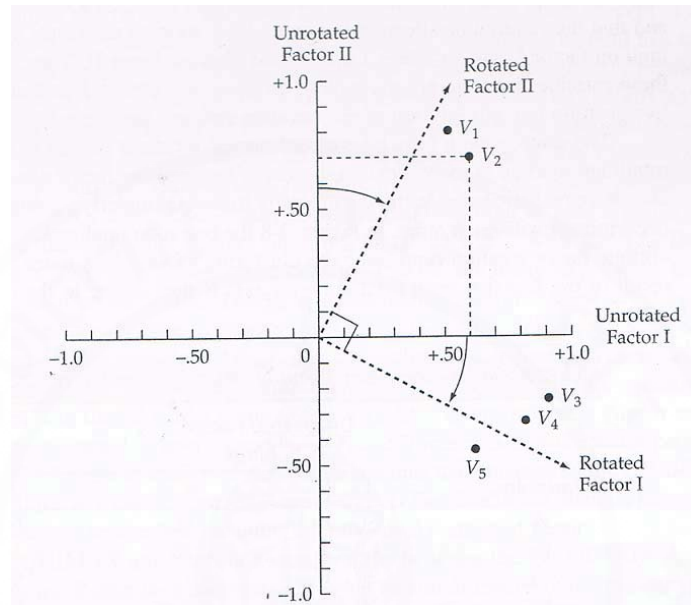


Figure 3.3 Orthogonal Factor Rotation [22]

From Figure 3.3 above five variables are depicted in two dimensional factor diagram. Vertical axis presents unrotated factor II. Horizontal axis presents unrotated factor I. Axes are labeled with 0 at origin and extend to +1.0 or -1.0. Numbers on axes presents the Factor Loadings. Five variables are labeled with V_1 , V_2 , V_3 , V_4 and V_5 . The Factor Loading for variable2 (V_2) on unrotated factor II is determined by drawing a dashed line horizontally from the data point to vertical axis for factor II. Vertical line is drawn from variable 2 to horizontal axis of unrotated factor I to determine the loading of variable 2 on factor I. Remaining variable follows the same procedure. It determines the Factor Loading for unrotated and rotated solutions. It will display in Table 3.1 below. On the unrotated first factor, all variables load fairly high. For the unrotated second factor, variables 1 and 2 are very high in the positive direction. Variable 5 is moderately high in negative direction. Variable 3 and 4 are considerably lower loadings in negative direction. By the way, after inspection, two clusters of variables are obvious. Variable 1 and 2 go together also with variable 3, 4 and 5. However, this pattern is not so obvious from unrotated Factor Loadings. By rotating original axes clockwise, the research obtains a completely different Factor Loading pattern. It shows that the axes are maintained at 90 degrees. This procedure

signifies that factors are mathematically independent and rotation has been orthogonal. After loading the factor axes, variables 3, 4 and 5 loads high on factor I and variables 1 and 2 load high on factor II. Therefore, the clustering or patterning of these variables into two groups is more obvious after rotation and before, even though the relative position or configuration of variables remains unchanged.

Oblique rotation is more flexible. But the factor axes need not be Orthogonal, then it is more realistic because theory underlies that dimensions are not assumed to be uncorrelated with each other.

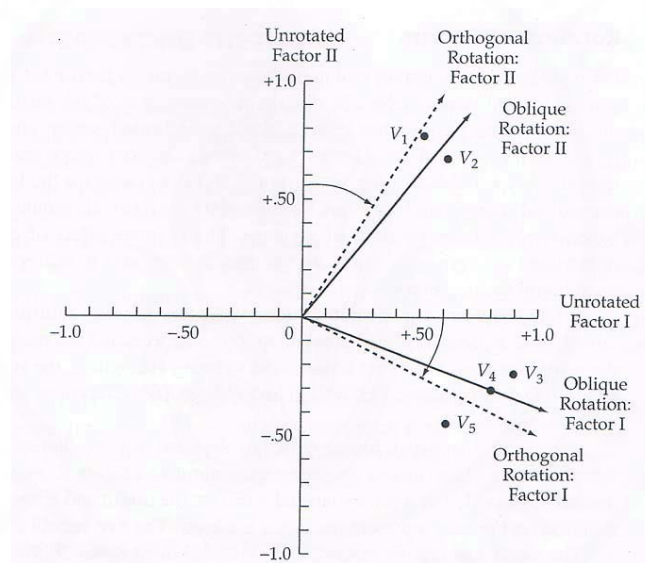


Figure 3.4 Oblique Factor Rotation [22]

In Figure 3.4, Oblique Factor Rotation presents that variables clustering are more accurately. Accuracy is a result of the fact of each rotated factor axis is now closer to the respective group of variables. Oblique method provides information about extent because the factors are actually correlated with each other.

By the way, unrotated solutions are not sufficient. In most cases rotation will improve the interpretation by reducing some of ambiguities which occur on initially analysis. Major option is to choose Orthogonal or Oblique rotation. Ultimate goal of rotation is to obtain theoretically meaningful factors also with the simplest structure. Orthogonal method is more widely used. Whereas, Oblique method are not well developed and still some controversy. Only limited numbers of Oblique procedure are available.

Table 3.1 Comparison between Rotated and Unrotated Factor Loadings

Variables	Unrotated Factor		Rotated Factor	
	Loadings		Loadings	
	I	II	I	II
V ₁	.50	.80	.03	.94
V ₂	.60	.70	.16	.90
V ₃	.90	-.25	.95	.24
V ₄	.80	-.30	.84	.15
V ₅	.60	-.50	.76	-.13

Selecting among rotational methods

Most programs have default rotation of VARIMAX but Orthogonal or Oblique rotation should be made on the basis of particular needs of a certain research.

Judging the Significance of Factor Loadings

In interpretation factors, a decision must be made on the Factor Loading with consideration and attention. The significant of practical and statistical as well as the number of variables can effect to the interpretation of Factor Loading too.

Principal Component Analysis

When variable p describes n individuals of a population are all numerical. Each individual can be presented by a point in a p -dimensional space R^p . Set of individuals is a “cloud points”. When $p \leq 2$, distance between individuals can be seen clearly by simple observation of the cloud; this observation becomes more difficult when $p = 3$ and is impossible when $p > 3$. It would be desirable to reduce space R^p to R^2 or R^3 . Problem is that the choice of two or three variables is naturally arbitrary and may result in a considerable loss of information from data, since there is not able to know in advance whether these are the most discriminating variables. [23]

Principal Component Analysis (PCA) is method for projecting the cloud of individuals on to subspaces with fewer dimensions while maintaining distance between individuals as much as possible. The research began by systematically centering all variables by subtracting their means, so that research was looking on variables with a zero mean. This simplifies calculation and geometrical representation

because the center of gravity of the cloud of individuals then coincides with origin 0 of axes and subspaces.

Determinations of subspaces are carried out for each axis in turn. Each individual x_i has a weight p_i . This weight is generally $p_i = 1/n$ for every i , but different weight can be given to individuals belonging to different sub-populations. Sum of squares of distances of individual's x_i from their center of gravity, multiplied by their weight p_i , is called the total inertia:

$$I = \sum p_i d(0, x_i)^2 \quad (3.3)$$

Aim of PCA is to find the axis for inertia projected on this axis is maximized. Inertia projected on an axis is sum of squares of coordinates v_i of individuals on axis. These squares are weighted by p_i . In other hand, research looked for axis which the sum:

$$\sum p_i v_i \quad (3.4)$$

The sum reached the maximum and possibly closes to I . This is equivalent to minimizing the difference between each individual and projection. Having done this, research looked for a second axis which will be the one which maximizes the inertia projected on it. This inertia projected on the second axis is less than projected on the first axis. A number of factor axes can thus be determined in succession with decreasing projected inertias. Because of their Orthogonal, the total inertia of the cloud of individuals is broken down into sum of inertias projected on each axis.

Then the concept of **distance** is used here. In the space of individuals, the simplest distance is the Euclidean distance, according to distance of two individuals $x = (x_1, x_2, \dots, x_p)$ and $y = (y_1, y_2, \dots, y_p)$ is:

$$d(x, y) = (x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_p - y_p)^2 \quad (3.5)$$

This distance is very useful in physical but less so in economics and social science, where data $x_1, x_2, \dots, x_p$ which be manipulated may be as unlike and incomparable as age, income, turnover, number of children etc.

In practice, "inverse of variances" is practically used as distance. It is defined as:

$$d(x, y) = ((x_1 - y_1) / \sigma_1)^2 + ((x_2 - y_2) / \sigma_2)^2 + \dots + ((x_p - y_p) / \sigma_p)^2 \quad (3.6)$$

With this new distance, which is a way of **reducing** variables, distance between 2 individuals are no longer depends on unit of measurement and more dispersed variables are not favored. Even if units of measurement are not the same for all variables, “inverse of variance” distance brought all of them to the same level. Normalized PCA opposed to non-normalized PCA in which variables are centered but not reduced. It can assume that we are dealing with “inverse of variance” distance but will occasionally point out certain special features which arise from the use of simple Euclidean distance.

Before examining the cloud of variables, the *covariance* $cov(X, Y)$ of two numeric variables X and Y is an indicator of their simultaneously variation, which is positive if Y increases whenever X increases and is zero if X and Y are independent. If the standard deviations of variables X and Y are denoted σ_x and σ_y , their means are denoted μ_x and μ_y . Then their values are denoted $(x_i)_i$ and $(y_i)_i$. Their linear correlation coefficient is r_{xy} , then the covariance is

$$C o v (X , Y) = \frac{1}{n} \sum (x_i - \mu_x) (y_i - \mu_y) \quad (3.7)$$

And we find that

$$cov(X,Y) = \sigma_x \cdot \sigma_y \cdot r_{xy} \quad (3.8)$$

If we use σ_{ij} as a simpler notation for covariance of x_i and x_j , then the covariance matrix is given by:

$$M_{cov} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \cdots & \sigma_2^2 & \cdots & \sigma_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{pmatrix} \quad (3.9)$$

This is a matrix which diagonal terms are the variance of variables and the trace is sum of the variances of variables. This matrix is also positive, semi-definite and symmetric, meaning that is diagonalizable with orthogonal eigenvectors and eigenvalues, all non-negative. By changing variables, therefore, it is possible to find a base which non-diagonal terms are all zero.

When variables are reduced, formula $\text{cov}(X, Y) = \mu_x \cdot \mu_y \cdot r_{xy}$ shows that the covariance matrix becomes:

$$M_{\text{corr}} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1n} \\ \cdots & 1 & \cdots & r_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ r_{n1} & r_{n2} & \cdots & 1 \end{pmatrix} \quad (3.10)$$

It is the matrix of the linear correlation coefficient (r_{xy}). Therefore, it called the correlation matrix. Its trace is equal to the number p of variables.

Use of PCA

There is no supported software relatively simple of PCA. Furthermore, there are still some pitfalls to avoid when interpreting of PCA as follow:

1. Individual space and variable space must not be over defined.
2. Proximity of two variables means nothing unless they are near the circle of correlation.
3. Intersection of first two factor axes is not only one which offers useful information.
4. Avoid individual or group of individuals have over contribute to the first axes. Axis may be almost entirely accounted for a single individual.

Choosing the number of Factor axes

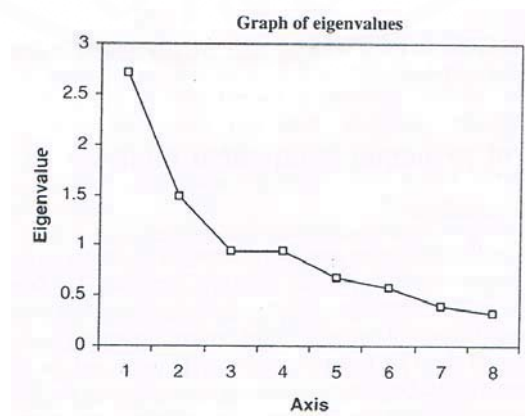


Figure 3.5 Choosing the number of Factor axes [23]

From Figure 3.5, research retained the first two eigenvalues. This test was stated in analytical terms as Cattell's scree test: the existence of a bend (at the i^{th}

eigenvalue) corresponds to the vanishing (at the $(i + 1)^{\text{th}}$ eigenvalue) of the second derivative of the function $f(k) = k^{\text{th}}$ eigenvalue, and stopped selecting new axes before this second derivative vanishes.

When this criterion is used to determine the number of axes to be retained, it must be careful because the fact that 40% of the inertia is provided by the first axis does not have the same meaning regardless of whether the research are dealing with 10 or 50 variables.

The most widely used form of PCA with rotation is **varimax PCA**. It is based on the principle of maximizing, for each factor, not sum of squares of the correlation coefficients of this factor with the set of variables, but variance of these correlation coefficients, with the result that each factor is strongly correlated with some variables and weakly correlated with the others. Thus some variables have a high contribution to each axis, while others have a very low contribution and axes are easy to interpret.

Below is the example of five wines, described in terms of acidity, sugar and alcohol content, matching with meat and desserts, the hedonic dimension and price.

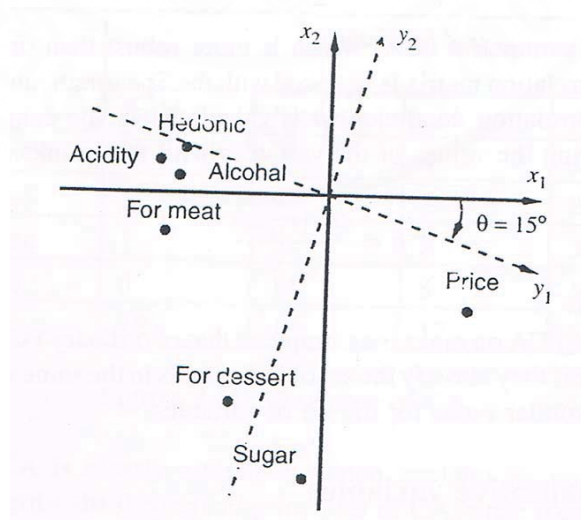


Figure 3.6 Example of VARIMAX rotation [23]

Figure 3.6 showed that VARIMAX PCA provided the best interpretation of the price and sugar axes.

Promax PCA is a hybrid method consisting of a **varimax rotation** followed by an **Oblique rotation**. So the high and low factor coordinates of variable space correspond to same variables but with low values of coordinates which are even weaker.

These variants of PCA are provided in IBM SPSS Statistics and in SAS/STAT FACTOR procedure which is more generally applicable but more complex and slower than PRINCOMP which is used for ordinary PCA. **Quartimax PCA** also forms the basis of VARCLUS procedure in SAS/STAT, used for clustering numeric variables. However, the forms of PCA with rotation preferred in English-speaking countries are not included in the SPAD software which relates more to French-style data processing.

3.1.2 Cluster Analysis

3.1.2.1 Definition

Clustering is the statistical operation of grouping objects (individual or variables) into a limited number of groups known as clusters (or segments), which have two properties. Cluster analysis is a multivariate statistical technique that allows the identification of groups, or clusters, of similar objects in a space that is typically assumed to be multi-dimensional. Groups of subjects are formed in such a way that objects in the same cluster are similar to one another and dissimilar to objects in other clusters. Clustering is a task that has been practiced by humans for thousands of years, and it has been fully automated in the last few decades due to the advancements in computing technology [3]. There is often a confusion regarding the usage of the terms clustering and classification. It is important to distinguish between these two related but different tasks. In classification, the classes are assumed to be pre-defined. The job of the classification module, or the classifier as it is often being called, is to identify a subset of the pre-defined classes to be assigned to the object. On the other hand, in clustering task, the classes are often not known in advance. The clustering module is then required to discover appropriate classes based on some similarity measure quantifying how similar or how closely related the objects are to each other. Objects that are closely related are usually grouped separate clusters. Thus, as is the case in many clustering techniques, the goal of clustering is to form clusters

of objects such that the intra-cluster object similarity is maximized, but the inter-cluster object similarity is minimized [4].

3.1.2.2 Methodology

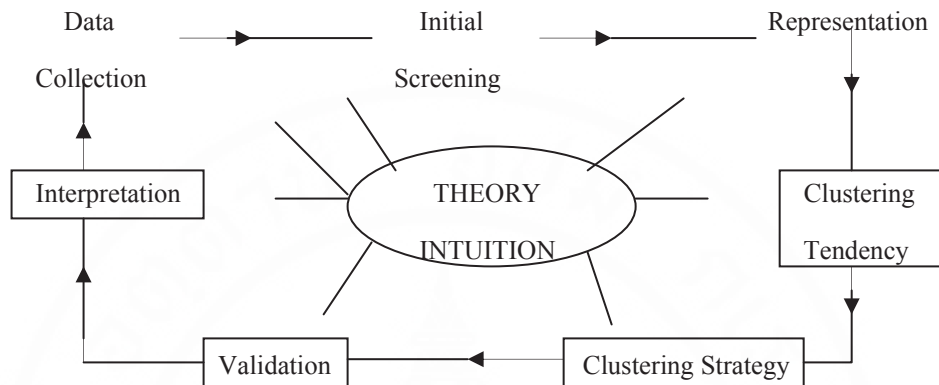


Figure 3.7 Clustering Methodology [24]

From Figure 3.7 above can explain all details as below:

1. Data Collection

Careful recording of data in accordance with standards area of application is the first important step in analysis. Previous work in the subject matter, resources available and patience of investigator must be considered. Amount and type of data will strongly influence the strategies available for analyzing data. So a few iterations through loop in Figure 3.7 might be necessary before a meaningful and compact procedure for data collection can be established. [24]

2. Initial Screening

Raw data usually need some screening before formal analysis. Screening method should suit data and researcher should have an understanding and confidence in the method chosen.

3. Representation

Suitable form of procedure should be choosing a proximity index, projecting data to suitable feature space, examining intrinsic dimensionality and performing a multidimensional scaling. Suitability of these procedures depends on problem at hand. End result should either be a **pattern matrix** or **proximity matrix**. Representation chosen will depend on data, application area, experience of researcher and availability of computer software.

4. Clustering Tendency

The information gained from this step can not only prevent inappropriate application of Cluster Algorithms but can also provide information on fundamental nature of data. If data cannot be shown to have tendency to cluster, then it is well advised to pursue analysis techniques other than Cluster Analysis.

5. Clustering Strategy

A major question is the choice between Hierarchical and Partitional procedures. Within each type of clustering must be given to several details: matching the algorithm to data, presentation of results and choice of parameters. Amount of data is a major factor. Researcher can also choose to cluster both the patterns and features.

6. Validation

Careful validation of clustering results is essential step that changes a qualitative analysis into hard evidence. Validation often involves Monte Carlo Analysis and statistical testing. It demands more computer resources, time and care than collecting data and clustering it. It is the price for generating reliable results. Researcher might choose to study stability of analysis by perturbing data slightly and repeating analysis. Stability is one basis for comparing clustering methods.

7. Interpretation

The more exploratory data analysis is used, the more confident one becomes in its use.

Figure 3.7 above shows the major steps to be considered when undertaking an **Exploratory Data Analysis** whose central component is a **Cluster Analysis**. It shows the process as an endless loop in which new insights are obtained and new ideas generated each time through the loop. End result could be design of an experiment that uses standard statistical tools to come to decisions about phenomenon being studies. Researcher might derive enough information about the phenomenon from an exploratory data analysis itself to draw informal conclusions. Exploratory data analysis remains a tool for discovery. However, the fact that cluster analysis is exploratory in nature does not mean that only ad hoc procedures can be adopted.

3.1.2.3 Clustering techniques

Different approaches to clustering data can be described in the taxonomy of clustering approaches in figure. At the top level, there is a distinction between hierarchical and partitioned approaches. Hierarchical methods produce a nested series of partitions, while partitioned methods produce only one. In this research, we focus on the hierarchical approaches only.

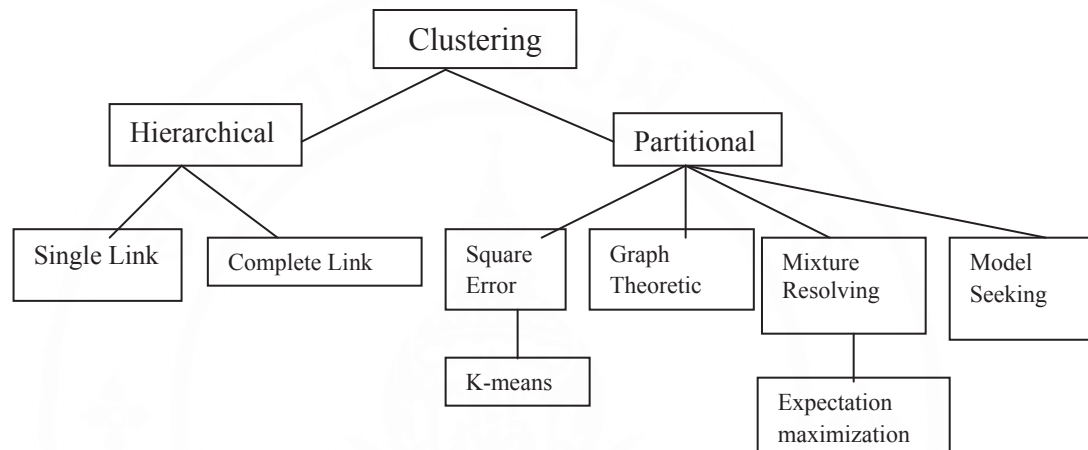


Figure 3.8 Taxonomy of clustering approaches [5].

3.1.2.3.1 Hierarchical Clustering Algorithms

A huge numbers of hierarchical clustering algorithms were different in single link, complete link and minimum variance algorithms. The single link, complete link, and group average were well-known. The distinction of these algorithms showed resemblance between a pair of clusters. Single link algorithm showed distance between two clusters and presented minimum of the distance between all pairs of patterns drawn from two clusters. In the complete link, distance between two clusters was the maximum of all pairs between patterns in two clusters.

One sample of a hierarchical clustering was the correspondence tree or dendrogram which showed how samples were grouped. The first level showed all samples x_i as singleton clusters. When we increased levels, then, more numbers of samples were clustered together in a hierarchical manner.

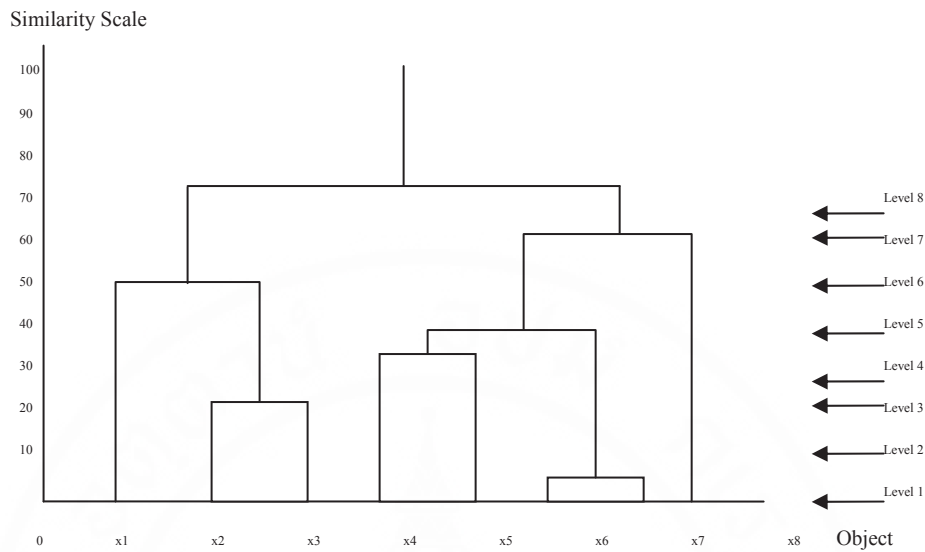


Figure 3.9 Example of a hierarchical clustering in a dendrogram [6].

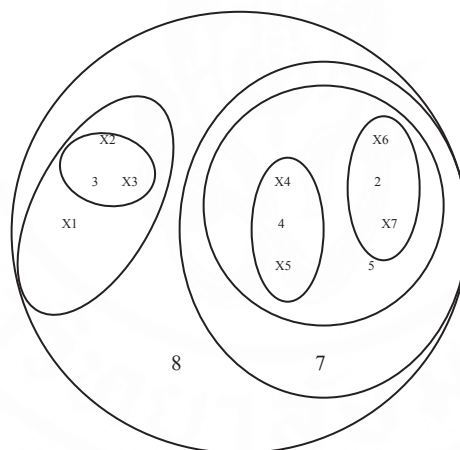


Figure 3.10 Example of a hierarchical clustering in Venn diagram [6]

The hierarchical clustering procedures can be divided into two different approaches: agglomerative and divisive. The agglomerative was a bottom-up clumping approach. It used by Single Link method which was the nearest neighbor algorithms and also by Complete Link method which was the farthest. The research began with n singleton clusters and merged more to be another one.

Thus the research used the divisive approach to the Minimal Spanning Tree. It was a top-down approach which started with one cluster and separated clusters to others. If any clusters were compressed and splitted then agglomerative algorithms presented same result. In other hand, if clusters were closed to another or if the shape were not hyper spherical, then different results were being

occurred. Complexities of space and time were $O(n^2)$ and $O(cn^2d^2)$ respectively, where c was number of clusters and d was distance.

Single Link Method (Nearest Neighbor)

Describe the nearest-neighbor clustering algorithm as below:

1. Plot the objects in n -dimensional space (where n is the number of attributes)
2. Calculate the distance from each object (point) to other points, using Euclidean distance measure and place the numbers in a distance matrix.
3. Identify two clusters with the shortest distance in the matrix, and merge them together. Re-compute the distance matrix, as those two clusters are now in a single cluster, (no longer exist by themselves).
4. Repeat Step 3 until all clusters are merged.

The algorithm of Single link method [7]

Complete Link Method (Farthest-Neighbor)

Describe the farthest-neighbor clustering algorithm as below:

1. Begin disjoint clustering level $L(0) = 0$ and sequence number $m = 0$.
2. Find the most similar pair of clusters in the current clustering, assumed pair $(r), (s)$ to $d[(r), (s)] = \max d[(I), (j)]$ where the maximum is over all pairs of clusters in the current clustering.
3. Increase the sequence number: $m = m + 1$. Merge clusters (r) and (s) into a single cluster for performing the next clustering m . Set the level of this clustering to $L(m) = d[(r), (s)]$.
4. Update the proximity matrix D by deleting rows and columns corresponding to clusters (r) and (s) and adding a row and column corresponding to currently prior cluster. The proximity between the new cluster, denoted (r, s) and prior cluster (k) is defined as $d[(k), (r, s)] = \max d[(k), (r)], d[(k), (s)]$.
5. If all objects are in the same cluster. Then stop and go to step 2.

The algorithm of Complete link method [7]

Average Linkage between Group Method

Describe the average linkage between group clustering algorithms as below:

Before the first merge, let $N_i=1$ for $I=1$ to N . Update str by

$$Str = sp\ r + sq\ r$$

Update N_t by

$$N_t = N_p + N_q$$

Then choose the most similar pair based on the value

$$s_{ij}/(N_i N_j)$$

The algorithm of Average linkage between group method [9]

Centroid clustering method

Describe the centroid clustering principle as below:

Centroid linkage uses the Euclidean distance between the centroids of the two clusters.

The principle of Centroid clustering Method [25]

Ward's method

Describe the Ward's method principle as below:

Ward's linkage uses the incremental sum of squares; that is, the increase in the total within-cluster sum of squares as a result of joining two clusters. The within-cluster sum of squares is defined as the sum of the squares of the distances between all objects in the cluster and the centroid of the cluster.

The principle of Ward's method [25]

3.1.2.3.2 Comparison of Hierarchical Clustering Methods

Copyright by Mahidol University

Single Link Method (Nearest Neighbor)

Similarity : Join the most similar pair of objects which on different cluster. Distance of two clusters is the distance between the closest pair of points.

Type of clusters : Chains, ellipsoidal

Time : Usually $O(N^2)$

Space : $O(N)$

Advantage : Good at handling on non-elliptical shapes

Disadvantage : Sensitive to noise and outliers [10].

Complete Link Method (Farthest Neighbor)

Similarity : Join least similar pair between each of two clusters.

Type of clusters : All entries in a cluster are linked to one another within some minimum similarity, so have small, tightly bound clusters.

Time : Worst case is $O(N^3)$ but sparse matrices require less

Space : Worst case is $O(N^2)$ but sparse matrices require less

Advantage : Less susceptible to noise and outliers

Disadvantage : Break large clusters and it favors globular shapes

Average Linkage between Group Method (Group average)

For the group average version of hierarchical clustering, the proximity of two clusters is defined as the average pairwise proximity among all pairs of points in the different clusters. This is an intermediate approach between the single and complete link approaches. Thus, for group average, the cluster proximity (C_i, C_j) of clusters C_i and C_j , which are of size m_i and m_j , respectively, is expressed by the following equation: [10]

$$\text{proximity}(C_i, C_j) = \frac{\sum_{x \in C_i, y \in C_j} \text{proximity}(x, y)}{m_i * m_j} \quad (3.11)$$

Centroid clustering method

Centroid methods calculate the proximity between two clusters by calculating the distance between the centroids of clusters.

Ward's method

For Ward's method, the proximity between two clusters is defined as the increase in the squared error that results when two clusters are merged. [10]

K-means

K-means defines a prototype in terms of a centroid which is usually the mean of a group of points. It is typically applied to objects in continuous n -dimensional space. K-means is one of the oldest and most widely used clustering algorithms. [10]

The basic K-Means Algorithm

The K-means clustering technique is simple. The research initially chooses K to be initial centroids, where K is a user specified parameter, namely, the number of clusters desired. Each point is then assigned to the closet centroid and each collection of points assigned to a centroid is a cluster. The centroid of each cluster is then updated based on the points assigned to the cluster. The research repeats the assignment and update steps until no point changes cluster or equivalently, until the centroids remain the same. K-means is formally described by Table 3.2 below:

Table 3.2 Basic K-means Algorithm

Basic K-means Algorithm

1. Select K points as initial centroids.
 2. repeat
 3. Form K clusters by assigning each to its closet centroid.
 4. Recomputed the centroid of each cluster.
 5. Until Centroids do not change.
-

The operation of K-means is still illustrated in following Figure 3.11 [10]

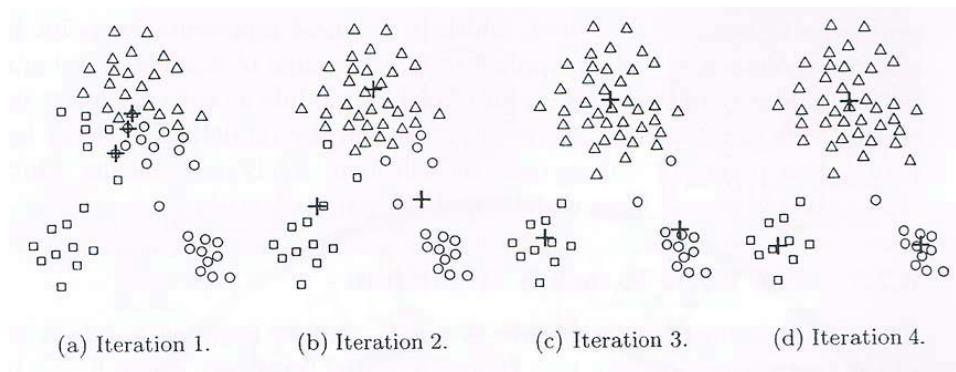


Figure 3.11 Using the K-means algorithm to find 3 clusters in sample data [10]

Figure 3.11 showed how to start from 3 centroids. Final clusters are found in 4 assignment-update steps. Each subfigure shows:

- (1) The centroids at the start of iteration.
- (2) The assignment of points to those centroids.

Centroids are indicated by “+” symbol. All points belong to the same clusters and have the same shape.

In the first step, points are assigned to the initial centroids which are all in the larger group of points. The research use mean as centroid. After points are assigned to a centroid, centroid is then updated. Again, figure of each step shows centroid at the beginning step and assignment of points to those centroids.

In the second step, points are assigned to the updated centroids, centroids are updated again.

In step 2, 3 and 4 which are shown in Figure 3.11(b), (c) and (d) respectively. Two of centroids move to the two small groups of points at the bottom of Figure. K-means Algorithm terminates in Figure 3.11(d) because no more changes occur. Centroids have identified natural groupings of points.

K-means always converges to a solution for combinations of proximity functions and types of centroids. K-means reaches a state in which no points are shifting from one cluster to another. Thus centroids don’t change because most of the convergence occurs in early steps.

Centroids and Objective Functions

Since centroid can vary, depending on proximity measure for data and goal of clustering. Goal of clustering is typically expressed by objective function that depends on proximities of points to one another or to cluster centroids. However, key point is once the research has specified a proximity measure and an objective function, centroid that we should choose can often be determined mathematically.

Data in Euclidean Space

The research use **Sum of the Squared Error (SSE)** or scatter. In other words, the research calculate the error of each data point such as its Euclidean distance to the closet centroid and then compute the total sum of the squared errors. Given 2 different sets of clusters that are produced by 2 different runs of K-means, the smallest squared error is better because prototypes of this clustering are better representation of points in their cluster.

Table 3.3 Table of notation

Symbol	Description
x	An object.
C_i	The i^{th} cluster.
c_i	The centroid of cluster C_i .
c	The centroid of all points.
m_i	The number of objects in the i^{th} cluster.
m	The number of objects in the data set.
K	The number of clusters.

From using notation in Table 3.3, then SSE is formally defined as follows:

$$\text{SSE} = \sum_{i=1}^K \sum_{x \in C_i} \text{dist}(c_i, x)^2 \quad (3.12)$$

Where dist is standard Euclidean (L_2) distance between 2 objects in Euclidean space.

Using the notation in Table 3.3, centroid (mean) of the i^{th} is defined by equation as below:

$$\mathbf{c}_i = \frac{1}{m_i} \sum_{\mathbf{x} \in C_i} \mathbf{x} \quad (3.13)$$

Centroid of a cluster contains the three two-dimensional points, (1,1), (2,3) and (6,2) is $((1+2+6)/3, ((1+3+2)/3) = (3,2)$.

K-means algorithm directly attempt to minimize the SSE by forming clusters with assigning points to their nearest centroid and then recomputing centroids so as to further minimize SSE. However these actions are only guaranteed to find a local minimum with respect to SSE since they are based on optimizing SSE for specific choices of centroids and clusters rather than for all possible choices.

General Case

Number of choices for proximity function, centroid and objective function can be used in the basic K-means algorithm and that are guaranteed to converge.

Table 3.4 K-means: Common choices for proximity, Centroids and objective functions

Proximity Function	Centroid	Objective Function
Manhattan (L_1)	median	Minimize sum of the L_1 distance of an object to its cluster centroid
Squared Euclidean (L_2^2)	mean	Minimize sum of the squared L_2 distance of an object to its cluster centroid
Cosine	mean	Maximize sum of the cosine similarity of an object to its cluster centroid
Bregman divergence	mean	Minimize sum of the Bregman divergence of an object to its cluster centroid

From Table 3.4 shows that Manhattan (L_1) distance and objective of minimizing the sum of distance. Then appropriate centroid is the median of points in a cluster. Bregman divergence is actually a class of proximity measures that includes squared Euclidean distance (L_2^2), Mahalanobis distance and cosine similarity. The importance of Bregman divergence functions is any function can be used as basis of a

K-means style clustering algorithm with the mean as centroid. If the research uses a Bregman divergence as proximity function then result of clustering algorithm has the usual properties of K-means with respect to convergence and local minimum. Furthermore, properties of a clustering algorithm can be developed for all possible Bregman divergences. K-means algorithms use cosine similarity or squared Euclidean distance is particular instances of a general clustering algorithm based on Bregman divergences.

For the rest of –K-means discussion, the research uses two-dimensional data because it is easy to explain K-means and properties for data. Anyway, K-means is very general clustering algorithm and can be used with a wide variety of data such as documents and time series.

Choosing Initial Centroids

Random initialization of centroids is used. Then different runs of K-means typically produce different total SSEs. Following Figure 3.17(a) shows a clustering solution that is the global minimum of SSE for three clusters. Figure 3.17(b) shows a suboptimal clustering that is only a local minimum.

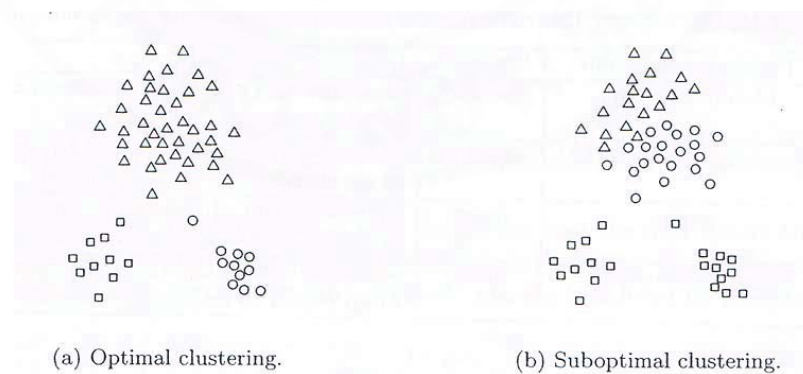


Figure 3.12 Three optional and non-optimal clusters [10]

Common approach is to choose the initial centroid randomly but the resulting clusters are often poor.

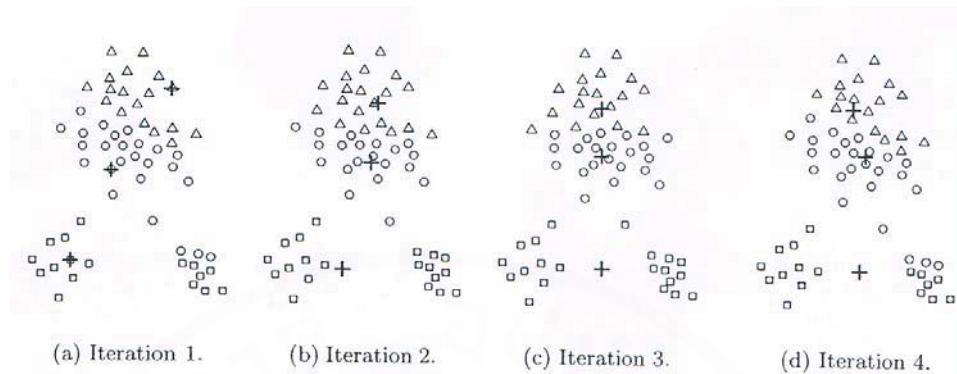


Figure 3.13 Poor starting centroids for K-means [10]

From Figure 3.13 above shows clusters that result from two particular choices of initial centroids. Though initial centroids seem to be better distributed, the research obtains a suboptimal clustering with higher squared error.

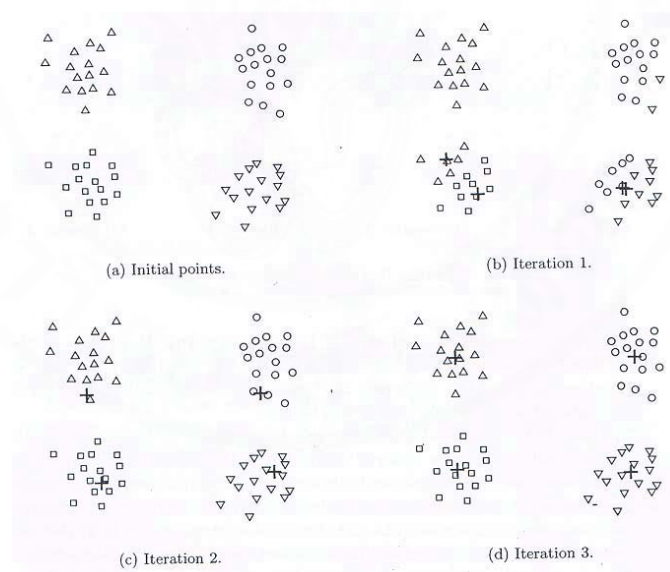


Figure 3.14 Two pairs of clusters with a pair of initial centroids within each pair of clusters [10]

From Figure 3.14 above Figure (a) the data consists of two pairs of clusters where the clusters in each pair (top-bottom) are closer to each other than to clusters in other pair. Figure (b-d) shows that if the research starts with two initial centroids per pair of clusters then both centroids are in a single cluster. For following Figure 3.15 showed that if a pair of clusters has only one initial centroid and other pair has three then two of true clusters will be combined and one true cluster will be split.

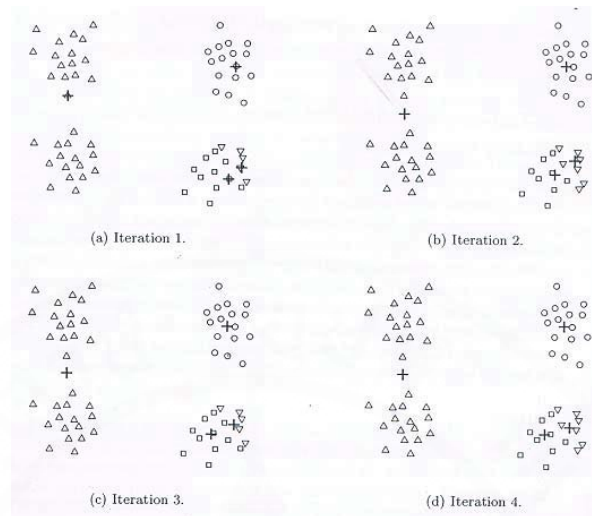


Figure 3.15 Two pairs of clusters with more or fewer than two initial centroids within a pair of clusters [10]

Time and Space Complexity

Space requirements for K-means are modest because only data points and centroids are stored. The storage required is $O((m + K)n)$ where m is number of points and n is number of attributes. Time requirements for K-means are also modest. Time required is $O(I * K * m * n)$ where I is number of iterations required for convergence. I is often small and can usually be safely bounded. Thus, K-means is linear in m (number of points) and is efficient as well as simple provided that K (number of clusters) is significantly less than m .

3.1.3. Similarity Measures

Cluster was basically defined by similarity. Then the research found that it was significant to all procedures of clustering to measure similarity between two patterns under the same feature. If the types and scales of features were different, then measurement of distance must be selected watchfully. We usually calculate on dissimilarity of two patterns by measurement of distance which defined on feature space [5].

Euclidean Distance

The most popular metric for continuous features was the Euclidean distance

$$d_2(d_i, d_j) = \sqrt{\sum_{k=1}^n (w_{i,k} - w_{j,k})^2} \quad (3.14)$$

Where d_i, d_j were the documents in a collection of documents D

k were set of occurring terms in d_i, d_j

$w_{i,k}$ were weighted value of term k^{th} in document d_i

The Euclidean distance was instinctive appeal. It was commonly used to evaluate the proximity of objects in two or three-dimensional spaces. It worked significantly when data set had compact or isolated clusters [5]. The well result of using Euclidean distance was on raw data anyway it might be very unsatisfactory if it was affected extremely by changing the scale of a variable [11].

3.2 Research methodology

The objective of this research was to analysis the O-NET, GAT scores for each faculty in Mahidol University for factor analysis and cluster analysis to recommend student to choose faculty in undergraduate education. Another objective was to help teacher to get appropriate student. The research methodology of this study was shown in Figure 3.16 as below:

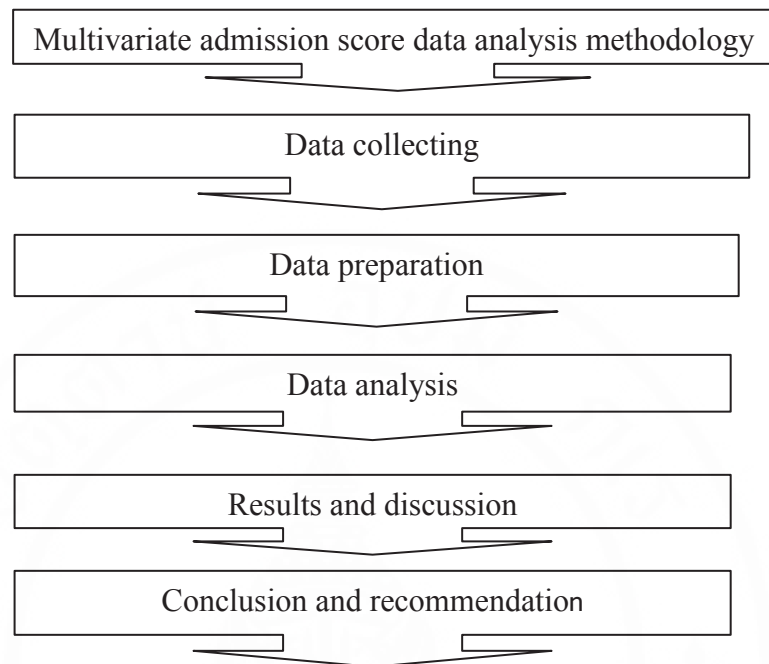


Figure 3.16 Research methodology

From figure above can describe as below:

- Multivariate admission score data analysis methodology
- Data collecting

The data used in this study is secondary data which collected information from Division of Information Technology (MUIT), Office of the President, Mahidol University.

- Data preparation

The researcher got data in digital file. Then researcher inputted data into spreadsheet program (Microsoft Excel 2007).

- Data analysis

After preparing data, bring data into SPSS Version 16 program. Factor Analysis along with Cluster Analysis was utilized.

- Results and discussion

After completion of data analysis by factor analysis and clustering analysis, then explain the result.

- Conclusion and recommendation

Summarized result and also suggested more useful experiment and interesting research for further education.

3.3 Materials and research tools

3.3.1 Hardware

Computer Name	:	Acer aspire
Processor	:	Intel Core 2 Duo 2.80 GHz
RAM	:	DDRII 2 GB
Hard Drive	:	320 GB

3.3.2 Software

Statistical tools	:	SPSS Version 16
Operating Systems	:	Microsoft Window Vista

3.4 Schedule of research

Table 3.5 Schedule of research

Schedule of research	2011-2012			
	July(2011)	Aug(2011)	Sep-Dec(2011)	Jan(2012)-Sep
1.Planning and collecting data	←————→			
2. Data analysis			←————→	
3. Result evaluation				←————→
4. Result of Factor analysis				←————→
5. Result of Cluster analysis				←————→
6. Documentation				←————→

CHAPTER IV

RESULT

4.1 Result of weight value of Mahidol University (Preliminary results)

4.1.1 Factor Analysis

Table 4.1 Descriptive Statistics from factor analysis

	Mean	Std. Deviation	Analysis N
GPA (Thai)	3.78	.68	36
GPA (Social)	3.67	1.04	36
GPA (Eng)	4.06	1.58	36
GPA (Math)	3.94	1.09	36
GPA (Sci)	4.00	1.49	36
GPA (Physical)	.56	2.32	36
O-NET(Thai)	7.50	.87	36
O-NET(Social)	7.50	.87	36
O-NET(Eng)	7.92	2.20	36
O-NET(Math)	6.67	2.08	36
O-NET(Sci)	6.67	2.08	36
A-NET(Thai)	1.25	6.02	36
A-NET(Eng)	7.92	6.58	36
A-NET(Math)	10.42	3.85	36
A-NET(Sci)	12.22	6.48	36
Engineering	1.94	4.01	36

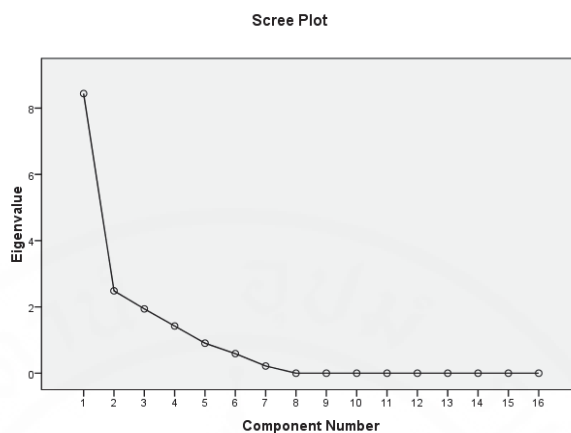


Figure 4.1 Scree Plot by factor analysis

Table 4.2 Rotated Component Matrix from factor analysis

	Component			
	1	2	3	4
O-NET(Thai)	.95			
O-NET(Social)	.95			
O-NET(Eng)	.95			
GPA (Sci)	-.91			
GPA (Math)	-.81			
O-NET(Sci)	-.80	-.29	.39	
O-NET(Math)	-.80	-.29	.39	
GPA (Eng)	.79	.48		
A-NET(Sci)	-.72			-.63
A-NET(Thai)	.69			-.35
GPA (Social)	.22	.97		
GPA (Physical)		-.95		
GPA (Thai)	.47	.83		
Engineering			.93	
A-NET(Eng)	.22		-.84	.31
A-NET(Math)	-.45			.78

Discussion on weight value of Mahidol University (Preliminary result)

Academic Year 2009

From the research found that four factors could be analyzed and blocked. Result could be described as shown in Table 1 about means and standard deviations. Variable of science, mathematics and English have highly means. It means that variable of science and English were influenced to exam into Mahidol University which is the science university.

Table 4.3 Means and Standard deviation of 16 variables

	Mean	Std. Deviation	Analysis N
Thai (GPA)	3.78	0.68	36
Social (GPA)	3.67	1.04	36
Foreign (GPA)	4.06	1.58	36
Math (GPA)	3.94	1.09	36
Science (GPA)	4.00	1.49	36
Health & Physical (GPA)	0.56	2.32	36
Thai (O-NET)	7.50	0.87	36
Social (O-NET)	7.50	0.87	36
English(O-NET)	7.92	2.20	36
Math (O-NET)	6.67	2.08	36
Science(O-NET)	6.67	2.08	36
Thai (A-NET)	1.25	6.02	36
English (A-NET)	7.92	6.58	36
Math(A-NET)	10.42	3.85	36
Science (A-NET)	12.22	6.48	36
Basic of Engineer	1.94	4.01	36

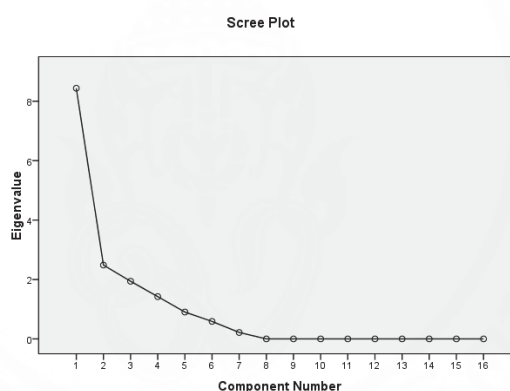
Sixteen variables of weight values, for exam subjects into Mahidol University, were analyzed by Eigen values which values were more than 1. The analysis found that four variables could be blocked as shown in Table 2. In Table 3 four rotated factors are presented Eigen values. All factors were still perpendicular when rotated. Varimax was the method of factor rotation. Eigen value, percentage of variance and cumulative % of 1st rotated factor were less than unrotated factors. Eigen value and % of Variance of 2nd factor and 3rd factor were more than 1st factor.

Table 4.4 Total Variance Explained (Initial Eigen values)

Component	Total	% of Variance	Cumulative%
1	8.43	52.73	52.73
2	2.48	15.53	68.26
3	1.94	12.13	80.40
4	1.42	8.90	89.30

Table 4.5 Total Variance Explained (Rotation Sums of Squared Loadings)

Component	Total	% of Variance	Cumulative%
1	7.74	48.40	48.40
2	3.04	19.02	67.43
3	2.04	12.78	80.22
4	1.45	9.08	89.30

**Figure 4.2** Relation of Eigen value for each factor and Component Number

From above figure found that Eigen value of component 1, 2, 3 and 4 were more than 1. Then it was the reason for considering on four factors. Furthermore, from result as shown in Table 1, 2, 3, 4 and 5 could insist result of analysis on four factors. From Table 4 found that after rotated the Factor Loading, then 1st factor consisted of 10 variables. The 2nd factor consisted of 3 variables. The 3rd factor consisted of 2 variables and the 4th factor consisted of one variable.

From Table 4.3 found that the Factor Score of all 16 variables could insist result of analysis for those four factors.

Table 4.6 Rotated Component Matrix

	Component			
	1	2	3	4
Thai (O-NET)	.95			
Social (O-NET)	.95			
English (O-NET)	.95			
Science (GPA)	-.91			
Math (GPA)	-.81			
Science (O-NET)	-.80	-.29	.39	
Math (O-NET)	-.80	-.29	.39	
Foreign Language (GPA)	.79	.48		
Science (A-NET)	-.72			-.63
Thai (A-NET)	.69			-.35
Social (GPA)	.22	.97		
Health & Physical (GPA)		-.95		
Thai (GPA)	.47	.83		
Basic Engineer			.93	
English (A-NET)	.22		-.84	.31
Math (A-NET)	-.45			.78

Table 4.7 Component Score Coefficient Matrix

	Component			
	1	2	3	4
Thai (GPA)	.00	.27	.02	.12
Social (GPA)	-.04	.34	.02	.03
Foreign Language (GPA)	.06	.11	-.04	-.09
Math (GPA)	-.11	.07	.04	.03
Science (GPA)	-.13	.04	.01	-.16
Health & Physical (GPA)	.11	-.37	-.01	.10
Thai (O-NET)	.15	-.07	.14	.02
Social (O-NET)	.15	-.07	.14	.02
English (O-NET)	.12	-.02	.01	-.03
Math (O-NET)	-.07	-.04	.15	.10
Science (O-NET)	-.07	-.04	.15	.10
Thai (A-NET)	.08	.00	.04	-.22
English (A-NET)	.00	-.01	-.41	.21
Math (A-NET)	-.04	.03	-.03	.52
Science (A-NET)	-.12	.02	-.04	-.46
Basic Engineer	.06	.03	.48	.13

From result that shown in above table found that all variables were variables of core subjects. Three factors in the 2nd factors as shown in Table 7 found that all variables were variable of weight value for learning subjects and then the 2nd factors were named as learning subjects. Two factors of the 3rd factors were variable of professional subjects and advanced. Then it has shown that professional

subjects were very significant to further education in every faculty. Then the 3rd factors were named as professional subjects and advanced English. Weight value of advanced Mathematics (A-NET) was the only variable which separated from other variables. It has shown that advanced Mathematics (A-NET) was important for university exam. Then the 4th factor was named advanced Mathematics.

Table 4.8 The 1st factor: Core subjects

Ordering	Subject name
1	Foreign Language (GPA)
2	Mathematics (GPA)
3	Science (GPA)
4	Thai (O-NET)
5	Social and Culture (O-NET)
6	English (O-NET)
7	Mathematics (O-NET)
8	Science (O-NET)
9	Thai (A-NET)
10	Science (A-NET)

Table 4.9 The 2nd factor: Learning subjects (GPA)

Ordering	Subject name
1	Social, Religion and Culture (GPA)
2	Thai Language (GPA)
3	Health & Physical (GPA)

Table 4.10 The 3rd factor: Professional subjects and Advanced English

Ordering	Subject name
1	Basic Engineer
2	Advanced English (A-NET)

Table 4.11 The 4th factor: Advanced Mathematics

Ordering	Subject name
1	Advanced Mathematics (A-NET)

4.1.2 Cluster Analysis

***** HIERARCHICAL CLUSTER ANALYSIS *****

Dendrogram using Average Linkage (Between Groups)

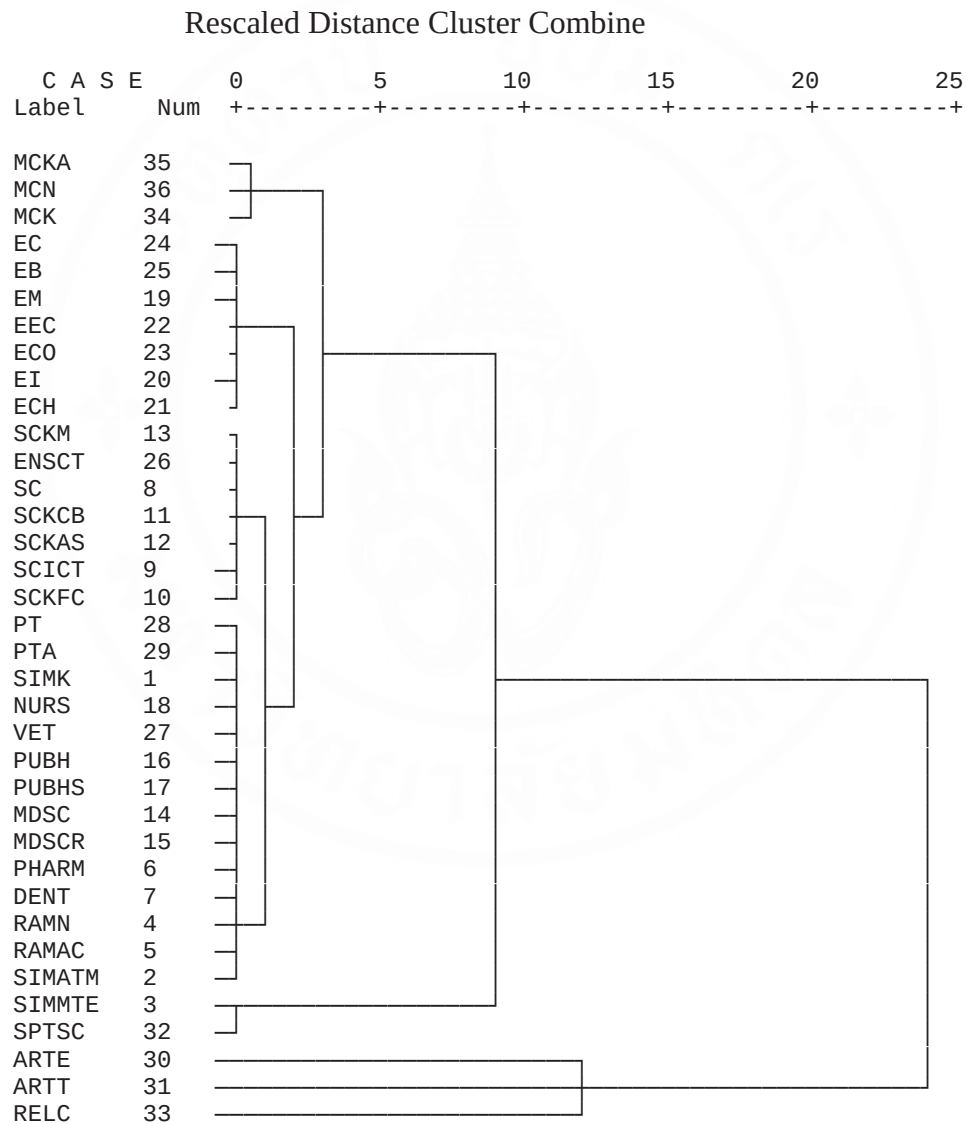


Figure 4.3 Hierarchical Cluster Analysis

Discussion on weight value of Mahidol University (Using Hierarchical Clustering) (Preliminary result)

Academic Year 2009

After completion of factor analysis and then applied cluster analysis, experiment shown that combination method of faculty/department, by using weight value of exam subjects and using characteristics of faculty/department of Mahidol University, was Average Linkage Between Group. Clusters can be grouped to 6 clusters as group of Administration, group of Engineering, group of Science, Group of Health Sciences & Medicine, group of Sport Science & Medical Technology and group of Fine Art.

4.2 Result of Admission score (New Data Sets) of Mahidol University

4.2.1 Factor Analysis

Academic year 2010 (New Data Sets)

From Factor Analysis on admission score of O-NET on all subjects and GAT on academic year 2010, then the results are below:

Table 4.12 Descriptive Statistics for factor analysis (academic year 2010)

Variable Name	Mean	Std. Deviation	Analysis (N)
O-NET(Thai)	67.50	8.34	1037
O-NET(Social)	50.46	8.03	1037
O-NET(Eng)	42.56	12.66	1037
O-NET(Maths)	55.94	16.85	1037
O-NET(Sci)	48.67	12.08	1037
O-NET(Physical)	52.16	7.14	1037
O-NET(Arts)	45.57	7.10	1037
O-NET(Jobs)	41.28	9.30	1037
GAT	203.37	29.21	1037

Table 4.13 Correlation Matrix for factor analysis (academic year 2010)

	O-NET	O-NET	O-NET	O-NET	O-NET	O-NET	O-NET	O-NET	GAT
	(Thai)	(Social)	(Eng)	(Maths)	(Sci)	(Physical)	(Arts)	(Jobs)	
O-NET(Thai)	1.00	.57	.45	.45	.50	.24	.24	.22	.52
O-NET(Social)	.57	1.00	.36	.41	.52	.18	.17	.17	.42
O-NET(Eng)	.45	.36	1.00	.35	.37	.15	.21	.14	.64
O-NET(Math)	.45	.41	.35	1.00	.69	.29	.16	.22	.39
O-NET(Sci)	.50	.52	.37	.69	1.00	.30	.18	.24	.41
O-NET(Physical)	.24	.18	.15	.29	.30	1.00	.20	.28	.14
O-NET(Arts)	.24	.17	.21	.16	.18	.20	1.00	.16	.18
O-NET(Jobs)	.22	.17	.14	.22	.24	.28	.16	1.00	.10
GAT	.52	.42	.64	.39	.41	.14	.18	.10	1.00

Table 4.14 Total Variance Explained for factor analysis (academic year 2010)

Total Variance Explained									
Initial Eigenvalues				Extract Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	%Variance	%Cumulative	Total	% Variance	%Cumulative	Total	%Variance	%Cumulative
1	3.70	41.15	41.15	3.70	41.15	41.15	3.13	34.79	34.79
2	1.19	13.22	54.37	1.19	13.22	54.37	1.76	19.57	54.37
3	.94	.94	10.53	.64	6.90				
4	.76	.76	8.54	.45	73.45				
5	.71	.71	7.95	.40	81.40				
6	.63	.63	7.01	.42	88.42				
7	.41	.41	4.57	.99	92.99				
8	.34	.34	3.77	.76	96.76				
9	.29	.29	3.23	100.00					

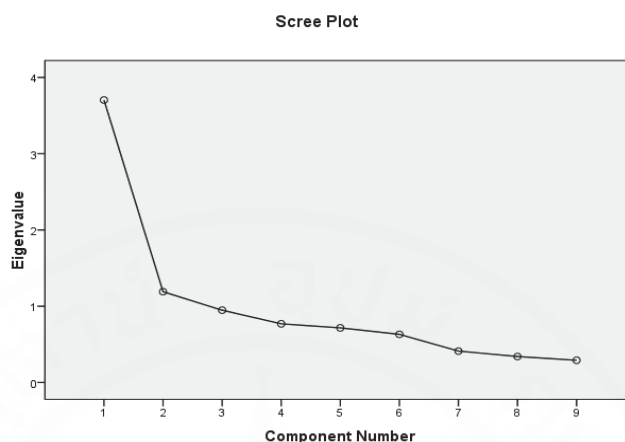


Figure 4.4 Scree Plot for factor analysis (academic year 2010)

Table 4.15 Rotated Component Matrix for factor analysis (academic year 2010)

	Component	
	1	2
GAT	.82	
O-NET (Eng)	.76	
O-NET (Thai)	.73	.26
O-NET (Social)	.69	.22
O-NET (Sci)	.64	.44
O-NET (Math)	.60	.43
O-NET (Phyisc)		.74
O-NET (Jobs)		.71
O-NET (Arts)		.43

Discussion of Admission score (New Data Sets)

Academic Year 2010

From experiment found that variable of **O-Net (Math)** and **O-NET (Sci)** were the most related. Correlation coefficient was 0.692. Then **O-NET (Math)** and **O-NET (Sci)** was arranged into the same factor. Anyway, variable of **O-NET (Thai)** was related to variable of **O-NET (Social)** ($r=0.573$), thus they were arranged to the same factor too.

From Table of Total Variance Explained should have two factors because the Eigen values were more than 1. The most important factor was the 1st factor because variance of data was 41.154%. The 2nd factor was fewer important.

From Table of Component Matrix found that the 1st component should consisted of variable of **O-NET (Sci)**, **O-NET (Thai)**, **O-NET**

(Math), GAT, O-NET (Social), O-NET (Eng) and O-NET (Arts). The 2nd component should consist of O-NET (Jobs) and O-NET (Physics). By the way, from the result found that the Factor Loading for variable of O-NET (Arts) in all factors was not significantly different. The three factors were still perpendicular when rotated. After rotated found that two factors were still arranged variables. The 1st component was consisted of variable of GAT, O-NET (Eng), O-NET (Thai), O-NET (Social), O-NET (Sci) and O-NET (Math). The 2nd component was consisted of variable of O-NET (Physics), O-NET (Jobs) and O-NET (Arts).

Academic Year 2011

From Factor Analysis on admission score of O-NET on all subjects and GAT on academic year 2011, then the results are below:

Table 4.16 Descriptive Statistics for factor analysis (academic year 2011)

	Mean	Std. Deviation	Analysis N
O-NET (Eng)	63.00	8.71	1123
O-NET (Math)	58.13	7.82	1123
O-NET (Sci)	40.44	13.47	1123
O-NET (Physsc)	41.72	19.80	1123
O-NET (Arts)	48.92	11.42	1123
O-NET (Jobs)	67.50	8.88	1123
O-NET (Eng)	40.44	7.12	1123
O-NET (Math)	53.09	8.98	1123
GAT	223.78	28.76	1123

Table 4.17 Correlation Matrix for factor analysis (academic year 2011)

	O-NET	O-NET	O-NET	O-NET	O-NET	O-NET	O-NET	O-NET	GAT
	(Thai)	(Social)	(Eng)	(Maths)	(Sci)	(Physical)	(Arts)	(Jobs)	
O-NET(Thai)	1.00	.43	.41	.39	.49	.11	.31	.17	.48
O-NET(Social)	.43	1.00	.31	.26	.50	.37	.08	.38	.40
O-NET(Eng)	.41	.31	1.00	.37	.41	.09	.21	.16	.74
O-NET(Maths)	.39	.26	.37	1.00	.63	.05	.18	.09	.36
O-NET(Sci)	.49	.50	.41	.63	1.00	.24	.25	.25	.45
O-NET(Physical)	.11	.37	.09	.05	.24	1.00	.01	.42	.17
O-NET(Arts)	.31	.08	.21	.18	.25	.01	1.00	.05	.21
O-NET(Jobs)	.17	.38	.16	.09	.25	.42	.05	1.00	.18
GAT	.48	.40	.74	.36	.45	.17	.21	.18	1.00

Table 4.18 Total Variance Explained for factor analysis (academic year 2011)**Total Variance Explained**

Component	Initial Eigenvalues			Extract Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	%Variance	%Cumulative	Total	%Variance	%Cumulative	Total	%Variance	%Cumulative
1	3.57	39.68	39.68	3.57	39.68	39.68	3.08	34.31	34.31
2	1.42	15.88	55.56	1.42	15.88	55.56	1.91	21.24	55.56
3	.91	10.13	65.70						
4	.87	9.74	75.44						
5	.63	7.06	82.51						
6	.57	6.40	88.91						
7	.46	5.11	94.02						
8	.29	3.28	97.31						
9	.2	2.69	100.00						

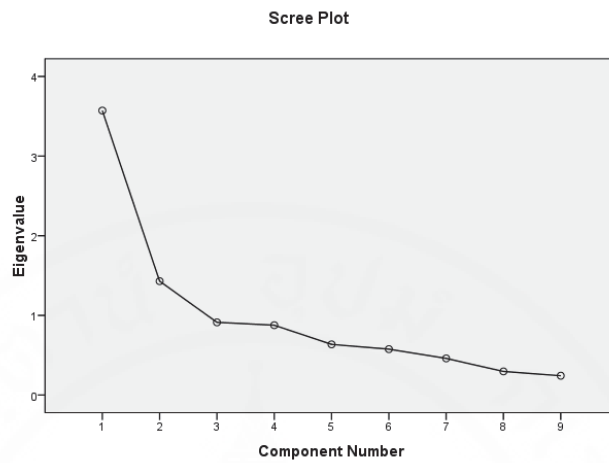


Figure 4.5 Scree Plot for factor analysis (academic year 2011)

Table 4.19 Rotated Component Matrix for factor analysis (academic year 2011)

	Component	
	1	2
GAT	.76	
O-NET (Eng)	.75	
O-NET (Sci)	.72	.33
O-NET(Thai)	.71	
O-NET (Math)	.70	
O-NET (Arts)	.47	
O-NET(Phycs)		.81
O-NET (Jobs)		.78
O-NET (Social)	.42	.65

From the experiment found that from Table of Correlation Matrix showed variable of GAT and variable of O-NET (Eng) were the most related. Correlation Coefficient was 0.746. Then variable of GAT and variable of O-NET (Eng) should be arranged to the same factor. Variable of O-NET (Math) and variable of O-NET (Science) were highly related ($r=0.636$), then should be arranged to the same factor too.

From Table of Total Variance Explained should have two factors because Eigen valued was more than 1. The most significant factor was the 1st factor because explanation of variance was 39.687%. The 2nd factor is fewer important.

From Table of Component Matrix found that the 1st component should consist of variable of O-NET (Science), GAT, O-NET (Thai), O-NET (Social), O-NET (Eng) and O-NET (Math). The 2nd component should consist of variable of O-NET (Jobs) and O-NET (Physics). But experiment shown that factor loading of variable of O-NET (Arts) in all factor was not significant different. Three factors were perpendicular when rotated. After rotated found that two factors were be arranged. The 1st component consisted of variable of GAT, O-NET (Eng), O-NET (Thai), O-NET (Science), O-NET (Math) and O-NET (Arts). The 2nd component consisted of variable of O-NET (Physics), O-NET (Jobs) and O-NET (Social).

Academic Year 2012 (New Data Sets)

From Factor Analysis on admission score of O-NET on all subjects and GAT on academic year 2012, then the results are below:

Table 4.20 Descriptive Statistics for factor analysis (academic year 2012)

	Mean	Std. Deviation	Analysis N
O-NET (Eng)	59.40	7.98	1478
O-NET (Math)	47.90	9.24	1478
O-NET (Sci)	41.84	13.81	1478
O-NET (Phyisc)	47.80	18.01	1478
O-NET (Arts)	45.01	10.88	1478
O-NET (Jobs)	63.45	7.01	1478
O-NET (Eng)	38.26	7.30	1478
O-NET (Math)	58.99	8.29	1478
GAT	229.13	28.60	1478

	Mean	Std. Deviation	Analysis N
O-NET (Eng)	59.40	7.98	1478
O-NET (Math)	47.90	9.24	1478
O-NET (Sci)	41.84	13.81	1478
O-NET (Phyisc)	47.80	18.01	1478
O-NET (Arts)	45.01	10.88	1478
O-NET (Jobs)	63.45	7.01	1478
O-NET (Eng)	38.26	7.30	1478
O-NET (Math)	58.99	8.29	1478

Table 4.21 Correlation Matrix for factor analysis (academic year 2012)

	O-NET (Thai)	O-NET (Social)	O-NET (Eng)	O-NET (Maths)	O-NET (Sci)	O-NET (Physical)	O-NET (Arts)	O-NET (Jobs)	GAT
O-NET(Thai)	1.00	.47	.38	.30	.43	.20	.32	.12	.45
O-NET(Social)	.47	1.00	.25	.20	.50	.31	.26	.03	.34
O-NET(Eng)	.38	.25	1.00	.39	.	.08	.27	.16	.72
O-NET(Maths)	.30	.20	.39	1.00	.62	.08	.15	.17	.36
O-NET(Sci)	.43	.50	.35	.62	1.00	.23	.25	.15	.39
O-NET(Physical)	.20	.31	.08	.08	.23	1.00	.16	.13	.14
O-NET(Arts)	.32	.26	.27	.15	.25	.16	1.00	.13	.27
O-NET(Jobs)	.12	.03	.16	.17	.15	.13	.13	1.00	.17
GAT	.45	.34	.72	.36	.39	.14	.27	.17	1.00

Table 4.22 Total Variance Explained for factor analysis (academic year 2012)**Total Variance Explained**

Component	Initial Eigenvalues			Extract Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	%Variance	%Cumulative	Total	%Variance	%Cumulative	Total	%Variance	%Cumulative
1	3.70	41.15	41.15	3.70	41.15	41.15	3.13	34.79	34.79
2	1.19	13.22	54.37	1.19	13.22	54.37	1.76	19.57	54.37
3	.94	.94	10.53	64.90					
4	.76	.76	8.54	73.45					
5	.71	.71	7.95	81.40					
6	.63	.63	7.01	88.42					
7	.41	.41	4.57	92.99					
8	.34	.34	3.77	96.76					
9	.29	.29	3.23	100.00					

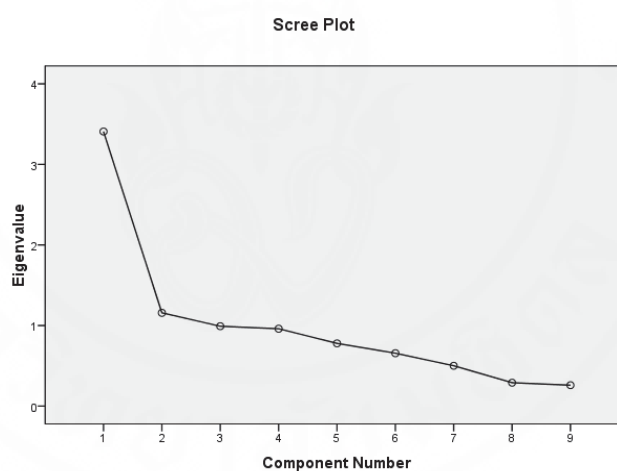
**Figure 4.6** Scree Plot for factor analysis (academic year 2012)

Table 4.23 Rotated Component Matrix for factor analysis (academic year 2012)

	Component	
	1	2
O-NET (Eng)	.83	
GAT	.79	.21
O-NET (Math)	.68	
O-NET (Jobs)	.33	
O-NET (Social)		.77
O-NET (Phyisc)		.73
O-NET (Thai)	.45	.56
O-NET (Sci)	.53	.53
O-NET (Arts)	.27	.44

From the experiment found that from Table of Correlation Matrix showed variable of O-NET (Math) and variable of GAT were the most related. Correlation Coefficient was 0.727. Then variable of GAT and variable of O-NET (Math) should be arranged to the same factor. Variable of O-NET (Math) and variable of O-NET (Science) were highly related ($r=0.620$), then should be arranged to the same factor too.

From Table of Total Variance Explained should have two factors because Eigen valued was more than 1. The most significant factor was the 1st factor because explanation of variance was 37.865%. The 2nd factor is fewer important.

From Table of Component Matrix found that the 1st component should consist of variable of O-NET (Science), GAT, O-NET (Math), O-NET (Eng), O-NET (Thai), O-NET (Social), O-NET (Arts) and O-NET (Jobs). The 2nd component should consist of variable of O-NET (Physics). By the way, experiment shown that factor loading of variable of O-NET (Social) and O-NET (Math) in all factors were not dramatically differ, thus three factors were perpendicular when rotated. After rotated found that two factors were be arranged. The 1st component consisted of variable of GAT, O-NET (Eng), O-NET (Thai), O-NET (Social), O-NET (Math) and O-NET (Jobs). The 2nd component consisted of variable of O-NET (Physics), O-NET (Social)

and O-NET (Arts). O-NET (Science) could not arrange to any component because it had different value of factor loading.

4.2.2 Cluster Analysis

Academic year 2010 (New Data Sets Using K-Means Clustering Algorithm)

From Cluster Analysis on admission score of O-NET on all subjects and GAT on academic year 2010, then the results are below:

Table 4.24 Final Cluster Centers for cluster analysis (academic year 2010)

	Cluster			
	1	2	3	4
Zscore O-NET (Thai)	.14	-.06	-1.00	.99
Zscore O-NET (Social)	.11	.56	-.88	.91
Zscore O-NET (Eng)	-.04	.66	-.78	1.21
Zscore O-NET (Math)	.12	-1.59	-.96	1.00
Zscore O-NET (Sci)	.11	-1.21	-1.00	1.08
Zscore O-NET (Phyisc)	.09	-7.30	-.53	.51
Zscore O-NET (Arts)	.00	-6.41	-.40	.57
Zscore O-NET (Jobs)	.08	-4.43	-.44	.40
Zscore GAT	.10	.48	-.92	.98

Table 4.25 Distances between Final Cluster Centers for cluster analysis (year 2010)

Cluster	1	2	3	4
1		11.04	2.67	2.46
2	11.04		10.26	12.13
3	2.67	10.26		5.09
4	2.46	12.13	5.09	

Table 4.26 ANOVA for cluster analysis (academic year 2010)

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Zscore O-NET (Thai)	162.98	3	.53	1033	307.75	.00
Zscore O-NET (Social)	131.06	3	.62	1033	210.63	.00
Zscore O-NET (Eng)	155.73	3	.55	1033	282.82	.00
Zscore O-NET (Math)	156.34	3	.54	1033	284.86	.00
Zscore O-NET (Sci)	174.87	3	.49	1033	353.25	.00
Zscore O-NET (Phyisc)	63.75	3	.81	1033	77.96	.00
Zscore O-NET (Arts)	50.34	3	.85	1033	58.76	.00
Zscore O-NET (Jobs)	36.83	3	.89	1033	41.11	.00
Zscore GAT	145.36	3	.58	1033	250.31	.00

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Table of Final Cluster Centers

It showed standardized averages. These averages were means of each cluster. Average on variable of O-NET (Thai) differed when they were in different cluster. It has dramatically differed when compared to other variables. Average on variable of O-NET (Thai) in the 1st cluster = 0.14275. It was more than sum of average = 0.14275 times of standard deviation. Average on variable of O-NET (Thai) in the 3rd cluster = -1.00181. It was less than sum of average = 1.00181 times of standard deviation. In the same way, other variables also had different average when they were in different cluster.

Table of Distances between final cluster centers

In this table all figures were distance between means of four clusters. In 1st cluster had the longest distance from 2nd cluster = 11.045 and it had the nearest distance to 4th cluster = 2.464. The 3rd cluster and 4th cluster had the nearest distance to 2nd cluster.

Table of ANOVA

This table showed Between-cluster Mean Square and Mean Square Error or Within-Cluster Mean Square. It showed statistical data at F. Anyway it would not use statistical at F and Significance on the last column of table. While testing on differentiation of average of each variable, which were on different cluster, found that average on variable of O-NET (Science) were dramatically differed. It had the most statistical $F = 353.250$. Variable of O-NET (Thai) was second rank. It had statistical $F = 307.751$. For variable O-NET (Jobs) had lest differentiation when they were on different cluster.

Academic year 2011(New Data Sets Using K-Means Algorithm)

From Cluster Analysis on admission score of O-NET on all subjects and GAT on academic year 2011, then the results are below:

Table 4.27 Final Cluster Centers by K-means algorithm (academic year 2011)

	Cluster			
	1	2	3	4
Zscore O-NET (Thai)	.80	1.23	.12	-1.04
Zscore O-NET (Social)	.85	.078	.08	-.98
Zscore O-NET (Eng)	.98	1.22	-.02	-.87
Zscore O-NET (Math)	1.00	.06	-.04	-.81
Zscore O-NET (Sci)	1.08	1.62	.04	-1.11
Zscore O-NET (Phyisc)	.37	-7.59	.09	-.55
Zscore O-NET (Arts)	.55	-5.67	-.02	-.42
Zscore O-NET (Jobs)	.39	-5.90	.11	-.61
Zscore GAT	.86	1.39	.10	-1.05

Table 4.28 Distances between Final Cluster Centers by K-means algorithm (year 2011)

Cluster	1	2	3	4
1		12.01	2.30	4.99
2	12.01		11.58	11.40
3	2.30	11.58		2.76
4	4.99	11.40	2.76	

Table 4.29 ANOVA by K-means algorithm (academic year 2011)

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Zscore O-NET (Thai)	151.43	3	.59	1119	253.78	.00
Zscore O-NET (Social)	144.79	3	.61	1119	235.63	.00
Zscore O-NET (Eng)	147.45	3	.60	1119	242.78	.00
Zscore O-NET (Math)	141.28	3	.62	1119	226.45	.00
Zscore O-NET (Sci)	205.88	3	.45	1119	456.82	.00
Zscore O-NET (Phyisc)	59.60	3	.84	1119	70.70	.00
Zscore O-NET (Arts)	51.84	3	.86	1119	60.02	.00
Zscore O-NET (Jobs)	60.83	3	.84	1119	72.45	.00
Zscore GAT	161.65	3	.56	1119	283.95	.00

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

From experiment found that in the table of Initial Cluster Centers showed average of each standardized variable on all clusters. Clusters had four groups which initially defined content and process.

Table of Final Cluster Centers

This table showed standardized average. These averages were means of each cluster. Average on variable of O-NET (Thai) was differing when they were on different clusters and they had dramatically differed when compared to other variables. Average on variable O-NET (Thai) in 1st cluster = 0.80162. It was more than sum of average. It was 0.80162 times to standard deviation. While variable O-

NET (Thai) in 4th cluster = -1.04244. It was less than sum of average. It was 1.04244 times to standard deviation. In the same time, other variables were differing when they were on different clusters.

Table of distances between final cluster centers

In this table showed distance between means of four clusters. The 1st cluster had longest distance from the 2nd cluster = 12.014 and it had shortest distance to the 3rd cluster = 2.308 while 4th cluster had shortest distance to 3rd cluster.

Table of ANOVA

Average on variable of O-NET (Science), when it was on different group, had dramatically differed because it had the most statistically F = 456.823. Variable of GAT were the second rank. It had statistically F = 283.957. Variable O-NET (Arts) had lest differentiation when it was on different clusters.

Academic year 2012 (New Data Sets Using K-Means

Algorithm)

From Cluster Analysis on admission score of O-NET on all subjects and GAT on academic year 2012, then the results are below:

Table 4.30 Final Cluster Centers by K-means algorithm (academic year 2012)

	Cluster			
	1	2	3	4
Zscore O-NET (Thai)	1.57	-.92	.91	.04
Zscore O-NET (Social)	1.71	-.78	.87	-.01
Zscore O-NET (Eng)	-.27	-.81	.91	-.01
Zscore O-NET (Math)	.51	-.65	.85	-.07
Zscore O-NET (Sci)	1.74	-.83	1.08	-.09
Zscore O-NET (Phyisc)	-9.04	-.52	.39	.10
Zscore O-NET (Arts)	-5.23	-.52	.68	-.04
Zscore O-NET (Jobs)	-7.11	-.42	.28	.09
Zscore GAT	.11	-.97	.88	.08

Table 4.31 ANOVA by K-means algorithm (academic year 2012)

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Zscore O-NET (Thai)	217.13	3	.56	1474	387.65	.00
Zscore O-NET (Social)	175.35	3	.64	1474	271.79	.00
Zscore O-NET (Eng)	191.29	3	.61	1474	312.22	.00
Zscore O-NET (Math)	148.37	3	.70	1474	211.94	.00
Zscore O-NET (Sci)	239.99	3	.51	1474	467.28	.00
Zscore O-NET (Phyisc)	85.61	3	.82	1474	103.43	.00
Zscore O-NET (Arts)	103.80	3	.79	1474	131.26	.00
Zscore O-NET (Jobs)	53.13	3	.89	1474	59.43	.00
Zscore GAT	223.63	3	.54	1474	408.93	.00

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

From experiment found that from table of Initial Cluster Centers showed average of each standardized variables on all clusters. They had four clusters which initially defined since starting analysis.

Table of final cluster centers:

This table showed standardized average. These averages were means of each cluster. Average on variable of O-NET (Social) was differ when it was on different clusters and had dramatically differed when compared to other variables. Average on variable O-NET (Social) in 1st cluster = 1.71319. It was more than sum of average = 1.71319 times of standard deviation. While average on variable O-NET (Social) in 4th cluster = -0.1265. It was less than sum of average = 0.01265 times of standard deviation. At the same time, other variables had different average when they were on different clusters. Especially for variable O-NET (Physics) had differentiation of variable = -9.04766. It was less than sum of average = 9.04766 times of standard deviation.

Table of ANOVA

This table showed Mean Square between clusters (Between-cluster Mean Square) and Mean Square Error or Within-Cluster Mean Square. It had statistical F but it would not have statistical F or Significance in the last column of table. From experiment found average on variable of O-NET (Science) had dramatically differed when they were on different clusters because it had the most statistically $F = 467.281$. Variable of GAT was the second rank. It had statistically $F = 408.939$. Variable O-NET (Jobs) had lest different when they were on different clusters.

CHAPTER V

DISCUSSION

5.1 Data Collection

Procedure of data collection initially started from collection of admission scores from Mahidol University. These data consisted of scores of O-NET, GAT and PAT of students on academic year 2010, 2011 and 2012 from Division of Information Technology, Mahidol University, Salaya campus. All data was spreadsheet Microsoft Excel 2007. Lost of data on certain parts i.e. PAT and GAT on academic year 2010 was occurred. Cooperation support from officer was overwhelmingly with official letter from university.

5.2 Factor Analysis

5.2.1 Discussion on factor analysis

5.2.1.1 Preliminary

From the research has found that faculty or department of university which has used the same examination subject and weight value would be grouped into the same cluster. Example: group of departments in faculty of Engineer were under the same cluster, at the same time group of departments in faculty of Science were under the same cluster also.

Thus researcher has discussed the result that student who study in certain department of university may study in others departments under the same faculty. Example: student of Computer Engineering can study in department of Electric Engineering etc.

5.2.1.2 New data sets

Academic year 2010

From the result of research has suggested that it is better to use examination subjects in two factors as below:

Factor 1: Core subject:

This factor should select from GAT, O-NET (Eng), O-NET (Thai), O-NET (Social), O-NET (Science) or O-NET (Math) into only one subject for examination. It would be convenience for student to prepare for entrance examination.

Factor 2: Profession subject

This factor should select from subject of Health & Physical and subject of Art to only one subject for examination. It would be simpler for entrance examination.

Academic year 2011

The suggestion of this year should be the same as suggestion of year 2010 but subject of Art can be taken into factor 1 and 2.

Academic year 2012

The suggestion of this year should be the same as suggestion of year 2010 and 2011 but factor of profession subject can be taken into factor 1 and factor 2.

Thus the research has summarized that Factor Analysis can group subjects for entrance examination into only two subjects as 1st subject is group of core subjects & professional subjects and 2nd subject is group of profession, health & physical and art.

5.3 Cluster Analysis

5.3.1 Discussion of cluster analysis

Academic year 2010

From cluster analysis on entrance examination score of junior student of Mahidol University has found that discussion on result of research can be grouped into 2 topics as:

1. The important subjects to prepare for faculty of Science were Mathematics, Science and English. Therefore, the important subjects for faculty of Art were Thai language, Social, Profession, Health and Art.

2. Student who has got high score on Mathematics, Science and English should study on faculty of Science, Engineering or Medical. Student who has got high score on Thai language, Social, English language, health, art, profession or general learning subject should study on faculty of Art or Administration.

Academic year 2011 and 2012

The interesting result of research was described into 2 topics:

1. Student who intends to study in faculty of Art should better prepare on Mathematic also.

2. Student who has got the high score on Mathematics and Science can also study in faculty of Art or Administration as well.

5.4 Tools and limitations in the research

Researcher utilized program of SPSS version 16 for analysis on Hierarchical Clustering to actual exam results of students in Mahidol University. The analysis found that this program were not capable to compute because excessive data quantity and imperfect in contents of data.

CHAPTER VI

CONCLUSION AND RECOMMENDATION

6.1 Conclusion

This research has objective to analyze O-NET and GAT, using in university exam on system of Admission, by methods of factor analysis and cluster analysis. The experiment has defined to 2 sections as:

Section 1: Preliminary result

This section researcher operated factor analysis and clustering to weight values of exam subjects into Mahidol University. The experiment found from factor analysis that blocked factor were four factors as below:

- Factor 1: Group of core subjects
- Factor 2: Group of learning subjects
- Factor 3: Group of professional subjects & advanced English
- Factor 4: Group of advanced Mathematics

From cluster analysis found that clustering of faculties and department can be grouped by method of Agglomerative Hierarchical Cluster Analysis to 6 groups as:

- Group 1: Group of administration
- Group 2: Group of Engineering & environment science
- Group 3: Group of science
- Group 4: Group of health science & medicine
- Group 5: Group of sport science & medical technology
- Group 6: Group of fine art

Section 2: Final result

This section researcher still operated factor analysis and cluster analysis to scores of O-NET and GAT, using sample of students in Mahidol University on academic year of 2010, 2011 and 2012. The research had found two blocked factors. The method of cluster analysis was **K-mean clustering**. Number of clustering was

four groups. Research found that variable of exam subject in each clustering which similarity contents were arranged into the same cluster.

6.2 Recommendation for factor analysis

Additional analytical factors, such as score report from school or required characteristics of upper secondary school student etc., should be taken to analysis procedure for more completion of result.

6.3 Recommendation for cluster analysis

Diversified number of clusters on clustering with method of **K-means clustering** should be defined.

6.4 Future work

- Additional method on collecting data procedure, such as questionnaire from students about preference on required faculty etc., should be added for enhancing to Multivariate Analysis as Discriminate Analysis, Linear Regression Analysis etc.
- Additional technique of Data Mining on part of supervised learning, as classification, should be analyzed.

REFERENCES

1. <http://www.cuas.or.th/info.html>.
2. กัลยา วาณิชยัญชา.การวิเคราะห์สถิติขั้นสูงด้วย SPSS for Windows. พิมพ์ครั้งที่9. กรุงเทพฯ: พิมพ์ที่บริษัท ธรรมสาร จำกัด, มิถุนายน 2554.
3. Tombros, A., The Effectiveness of Query-based Hierarchical Clustering of Documents for Information Retrieval, Ph.D. thesis, Faculty of Computing Science, Mathematics and Statistics, University of Glasgow,2002.
4. Savio, L., Learned Text Categorization by Backpropagation Neural Network, The Hong Kong University of Science and Technology, 1996.
5. Jain, A. K., Murty, M.N. and Flynn, P.J., Data Clustering: A Review, ACM Computing Surveys, 31(3): 264-323, 1999.
6. Likhatchev, A., Cluster Analysis: Determining the Number of Clusters, <http://www.cs.mcgill.ca/tolik/CS-644/cluster.html>,1999.
7. faculty.uscupstate.edu/.../SingleLinkExample.doc.
8. http://en.wikipedia.org/wiki/Complete-linkage_clustering.
- 9.http://publib.boulder.ibm.com/infocenter/spssstat/v20r0m0/index.jsp?topic=%2Fcom.ibm.spss.statistics.help%2Falg_cluster_methods.htm.
10. Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Data Mining, Pearson Addison Wesley,2006.
11. He, Q., A Review of Clustering Algorithms as Applied in IR, Graduate School of Library and Information Science, University of Illinois at Urbana Champaign, 1999.
12. <http://www-01.ibm.com/software/analytics/spss/products/statistics/>.
13. สุพรรณยา อเนกบุญย์, วิชุดา ไชยศิริวามงคล, การจัดกลุ่มโรงเรียนมัธยมศึกษาในจังหวัดขอนแก่น , การประชุมทางวิชาการเสนอผลงานวิจัยระดับบัณฑิตศึกษาครั้งที่11.

14. กิจติพงษ์ ทะนงลักษณ์, สมชาย ปราการเจริญ และเกียรติศักดิ์ โยชนะนัง, ระบบวิเคราะห์และตรวจสอบชุดข้อมูลการโจมตีบนเครือข่าย WAN กรณีศึกษา ธนาคารเพื่อการเกษตรและสหกรณ์การเกษตร, The 6th National Conference on Computing and Information Technology NCCIT2010-221.
15. ณรงค์ศักดิ์ คงทิม และ จิรัฏฐา ญบุญชอบ, การประยุกต์ใช้เอฟพี-กโรธ กับงานแนะแนวการศึกษาต่อในระดับอุดมศึกษา, CIT2011&UNINOMS.
16. Hsu. et. al., “The Hybrid of Association Rule Algorithms and Genetic Algorithms for Tree Induction: an Example of Predicting the student course performance.” Expert Systems with Application.25, 1. (2003):51-62.
17. กฤษณะ ไวยมัย, ชิดชนก ส่งศิริ และชนาวินท์ รักธรรมานนท์, “การใช้เทคนิคดาต้าไมน์นิ่งเพื่อพัฒนาคุณภาพการศึกษาคณะวิศวกรรมศาสตร์ คณะวิศวกรรมศาสตร์”, NECTEC technical Journal vol. III, No.11 หน้า 134-142. 2001.
18. Wei Zong, “Application of SPSS for evaluation the curriculum designing and analysis the teaching effect in food engineering specialty”, School of Food and Biological Engineering, p.p.15.
19. Yang Yang, et. al., “Assessment of surface water quality using multivariate statistical techniques: A case study of the lakes in Wuhan, China”.
20. Jun Qu, et. al. , “A Hierarchical Clustering Based on Overlap Similarity Measure”, Software School, Department of Computer Science, Xiamen University, China, p.p.905.
21. Jinliang Zhang, “The Application of Hierarchical Clustering Analysis Technique in Reservoir for Flow Unit Study”, College of Marine Geo-science, Ocean University of China, p.p.43.
22. Joseph F. Hair JR., William C. Black, Barry J. Babin, Rolph E. Anderson, Multivariate Data Analysis, seventh edition.
23. Stephen Tuffery, Data Mining and Statistics for Decision Making, University of Rennes, France, Wiley series in computational statistics, ISBN 978-0-470-68829-8.
24. Anil K. Jain., Richard C. Dubes, Algorithms for Clustering Data, Michigan State University, Prentice Hall Englewood Cliffs, New Jersey 07632.
25. www.mathsworks.com.

The seal of Mahidol University is a large, faint, circular watermark in the background. It features a central emblem with a crown and a lotus flower, surrounded by Thai script and the words "MAHIDOL UNIVERSITY" in English.

APPENDICES

APPENDIX A

The Application of Using Hierarchical Clustering with Admission Score

Rajjakrij Wasubhaddaradilok, Waranyu Wongseree

**Department of Technology of Information System Management
Faculty of Engineering, Mahidol University, Nakorn Pathom
E-mails: rajjakrijoak@hotmail.com, E-mails: waranyu.won @ mahidol.ac.th**

Abstract

This paper presents the application of using factor analysis and hierarchical clustering to analyze factors of weight values for required subjects in university admission and for examinations to each faculty or 36 departments in Mahidol University. This research has utilized data of 16 examined subjects. The result of research found that factor analysis will decrease numbers of factor of weight values into only four factors as group of core subjects, group of learning subjects, group of professional subject & advanced English and group of advanced Mathematics. The clustering result has arranged faculties into 6 majors groups of Mahidol University as group of administration, group of engineering, group of science, group of health sciences & medicine, group of sport science & medical technology and group of fine art. The usefulness is trying to help secondary school students grade 6 when further education in university or the freshly students who had the problems of less academic proficiency while initially studying in university. At the same time, faculties could properly meet more required students as well.

Keywords: cluster analysis, factor analysis, university admission

1. Introduction

The university admission has two main objectives are 1) University can directly get the omniscient and skillful students to their faculties. 2) Learning procedures of secondary school student grade 6 can be improved to meet philosophy and objective of National Education

Act. The components of university admission are related to 1) Sum of Grade Point Average from the entire course of secondary school grade 6 (GPAX) is counted for 10%. 2) Ordinary National Educational Test (O-NET) is counted for 35-70%. 3) Grade Point Average on group of learning subjects is counted for 20%. 4) Advanced National Educational Test (A-NET) is counted for 0-35%. 5) Test of interview and physical examination is not counted. Under this procedure then faculties/departments of every university will specify on different examined subjects on core subjects and then on special subjects.

Anakboon and Chaisivamongkol [3] presented the clustering of secondary school in Khon Kaen and invented model for clustering by using the basic data and exterior standard evaluation of 100 secondary schools from educational service area of Khon Kaen province and the research found that clustering of secondary schools could divided into four groups as group of district, group of out of district, group of provincial and group of student's characteristics. The result of discriminate analysis found that accuracy of prediction was 91.9%.

Kongtim and Phuboonoop [4] presented the applying of association rule for guiding student to further education in university. Grades point average of the core 7 subjects as Mathematics, Chemistry, Biology, Physics, Thai Language, Social and English Language which analyzed by FP-Growth algorithm found that accuracy of means was 89.87%.

Tanonglak et al. [5] presented the analysis and inspection of attacking data on WAN system by using cluster analysis and discriminate analysis found that accuracy of grouping by system was 85.87%.

Hsu et al. [6] presented the techniques of suggestion for education in bachelor degree, by applying procedure of relation on association rule with genetic algorithms; found that efficiency of forecasting was better and less time of processing than simple genetic algorithms.

Waiyamai et al. [7] presented the searching on interesting topic of student's data for rule of relation, classification and forecast could support student grade 6 on determination of choosing appropriate faculty in university. At the same time, it could predict on exam result on next semester. The research used tree diagram to be the center of model for data classification.

From the university admission test found two problems were: 1) Students who passed the university exam could not meet the initially educational proficiency of faculties and then could not further education in university. Problems above came from The Central University Admission System differently defined the weight values of subjects for faculties. Then this research had analyzed the scores of university admission with the techniques of data clustering of all faculties in Mahidol University. It would help supporting for faculty to exam secondary school students grade 6 and further more it also help the freshly university students who could not further education in university. At the same time faculty could get more required students.

2. Related theories

2.1 Factor Analysis

Factor Analysis was the technique of decreasing number of factors or changing related factors to be the new unrelated factor. The resulted factors were linear combination of original factors [2]. The forecasting factor equation was

$$F_j = W_{j1}X_1 + W_{j2}X_2 + \dots + W_{jp}X_p + e$$

when X_j was factor j , W_j was coefficient of Factor j and relationship of the variable X_j which be the linear combination of all factors as:

$$Z_p = L_{p1}F_1 + L_{p2}F_2 + \dots + L_{pm}F_m + e_p$$

where Z_p was variable which be standardized, p were the number of variable, m were the number of factor, F_m was common factors, e_p was unique factor, L_{pm} was coefficient or factor loading.

2.2 Hierarchical Clustering

Hierarchical clustering analysis was usually used for clustering or factors clustering.

This technique must have identical function which utilized in a form of distance function. This research used distance function of Euclidean distance.

It started from the same numbers of clusters and samples. Then it determined the combination of selected couple of data by utilizing the least differentiations to be the same group. After that determine the method of combination and combine into one group [2]. The steps were as below:

Step 1: Combined two cases into one group or one cluster by determining on distance or resemblance.

Step 2: Determined about combining the 3rd case into two cases or took two new cases into new group by determining on distance or resemblance.

Step 3: Done the same as step 2. Each step might combine new case into original cluster or might combine two new cases into new cluster.

Step 4: Done the same until all cases are in the same cluster. Finally it would have only one group or one cluster.

This research would use the data combination of Single Linkage, Complete Linkage, Average Linkage Between Group and Centroid clustering. Each linkage had details as below:

Single Linkage method would combine two clusters into one cluster by determining on the least distance. Complete Linkage would combine two clusters into one cluster by determining on the maximum distance. Average Linkage method would calculate the average distance of every couple cases. One case would in cluster i and another case would in cluster j . If cluster i had the averaged distance from cluster j which shorter than other clusters, then it would combine Cluster i and j into same cluster. Centroid clustering method would combine two clusters into the same cluster by determining from distance of the center of cluster. It would calculate distance from center of two clusters. If any two clusters had the shortest distance would combine those two clusters into the same cluster.

3. Method of research

The process was started by forming data into weight values of required subjects in university admission of Mahidol University. Then 16 subjects were analyzed from 36 faculties.

The statistics in this research were as below:

1. All 16 variable values of the weight values were clustering by utilizing technique of factor analysis. It would be shown for factor score which ordering group of faculties, then screened

that factor and selected from the eigenvalue which was more than or equal to 1.

2. Faculties were clustering by technique of clustering analysis. It would use distance and calculated distance by squared Euclidean distance, then took that resulted distance into clustering by techniques of hierarchical clustering. This techniques had 4 methods were 1) Average Linkage 2) Single Linkage 3) Complete Linkage 4) Centroid Clustering. Dendrogram was used for presenting the results.

4. Results and Discussion

Means and standard deviation for all 16 variables were shown in Table 1. It found that the variable values on subjects of Science, Mathematics and English Language were having rather high scores of arithmetic means. It meant that the variable values on subjects of Science and English Language were affected to university admission of Mahidol University which was the science university.

The research analyzed 16 variables of the weighted values, which were tested in the admission to Mahidol University, and determined on eigenvalues which values were more than 1. It found that all 16 variables could screen 4 factors which were shown in Table 2.

Table 1 Means and Standard Deviation of 16 variables

	Mean	SD.
Thai Language (GPA)	3.78	0.68
Social Science (GPA)	3.67	1.04
Foreign Languages (GPA)	4.06	1.59
Mathematics (GPA)	3.94	1.10
Science (GPA)	4.00	1.49
Health & Physical (GPA)	0.56	2.32
Thai Language (O-NET)	7.50	0.88
Social Science (O-NET)	7.50	0.88
English Language (O-NET)	7.92	2.21
Mathematics (O-NET)	6.67	2.08
Science(O-NET)	6.67	2.08
Thai Language (A-NET)	1.25	6.02
English Language (A-NET)	7.92	6.59
Mathematics (A-NET)	10.42	3.85
Science (A-NET)	12.22	6.49
Basic of Engineer	1.94	4.01

Table 2 Total Variance Explained (Initial Eigen values)

Component	Total	% of Variance	Cumulative%
1	8.43	52.73	52.73
2	2.48	15.53	68.26
3	1.94	12.13	80.40
4	1.42	8.90	89.30

Table 3 Total Variance Explained (Rotation Sums of Squared Loadings)

Component	Total	% of Variance	Cumulative%
1	7.74	48.40	48.40
2	3.04	19.02	67.43
3	2.04	12.78	80.22
4	1.45	9.08	89.30

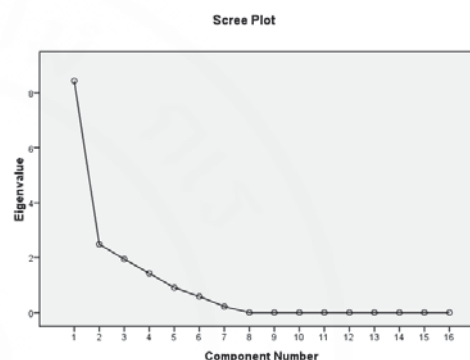


Figure 1 The relation between Eigen value of each factor and Component Number.

Table 3 was shown after rotated eigenvalue of 4 factors. When rotated the axle of factors into all types while all factors were still perpendicular. Varimax was the method while rotated the axle of factors. It found that the eigenvalue, percentage of variance and cumulative percentage of factor when rotated the axle were less than when stopped rotating while eigenvalue and percentage of variance of 2nd factor and 3rd factor were more than the value of 1st factor.

From Figure 1 found that component 1, 2, 3 and 4 had eigen value more than 1. Then factor should be 4 factors.

From the experiment which shown in Table 1, 2, 3, 4 and 5 are assured the result of Factor Analysis should be 4 factors.

From the experiment which shown in Table 4, after rotated factor loading then the 1st factor consisted of ten variables. 2nd factor consisted of three variables. 3rd factor consisted of two variables and 4th factor consisted of only one

From the experiment which shown in Table 5, factor score of 16 variables were assured result of factor analysis for all of four factors.

From the experiment as shown in Table 6 found that every variable were variables from group of core subjects. Three of variables in the 2nd factor as shown in Table 7 were involved in weight values of core subjects. Then the name of 2nd factor was group of learning subjects. Every variable of 3rd factor were the professional subjects & advanced English. Furthermore, it

shown that the professional subjects were the most influent for passing through university exam of every faculty. Therefore the name of 3rd variable value was group of professional subjects & advanced English. For the weight values of the advanced Mathematics (A-NET) was the only variable value that separated from others. It shown that advanced Mathematics (A-NET) was important. Furthermore it significantly impacted to university admission. Then the name of 4th variable value was advanced Mathematics.

Table 4 Rotated Component Matrixes

	Component			
	1	2	3	4
Thai Language (O-NET)	.95			
Social Science (O-NET)	.95			
English (O-NET)	.95			
Science (GPA)	-.91			
Math (GPA)	-.81			
Science (O-NET)	-.80	-.29	.39	
Math (O-NET)	-.80	-.29	.39	
Foreign Language (GPA)	.79	.48		
Science (A-NET)	-.72			-.63
Thai Language (A-NET)	.69			-.35
Social Science (GPA)	.22	.97		
Health & Physical (GPA)		-.95		
Thai Language (GPA)	.47	.83		
Basic of Engineer			.93	
English (A-NET)	.22		-.84	.31
Math (A-NET)	-.45			.78

Table 5 Component Score Coefficient Matrix

	Component			
	1	2	3	4
Thai Language GPA)	.00	.27	.02	.12
Social Science (GPA)	-.04	.34	.02	.03
Foreign Language (GPA)	.06	.11	-.04	-.09
Math (GPA)	-.11	.07	.04	.03
Science (GPA)	-.13	.04	.01	-.16
Health & Physical (GPA)	.11	-.37	-.01	.10
Thai Language (O-NET)	.15	-.07	.14	.02
Social Science (O-NET)	.15	-.07	.14	.02
English (O-NET)	.12	-.02	.01	-.03
Math (O-NET)	-.07	-.04	.15	.10
Science (O-NET)	-.07	-.04	.15	.10
Thai Language (A-NET)	.08	.00	.04	-.22
English (A-NET)	.00	-.01	-.41	.21
Math (A-NET)	-.04	.03	-.03	.52
Science (A-NET)	-.12	.02	-.04	-.46
Basic of Engineer	.06	.03	.48	.13

Table 6 Factor 1: Group of core subjects

Ordering	Subject Name
1	Foreign Language (GPA)
2	Math (GPA)
3	Science (GPA)
4	Thai Language (O-NET)
5	Social Science (O-NET)
6	English (O-NET)
7	Math (O-NET)
8	Science (O-NET)
9	Thai Language (A-NET)

10	Science (A-NET)
----	-----------------

Table 7 Factor 2: Group of learning subjects (GPA)

Ordering	Subject Name
1	Social Science (GPA)
2	Thai Language (GPA)
3	Health & Physical (GPA)

Table 8 Factor 3: Group of professional subjects

Ordering	Subject Name
1	Basic of Engineer
2	English (A-NET)

Table 9 Factor 4: Group of advanced Maths

Ordering	Subject Name
1	Math (A-NET)

After completed of factor analysis, then applying of cluster analysis was occurred. Four of combinations were 1) Average Linkage 2) Single Linkage 3) Complete Linkage 4) Centroid clustering.

Then researcher selected the method of combination. Selected method was Average Linkage Between Group. Average Linkage between group could clearly explain about relations and characteristics of faculty than other methods. Calculation on averaged distances of all faculties in all clusters was principal of this method. Clusters can be grouped to 6 clusters as group of Administration, group of Engineering, group of Science, Group of Health Sciences & Medicine, group of Sport Science & Medical Technology and group of Fine Art.

5. Conclusion

The research concluded that clustering of the weight value for the exam subjects into Mahidol University on 16 variable values could be grouped into 4 factors as below:

Factor 1: Group of Core Subjects

Factor 2: Group of Learning Subjects

Factor 3: Group of Professional & Advanced English

Factor 4: Group of Advanced Mathematic

Weight value of exam subjects were used from 16 variable values. Factor score was from combination of weight values on four factors. Then clustering of faculty and department, by using agglomerative hierarchical cluster analysis, can be grouped into 6 groups as:

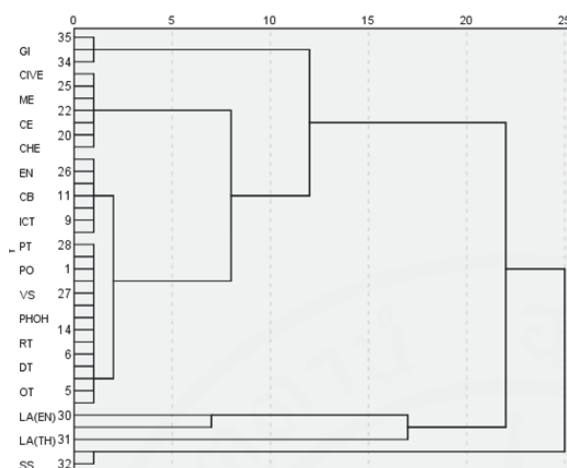


Figure 2 The diagram of dendrogram for the clustering of faculty and department in Mahidol University by method of Average Linkage between groups

Group 1: Group of Administration

Group 2: Group of Engineering

Group 3: Group of Science

Group 4: Group of Health Sciences & Medicine

Group 5: Group of Sport Science & Medical
Technology

Group 6: Group of Fine Art

The earliest score test of admission to Mahidol University should be applied for clustering of the faculty and department for more completion analysis. Furthermore, the others factor as gender should be determined as well.

6. References

- [1] Office of *Bunditnaenaew*, "Manual of Guiding for Selection of Department in Admissions", 1st edition, Rungrongsarn Printing, Bangkok, 2008.
- [2] Associate Professor Kalaya Vanichbancha, "Advanced Statistic Analysis with SPSS for Windows", 9th edition, Thammasarn Co., Ltd., Bangkok, June 2011.
- [3] Supansa Anakboon, Vishuda Chaisivamongkol, "The High School Clustering in Khon Kaen Province", The 11th Graduate Research Conference.
- [4] Narongsak Kongtim and Jiratta Phuboonoop, "Applied FP-Growth Algorithm and Guide to the Higher Education Study in university", CIT2011&UNINOMS.
- [5] Kijtipongs Tanonglak, Somchai Prakarncharoen and Kiattisak Yochanang, "the Analysis and Inspection on attacking data for WAN system", Case Study for Bank of Agriculture and Agricultural Cooperatives, The 6th National Conference on Computing And Information Technology NCCIT2010-221.
- [6] Hsu, P.L., Lai, R., Chiu, C.C. and Hsu, C.I., "The Hybrid of Association Rule Algorithms and Genetic Algorithms for Tree Induction: an Example of Predicting the student course performance", Expert Systems with Application, 25,1,2003, p.p.51-62.

[7] Krisana Waiyamai, Chidchanok Songsiri and Thanavin Raktammanon, "The Technique of Data Mining for Development of Educational Quality for Faculty Of Engineering", NECTEC Technical Journal vol.III, No.11, 2001, p.p. 134-142.

APPENDIX B

CHAPTER IV RESULT

4.1 Result of weight value of Mahidol University (Preliminary results)

4.1.1 Factor Analysis

Communalities from factor analysis

	Initial	Extraction
GPA (Thai)	1.00	.94
GPA (Social)	1.00	.99
GPA (Eng)	1.00	.92
GPA (Math)	1.00	.69
GPA (Sci)	1.00	.90
GPA (Physical)	1.00	.96
O-NET(Thai)	1.00	.95
O-NET(Social)	1.00	.95
O-NET(Eng)	1.00	.94
O-NET(Math)	1.00	.92
O-NET(Sci)	1.00	.92
A-NET(Thai)	1.00	.63
A-NET(Eng)	1.00	.85
A-NET(Math)	1.00	.82
A-NET(Sci)	1.00	.94
Engineering	1.00	.92

Component Matrix^a from factor analysis

	Component			
	1	2	3	4
O-NET(Eng)	.95			
GPA (Sci)	.92			
O-NET(Sci)	-.90		.26	
O-NET(Math)	-.90		.26	
O-NET (Arts)	-.90			-.25
O-NET(Thai)	.88	-.25	.32	
O-NET(Social)	.88	-.25	.32	
O-NET (Physsc)	-.78	.28		
A-NET(Thai)	.71			-.30
O-NET (Eng)	.69	.66		
A-NET(Sci)	-.69			-.68
O-NET (Jobs)		-.96		
GPA (Math)	.50	.86		
Engineering			.90	.25
A-NET(Eng)	.31		-.83	.25
A-NET(Math)	-.45			.73

Component Transformation Matrix from factor analysis

Component	1	2	3	4
1	.94	.30	-.12	-.05
2	-.29	.94	.09	.07
3	.14	-.04	.98	-.09
4	.08	-.06	.08	.99

Component Score Coefficient Matrix from factor analysis

	Component			
	1	2	3	4
O-NET (Eng)	.00	.27	.02	.12
O-NET (Math)	-.04	.34	.02	.03
O-NET (Sci)	.06	.11	-.04	-.09
O-NET (Physc)	-.11	.07	.04	.03
O-NET (Arts)	-.13	.04	.01	-.16
O-NET (Jobs)	.11	-.37	-.01	.10
O-NET(Thai)	.15	-.07	.14	.02
O-NET(Social)	.15	-.07	.14	.02
O-NET(Eng)	.12	-.02	.01	-.03
O-NET(Math)	-.07	-.04	.15	.10
O-NET(Sci)	-.07	-.04	.15	.10
A-NET(Thai)	.08	.00	.04	-.22
A-NET(Eng)	.00	-.01	-.41	.21
A-NET(Math)	-.04	.03	-.03	.52
A-NET(Sci)	-.12	.02	-.04	-.46
Engineering	.06	.03	.48	.13

Component Score Covariance Matrix from factor analysis

Component	1	2	3	4
1	1.00	.00	.00	.00
2	.00	1.00	.00	.00
3	.00	.00	1.00	.00
4	.00	.00	.00	1.00

4.1.2 Cluster analysis

Agglomeration Schedule for cluster analysis

Stage	Cluster Combined			Stage Cluster First Appears		
	Cluster 1	Cluster 2	Coefficients	Cluster 1	Cluster 2	Next Stage
1	35	36	.00	0	0	2
2	34	35	.00	0	1	31
3	3	32	.00	0	0	32
4	28	29	.00	0	0	5
5	1	28	.00	0	4	14
6	18	27	.00	0	0	14
7	13	26	.00	0	0	19
8	24	25	.00	0	0	9
9	19	24	.00	0	8	11
10	22	23	.00	0	0	11
11	19	22	.00	9	10	13
12	20	21	.00	0	0	13
13	19	20	.00	11	12	30
14	1	18	.00	5	6	16
15	16	17	.00	0	0	16
16	1	16	.00	14	15	18
17	14	15	.00	0	0	18
18	1	14	.00	16	17	25
19	8	13	.00	0	7	21
20	11	12	.00	0	0	21
21	8	11	.00	19	20	23
22	9	10	.00	0	0	23
23	8	9	.00	21	22	29
24	6	7	.00	0	0	25
25	1	6	.00	18	24	27
26	4	5	.00	0	0	27
27	1	4	.00	25	26	28
28	1	2	.00	27	0	29
29	1	8	7.27	28	23	30
30	1	19	14.41	29	13	31

31	1	34	18.44	30	2	32
32	1	3	49.28	31	3	35
33	30	31	62.02	0	0	34
34	30	33	64.37	33	0	35
35	1	30	126.58	32	34	0

4.2 Result of admission scores (new data sets)

4.2.1 Factor analysis

Academic year 2010

Component Matrix for factor analysis (academic year 2010)

	Component	
	1	2
O-NET (Sci)	.78	
O-NET (Thai)	.77	
O-NET (Math)	.73	
GAT	.71	-.41
O-NET(Social)	.71	
O-NET (Eng)	.67	-.36
O-NET (Arts)	.37	.28
O-NET (Jobs)	.37	.61
O-NET (Phyisc)	.43	.61

Component Transformation Matrix for factor analysis (academic year 2010)

Component	1	2
1	.87	.47
2	-.47	.87

Component Score Coefficient Matrix for factor analysis (yr 2010)

	Component	
	1	2
O-NET (Eng)	.23	.01
O-NET (Math)	.22	-.00
O-NET (Sci)	.30	-.18
O-NET (Physsc)	.13	.16
O-NET (Arts)	.15	.16
O-NET (Jobs)	-.14	.50
O-NET (Eng)	-.02	.26
O-NET (Math)	-.15	.50
GAT	.33	-.21

Academic year 2011

Communalities for factor analysis

	Initial	Extraction
O-NET (Eng)	1.00	.54
O-NET (Math)	1.00	.60
O-NET (Sci)	1.00	.58
O-NET (Physsc)	1.00	.49
O-NET (Arts)	1.00	.63
O-NET (Jobs)	1.00	.65
O-NET (Eng)	1.00	.23
O-NET (Math)	1.00	.62
GAT	1.00	.61

Component Matrix^a for factor analysis

	Component	
	1	2
O-NET (Sci)	.79	
GAT	.76	
O-NET (Thai)	.71	
O-NET (Eng)	.71	-.27
O-NET (Social)	.68	.37
O-NET (Math)	.64	-.29
O-NET (Arts)	.37	-.31
O-NET (Physsc)	.37	.72
O-NET (Jobs)	.42	.66

Component Transformation Matrix for factor analysis

Component	1	2
1	.88	.47
2	-.47	.88

Component Score Coefficient Matrix for factor analysis

	Component	
	1	2
O-NET (Eng)	.23	-.01
O-NET (Math)	.04	.32
O-NET (Sci)	.26	-.07
O-NET (Physsc)	.25	-.09
O-NET (Arts)	.21	.07
O-NET (Jobs)	-.14	.49
O-NET (Eng)	.19	-.14
O-NET (Math)	-.11	.46
GAT	.25	-.01

Component Score Covariance Matrix for factor analysis

Component	1	2
1	1.00	.00
2	.00	1.00

Academic year 2011

Communalities for factor analysis

	Initial	Extraction
O-NET (Eng)	1.00	.52
O-NET (Math)	1.00	.64
O-NET (Sci)	1.00	.70
O-NET (Phyisc)	1.00	.48
O-NET (Arts)	1.00	.58
O-NET (Jobs)	1.00	.55
O-NET (Eng)	1.00	.27
O-NET (Math)	1.00	.11
GAT	1.00	.67

Component Matrix^a for factor analysis

	Component	
	1	2
O-NET (Sci)	.75	
GAT	.75	-.32
O-NET (Eng)	.70	-.45
O-NET (Thai)	.70	
O-NET (Social)	.63	.48
O-NET (Math)	.63	-.30
O-NET (Arts)	.49	
O-NET (Jobs)	.29	
O-NET (Phyisc)	.35	.65

Component Transformation Matrix for factor analysis

Component	1	2
1	.78	.62
2	-.62	.78

Component Score Coefficient Matrix for factor analysis

	Component	
	1	2
O-NET (Eng)	.08	.23
O-NET (Math)	-.11	.44
O-NET (Sci)	.40	-.17
O-NET (Phyisc)	.30	-.08
O-NET (Arts)	.12	.19
O-NET (Jobs)	-.26	.50
O-NET (Eng)	.01	.21
O-NET (Math)	.15	-.06
GAT	.35	-.08

Component Score Covariance Matrix for factor analysis

Component	1	2
1	1.00	.00
2	.00	1.00

APPENDIX C

Abbreviation of faculties in Mahidol University

MCKA	= Faculty of management college accounting Karnchanaburi campus
MCN	= Faculty of management college Nakornsawan campus
MCK	= Faculty of management college Karnchanaburi campus
EC	= Faculty of Civil Engineering
EB	= Faculty of Biomedical Engineering
EM	= Faculty of Mechanical Engineering
EEC	= Faculty of Electrical Engineering
ECO	= Faculty of Computer Engineering
EI	= Faculty of Industrial Engineering
ECH	= Faculty of Chemical Engineering
SCKM	= Faculty of Radiological Science Karnchanaburi campus
ENSCT	= Faculty of Environmental Science and Resource Studies
SC	= Faculty of Science
SCKCB	= Faculty of Conservational Biology Karnchanaburi campus
SCKAS	= Faculty of Agricultural Science Karnchanaburi campus
SCICT	= Faculty of Information Communication and Technology
SCKFC	= Faculty of Food Technology Karnchanaburi campus
PT	= Faculty of Physical Therapy
PTA	= Faculty of Physical and Occupational Therapy
SIMK	= Faculty of Siriraj Medicine Prosthetics and Orthotics
NURS	= Faculty of Nursing
VET	= Faculty of Veterinarian Science
PUBH	= Faculty of Public Health
PUBHS	= Faculty of Public Health (Sanitary and Safety)
MDSC	= Faculty of Medical Technique

MDSCR	= Faculty of Radio Technique
PHARM	= Faculty of Pharmacy
DENT	= Faculty of Dentist
RAMN	= Faculty of Nursing Ramathibodi hospital
RAMAC	= Faculty of abnormal communication Ramathibodi Hospital
SIMATM	= Faculty of Siriraj Medicine Applied Thai Medicine
SIMMTE	= Faculty of Siriraj Medicine Medical Education
SPTSC	= Faculty of Sport Science
ARTE	= Faculty of Arts (English)
ARTT	= Faculty of Arts (Thai)
RELC	= Faculty of Religious College

BIOGRAPHY

NAME	Mr.Rajjakrij Wasubhaddaradilok
DATE OF BIRTH	30 January1985
PLACE OF BIRTH	Bangkok, Thailand
INSTITUTIONS ATTENDED	Mahidol University, 2003-2007 Bachelor Degree of Science (Mathematics) Mahidol University, 2008-2012 Master of Science (Technology of Information System Management)
RESEARCH AREAS	Clustering, Data mining
PUBLICATION / PRESENTATION	ICSEC2012 (International Computer Science and Engineering Conference)
HOME ADDRESS	164/808, T. Pimolraj, Bangbuathong District, Nonthaburi 11110, Thailand
TEL	089-095-9340
E-MAIL	rajjakrijoak@hotmail.com