การจำแนกกลุ่มตัวแปรด้วยเทคนิค

Cluster Analysis

สมโภชน์ ศรีสมุทร

1. ความนำ

Cluster Analysis เป็นเทคนิคที่ใช้จำแนกหรือแบ่ง Case (หมายถึง คน สัตว์ สิ่งของ หรือ องค์กร ฯลฯ) หรือแบ่งตัวแปรออกเป็นกลุ่มย่อย ๆ ตั้งแต่ 2 กลุ่มขึ้นไป

Case ที่อยู่ในกลุ่มเดียวกันจะมีลักษณะที่เหมือนกันหรือคล้ายกัน ส่วน Case ที่อยู่ต่างกลุ่มกันจะมีลักษณะที่แตกต่างกัน ดังนั้น การพิจารณาเลือกลักษณะหรือตัวแปรที่จะนำมาใช้ในการแบ่งกลุ่ม Case จึงมีความสำคัญ นอกจากนั้น Case ใด Case หนึ่งจะต้องอยู่ในกลุ่มหนึ่งเพียงกลุ่มเดียว

ถ้านำเทคนิค Cluster Analysis มาใช้ในการแบ่งกลุ่มตัวแปร จะให้ตัวแปรอยู่ในกลุ่มเดียวกันมีความสัมพันธ์กันมากกว่าตัวแปรที่อยู่ต่างกลุ่มกัน ตัวแปรที่อยู่ต่างกลุ่มกันมีความสัมพันธ์กันน้อยหรือไม่ มีความสัมพันธ์กันเลย ส่วนใหญ่การแบ่งกลุ่มตัวแปรจะใช้เทคนิค Factor ที่กล่าวไว้ในบทที่ 2 ส่วนการแบ่งกลุ่ม Case (คน สัตว์ สิ่งของ) จะใช้เทคนิค Cluster Analysis

2. วัตถุประสงค์ของ Cluster Analysis

ชื่อ นามสกุล( ) ได้กล่าวถึงวัตถุประสงค์ของ เทคนิควิธี Cluster Analysis ว่า เทคนิค Cluster Analysis มีวัตถุประสงค์ที่สำคัญอยู่ 2 ประการ คือ การจัดกลุ่มหน่วยวิเคราะห์ การจัดกลุ่มตัวแปร ซึ่งมีความสอดคล้องกับ ของ กัลยา วานิชย์บัญชา (2548) และสามารถกล่าวโดยรวมได้ดังนี้

เพื่อจัดกลุ่ม Case ซึ่งจะเป็นประโยชน์ในงานด้านต่าง ๆ เช่น การตลาด การแพทย์ การปกครอง ฯลฯ ดังตัวอย่างต่อไปนี้

ตัวอย่างที่ 1 ใช้ศึกษาพฤติกรรมการบริโภคของกลุ่มผู้บริโภคที่อยู่ต่างกลุ่มกัน ซึ่งจะทำให้สามารถวางกลยุทธ์ทางการตลาดได้อย่างมีประสิทธิภาพมากขึ้น การที่จะสามารถแยกกลุ่มผู้บริโภคออกเป็นกลุ่มย่อยได้ จะต้องพิจารณาถึงตัวแปรที่ใช้ในการแบ่งกลุ่มผู้บริโภค ที่จะทำให้ผู้ที่อยู่ต่างกลุ่มกันมีพฤติกรรมการบริโภคที่แตกต่างกัน ตัวแปรดังกล่าวอาจจะประกอบด้วยอาชีพ อายุ รายได้ เป็นต้น

ตัวอย่างที่ 2 ใช้วางแผนเพื่อการทดสอบตลาด เช่น อาจจะมีการแบ่งกลุ่มพื้นที่ หรือจังหวัดโดยรวมพื้นที่ หรือจังหวัดที่คล้ายกันไว้ด้วยกัน เพื่อจะได้กำหนดกลยุทธ์ทางการตลาดที่แตกต่างกันสำหรับพื้นที่ที่อยู่ที่ต่างกลุ่มกัน สำหรับตัวแปรที่ควรนำมาพิจารณาในการแบ่งกลุ่มอาจจะเป็นจำนวนประชากร รายได้เฉลี่ย อาชีพของคนในพื้นที่ พฤติกรรม ทัศนคติของคนในพื้นที่ เป็นต้น

ตัวอย่างที่ 3 การเปรียบเทียบรถยนต์ยี่ห้อต่าง ๆ โดยที่ 1 Case คือ รถยนต์ 1 ยี่ห้อ ซึ่งพิจารณาจากตัวแปร เช่น ความถี่ในการซ่อม ลูกสูบ ระบบเบรก ค่าใช้จ่ายต่อกิโลเมตร ราคา เป็นต้น

ตัวอย่างที่ 4 การแบ่งกลุ่มประเทศ อาจใช้ดัชนีทางด้านสาธารณสุข เป็นตัวแปรที่ใช้ในการแบ่งกลุ่ม เช่น จำนวนแพทย์ เภสัชกร พยาบาล จำนวนเตียงในโรงพยาบาล สัดส่วนของไขมัน และแป้งในอาหาร ในที่นี้ 1 Case คือ 1 ประเทศ โดยให้ประเทศที่มีระบบสาธารณสุขคล้ายกันอยู่ด้วยกัน ถ้าประเทศที่มีระบบสาธารณสุขต่างกันจะอยู่ต่างกลุ่มกัน

หมายเหตุ

1. จากตัวอย่างที่ 1 และ 2 ข้างต้น จะพบว่าการเลือกตัวแปรเพื่อนำมาใช้แบ่งกลุ่ม Case มีความสำคัญมาก เพราะถ้าผู้วิจัยเลือกตัวแปรที่ไม่ได้ทำ Case แตกต่างกันแล้ว จะทำให้ไม่สามารถแบ่งกลุ่มได้ถูกต้อง การเลือกจะต้องพิจารณาว่าตัวแปรใดบ้างที่มีอิทธิพลทำให้เกิดความแตกต่าง ในตัวอย่างที่ 2 การแบ่งกลุ่มจังหวัด ถ้าไม่ได้นำตัวแปร จำนวนประชากร รายได้ อาชีพ เข้ามาพิจารณาแบ่งกลุ่มก็อาจไม่สามารถสร้างเกณฑ์ในการแบ่งกลุ่มได้ถูกต้อง

2. เมื่อแบ่ง Case เป็นกลุ่มย่อยแล้ว จะสามารถศึกษาถึง Profile หรือลักษณะของกลุ่มย่อยแต่ละกลุ่มได้ เพื่อนำมาใช้วางแผนด้านการตลาดต่อไป (กรณีที่เป็นเรื่องการศึกษาพฤติกรรมผู้บริโภค)

3. เมื่อใช้จัดกลุ่มตัวแปร การจัดกลุ่มตัวแปรที่มีความสัมพันธ์กันไว้ด้วยกัน จะเป็นการลดจำนวนข้อมูลที่มีจำนวนมากให้น้อยลง ทำให้ง่ายต่อการวิเคราะห์ เช่น เดิมมี 100 Case 20 ตัวแปร รวมข้อมูลทั้งหมด 2,000 ค่า (100×20) แต่ถ้าจัดกลุ่มตัวแปร 20 ตัว เหลือเพียง 3 กลุ่ม จะทำให้ข้อมูลลดลงเหลือเพียง 300 ค่า (3×100)

นอกจากนั้น การจัดกลุ่มตัวแปรทำให้ทราบว่าตัวแปรใดบ้างที่มีความสัมพันธ์กัน การเปลี่ยนแปลงของตัวแปรบางตัวย่อมมีผลกระทบต่อตัวแปรอื่น ๆ ที่มีความสัมพันธ์กับตัวแปรดังกล่าว

หมายเหตุ : ส่วนใหญ่จะใช้เทคนิค Cluster Analysis ในการจัดกลุ่ม Case มากกว่าการจัดกลุ่มตัวแปร การจัดกลุ่มตัวแปรจะใช้เทคนิค Factor Analysis ในที่นี้จึงจะแสดงตัวอย่างเฉพาะการจัดกลุ่ม Case

3. คุณสมบัติของเทคนิควิธี Cluster Analysis

1. ความต้องการทางด้านข้อมูล สำหรับการวิเคราะห์จัดกลุ่มหน่วยวิเคราะห์ (cast) ผู้วิจัยอาจใช้ข้อมูลที่ระบุหน่วยวิเคราะห์และตัวแปรตามที่จัดเก็บมาได้เลย เช่นการวิเคราะห์ที่ได้กล่าวมาแล้วข้องต้น ส่วนการวิเคราะห์จัดกลุ่มตัวแปร ผู้วิจัยไม่อาจจะใช้แฟ้มข้อมูลดังกล่าวได้ โดยใช้เมตริกแสดงความสัมพันธ์ระหว่างตัวแปร แทนได้

2. แนวคิดพื้นฐาน สิ่งสำคัญที่สุดของการวิเคราะห์การจัดกลุ่มคือ ตัวแปรที่ใช้ หากผู้วิจัยไม่ได้เก็บข้อมูลเกี่ยวกับตัวแปรที่สำคัญๆ ผลที่ได้ก็จะไม่ดีหรือทำให้ไขว้เขวได้ ทั่งนี้เพราะตัวแปรที่เลือกไว้ตั้งแต่แรกจะเป็นสิ่งที่กำหนดคุณสมบัติของสิ่งที่ระบุความเป็นกลุ่มย่อย เช่น ในการจัดกลุ่มโรงเรียนในเมือง หากผู้วิจัยไม่เก็บข้อมูลเกี่ยวกับจำนวนนักเรียนและครู ขนาดของโรงเรียนก็ไม่อาจเป็นเกณฑ์ในการแบ่งกลุ่มได้

3. ความคล้ายกันของหน่วย ความคิดเกี่ยวกับความคล้ายของหน่วยศึกษา เป็นเทคนิคของการวิเคราะห์ทางสถิติหลายวิธี โดยทั่วไปการวัดความคล้ายจะพิจารณาจากความห่าง ระหว่าวัตถุหรือพิจารณาจากความคล้ายกัน ซึ่งจะกล่าวโดยละเอียดในหัวข้อต่อไป

4. การวัดความห่าง วิธีการวัดความห่างสามารถวัดได้หลายวิธี วิธีการหนึ่งที่นิยมวัดกันมากก็คือ วิธีที่เรียกว่า ระยะห่างเชิงยูคลิดยกกำลังสอง (Squared Euclidean distance) คือผลรวมของผลต่างยกกำลังสองของทุกตัวแปร เช่นต้องการดูความห่างกันของเบียร์ 2 ยี่ห้อ ซึ่งเราทราบราคาต้นทุน และแคลอรี่ของเบียร์ทั้ง 2

ตารางที่ 3.1 ค่าของแคลอรีและต้นทุน

	แคลอรี่	ต้นทุน
บัดไวเซอ	114	43
โลเวนบราว	157	48

ความแตกต่างระหว่างเบียร์ทั้ง 2 คือ ($(114-157)^2+(43-48)^2$ เท่ากับ 13^2+5^2 หรือ 194

อย่างไรก็ดี ความแตกต่างระหว่างหน่วยของการวัดในแต่ละตัวแปรก็จะเป็นปัญหาในการวัดค่าความห่าง ดังนั้น จึงจำเป็นที่จะต้อง ทำให้ตัวแปรทุกตัวอยู่ในมาตรวัดเดี่ยวกัน คือการทำให้ตัวแปรทุกตัวมีค่าเฉลี่ยเป็น 0 และส่วนเบี่ยงแบนมาตาฐานเป็น 1 ซึ่งผลที่ได้คือ ค่าคะแนนมาตรฐาน ซึ้งจะได้เป็นค่า ดังตารางที่ 2

ตารางที่ 3.2 คะแนนมาตราฐานของค่าของแคลอรีและต้นทุน

	แคลอรี่	ต้นทุน
บัดไวเซอ	0.38	-0.46
โลเวนบราว	0.81	-0.11

ไม่ว่าจะทำการคำนวณหาความห่างหรือความคล้ายด้วนวิธีใดก็ตาม ผู้วิจัยจะต้องตัดสินใจว่าจะปรับสเกลตัวใดบ้าง เพื่อทำให้ตัวแปรมีสเกลเหมือนกัน มีฉะนั้นแล้วค่าความห่างหรือความต่างจะขึ้นอยู่กับขนาดของมาตรวัดของตัวแปรที่มีขนาดใหญ่กว่า ซึ่งการปรับทำได้หลายวิธีเช่น การหารด้วยค่าเบี่ยงเบนมาตรฐาน ค่าพิสัย ค่าเฉลี่ย

เมื่อทำการปรับค่ามาตรฐานแล้ว จึงคำนวณหาค่าความต่างหรือความคล้ายกัน ชนิดต่างๆ ซึ่งวิธีต่างๆนั้นจะให้น้ำหนักของข้อมูลที่ต่างกัน ซึ่งจะกล่าวถึงรายละเอียดของสูตรที่ใช้ในการวิเคราะห์แต่ละวิธีต่อไป

4. ประเภทของเทคนิค Cluster Analysis

เทคนิค Cluster Analysis แบ่งเป็นหลายประเภทหรือเทคนิคย่อย โดยเทคนิคที่ใช้กันมากมี 2 เทคนิค คือ

4.1 Hierarchical Cluster Analysis

4.2 K-Means Cluster Analysis

นอกจากนี้ ยังมีเทคนิค 2 Step Cluster Analysis และเทคนิคดังกล่าวมีวัตถุประสงค์และวิธีการที่แตกต่างกัน ซึ่งจะได้กล่าวถึงเทคนิค Hierarchical Cluster Analysis โดยละเอียดในหัวข้อ 5 และเทคนิค K-Means Cluster Analysis ในหัวข้อ 6 ตามลำดับ

5. เทคนิค Hierarchical Cluster Analysis

เป็นเทคนิคที่นิยมใช้กันมากในการแบ่งกลุ่ม Case หรือแบ่งกลุ่มตัวแปร โดยมีเงื่อนไขดังนี้

1. ในกรณีที่ใช้ในการแบ่ง Case นั้น จำนวน Case ต้องไม่มากนัก (จำนวน Case ควรต่ำกว่า 200 ถ้าตั้งแต่ 200 ขึ้นไปใช้ K-Means Cluster) และจำนวนตัวแปรต้องไม่มากเช่นกัน

2. ไม่จำเป็นต้องทราบจำนวนกลุ่มมาก่อน

3. ไม่จำเป็นเป็นต้องทราบว่าตัวแปรใดหรือ Case ใดอยู่กลุ่มใดก่อน

หมายเหตุ : เงื่อนไขในข้อ 2 และข้อ 3 จะตรงข้ามกับเงื่อนไขของเทคนิค Discriminant ในบทที่ 3 ซึ่งจำเป็นต้องทราบจำนวนกลุ่มมาก่อนและต้องทราบ Case ใดอยู่กลุ่มไหนมาก่อน

5.1 ขั้นตอนของเทคนิค Hierarchical Cluster สำหรับการแบ่งกลุ่ม Case

ขั้นที่ 1 เลือกตัวแปรหรือปัจจัยที่คาดว่ามีอิทธิพลที่ทำให้ Case ต่างกัน นั่นคือ ตัวแปรนั้นจะทำให้สามารถแบ่งกลุ่ม Case ได้ชัดเจน ขั้นตอนนี้เป็นขั้นตอนที่สำคัญดังได้กล่าวแล้วในหัวข้อ 1

ขั้นที่ 2 เลือกวิธีการวัดระยะห่างระหว่าง Case แต่ละคู่ หรือเลือกวิธีการคำนวณเพื่อวัดค่าความคล้ายของ Case แต่ละคู่ ซึ่งจะกล่าวถึงแต่ละวิธีในหัวข้อ 5.2

ขั้นที่ 3 เลือกหลักเกณฑ์ในการรวมกลุ่ม หรือรวม Cluster ซึ่งจะกล่าวถึงแต่ละหลักเกณฑ์ในหัวข้อ 5.3

5.2 การวัดความคล้าย (Similarity Measure)

ดังที่ได้กล่าวมาแล้วถึงหลักเกณฑ์ของเทคนิค Cluster ว่าจะใช้ในการจัด Case ที่คล้ายกันไว้ในกลุ่มเดียวกัน หรือจัดกลุ่มตัวแปรที่สัมพันธ์กันไว้ในกลุ่มเดียวกัน นั่นคือ จะมีการวัดความคล้ายกันของ Case ทีละคู่ ในกรณีที่เป็นการจัดกลุ่ม Case ส่วนการจัดกลุ่มตัวแปร การวัดความคล้ายจะเป็นการวัดความคล้ายของตัวแปรแต่ละคู่ คือ การหาค่าสัมประสิทธิ์สหสัมพันธ์ เมื่อต้องการจัดกลุ่ม Case จะต้องหาความคล้ายของ Case ถึง ${}^{n}C_{2}$ คู่ เมื่อมีข้อมูล Case = n แต่ถ้าต้องการจัดกลุ่มตัวแปรจะต้องหาความสัมพันธ์ของตัวแปรทีละคู่รวมถึง ${}^{k}C_{2}$ คู่ เมื่อมีตัวแปร k ตัว

การวัดความคล้ายของ Case แต่ละคู่อาจจะวัดด้วยระยะห่าง (Distance) หรือวัดด้วยค่าความคล้าย (Similarity) แต่การวัดความสัมพันธ์ของตัวแปรจะวัดด้วยค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน (Pearson correlation)

สำหรับวิธีการคำนวณระยะห่าง หรือค่าความคล้ายของ Case แต่ละคู่ จะแตกต่างกันเมื่อชนิดของข้อมูลต่างกัน ซึ่งชนิดของข้อมูลหรือตัวแปรที่สามารถใช้เทคนิค Hierarchical Cluster ได้ มี 3 ประเภท คือ

1. ข้อมูลเป็นสเกลอันตรภาค (Interval scale) หรือสเกลอัตราส่วน (Ratio scale)
2. ข้อมูลที่อยู่ในรูปความถี่ (Count Data)
3. ข้อมูลอยู่ในรูป Binary นั่นคือ มีได้ 2 ค่า คือ 0 กับ 1

หรือกล่าวได้ว่า ข้อมูลที่นำมาใช้ในเทคนิค Hierarchical จะเป็นข้อมูลชนิดตัวเลข หรือเป็นเชิงปริมาณ (Interval หรือ Ratio scale) หรือข้อมูลอยู่ในรูปความถี่ หรือ Binary

- **กรณีที่วัดความคล้ายด้วยระยะห่าง**

ถ้าระยะห่างระหว่าง Case คู่ใดต่ำ แสดงว่า Case คู่นั้นอยู่ใกล้กัน หรือมีความคล้ายกัน ควรจะจัดให้อยู่ในกลุ่ม หรือ Cluster เดียวกัน สำหรับวิธีการคำนวณจะขึ้นอยู่กับชนิดของข้อมูลทั้ง 3 ชนิดข้างต้น ซึ่งจะได้กล่าวถึงในหัวข้อ 5.4

- **กรณีที่วัดความคล้ายด้วยของ Case**

ถ้าค่าความคล้ายของ Case คู่ใดมีค่ามาก แสดงว่า Case คู่นั้นคล้ายกันมาก จึงควรจัดให้อยู่ในกลุ่มเดียวกัน การคำนวณค่าความคล้ายจะแตกต่างกัน ถ้าชนิดของข้อมูลแตกต่างกัน ซึ่งจะกล่าวถึงวิธีการคำนวณค่าความคล้ายตามชนิดของข้อมูลในหัวข้อ 5.4

- **กรณีที่วัดความคล้ายของตัวแปรด้วยค่าสัมประสิทธิ์สหสัมพันธ์**

ถ้าตัวแปรคู่ใด มีค่าสัมประสิทธิ์สหสัมพันธ์มาก แสดงว่าคู่นั้นสัมพันธ์กันมาก ควรจัดไว้ในกลุ่มเดียวกัน

5.3 หลักการการรวมกลุ่ม (Methods for Combining Cluster)

สำหรับหลักการในการรวมกลุ่มของเทคนิค Hierarchical Cluster นั้นมีหลายวิธี วิธีที่นิยมกันมาก คือ Agglomerative Hieratchical Cluster Analysis หรือในโปรแกรม SPSS เรียกว่า Agglomerative Schedule

Agglomerative Schedule

หลักการเกณฑ์ของ Agglomerative schedule จะทำการรวมกลุ่ม Cluster อย่างเป็นขั้นตอนดังนี้

ก่อนทำการวิเคราะห์จะกำหนดให้ 1 กลุ่ม หรือ 1 Cluster มี Case 1 Case นั่นคือ ถือว่าแต่ละ Case เป็น 1 Cluster จึงมีจำนวน Cluster เท่ากับจำนวนข้อมูลหรือจำนวน Case กรณีที่มีจำนวนข้อมูล n Case จะมี n Cluster หรือ n กลุ่ม

ขั้นที่ 1 : รวม Case 2 Case ให้อยู่ในกลุ่มเดียวกัน หรือ Cluster เดียวกัน โดยพิจารณาจากค่าระยะห่างหรือค่าความคล้าย

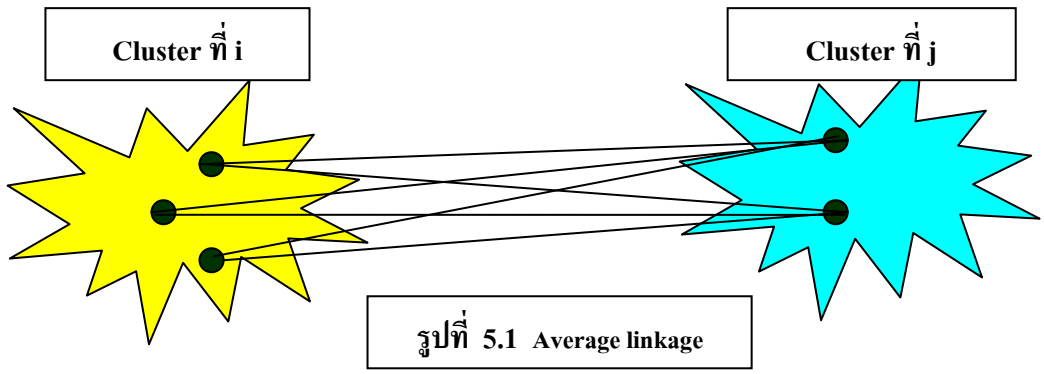
ขั้นที่ 2 : พิจารณาว่าควรจะรวม Case ที่ 3 เข้าอยู่ในกลุ่มเดียวกับ 2 Case แรก หรือควรจะรวม 2 Case ใหม่เข้าอยู่ในกลุ่มใหม่อีกกลุ่มหนึ่ง โดยพิจารณาจากค่าระยะห่าง หรือค่าความคล้าย

ทำขั้นที่ 3, 4 , ... โดยใช้เกณฑ์เดียวกับขั้นที่ 2 นั่นคือ ในแต่ละขั้นอาจจะรวม Case ใหม่เข้าไปในกลุ่มที่มีอยู่แล้ว หรือรวม Case ใหม่ 2 Case เป็นกลุ่มใหม่ ทำเช่นนี้ไปเรื่อย ๆ จนกระทั่งได้ ทุก Case อยู่ในกลุ่มเดียวกัน นั่นคือ สุดท้ายมีเพียง 1 กลุ่มหรือ 1 Cluster และ Case ใดที่ถูกจัดกลุ่มแล้วจะไม่มีการเปลี่ยนแปลง

หลักเกณฑ์ในการรวมกลุ่ม

หลักเกณฑ์ในการรวมกลุ่มในแต่ละขั้นตอนข้างต้นมีหลายวิธี ในที่นี้จะกล่าวถึงเฉพาะวิธีที่มีในโปรแกรม SPSS ซึ่งจะปรากฏในคำสั่ง Method ดังนี้

1. Between – groups Linkage หรือเรียกว่าวิธี Average Linkage Between Groups หรือเรียกกว่า UPGMA (Unweightede Pair-Group Method Using Arithmetic Average)



รูปที่ 5.1 Average linkage

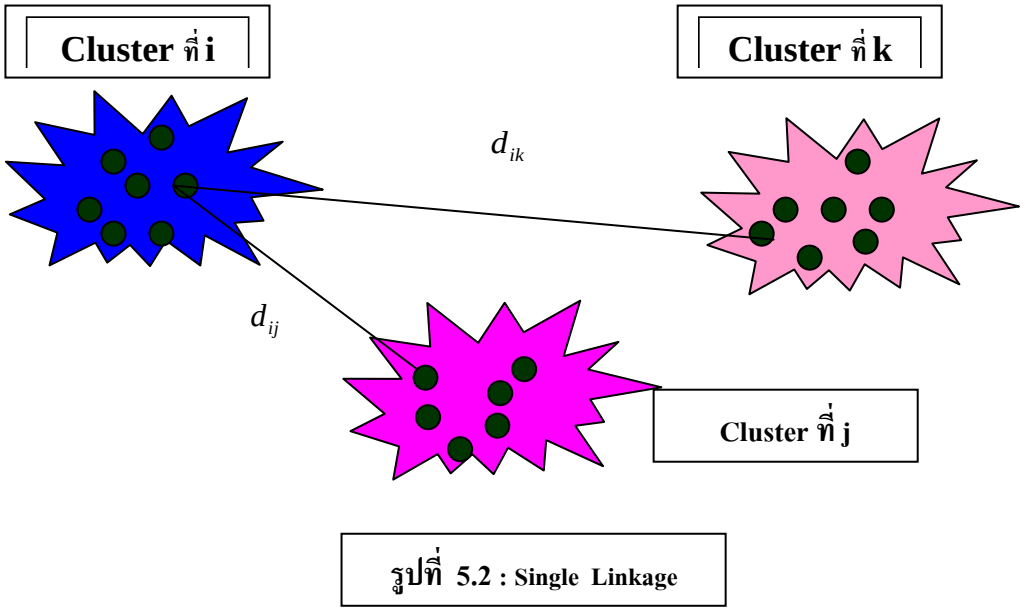
วิธีนี้จะคำนวณหาระยะห่างเฉลี่ยของทุกคู่ของ Case โดยที่ Case หนึ่งอยู่ใน Cluster ที่ i ส่วนอีก Case หนึ่งอยู่ใน Cluster ที่ j, $i \neq j$

ถ้า Cluster ที่ i มีระยะห่างเฉลี่ยจาก Cluster ที่ j สั้นกว่าระยะห่างจาก Cluster อื่นจะนำ Cluster ที่ i และ j รวมกันเป็น Cluster เดียวกัน

2. Within-group Linkage Technique หรือเรียกว่า Average Linkage Within Groups Method

วิธีนี้จะรวม Cluster เข้าด้วยกันถ้าระยะห่างเฉลี่ยระหว่างทุก Case ใน Cluster นั้น ๆ มีค่าน้อยที่สุด

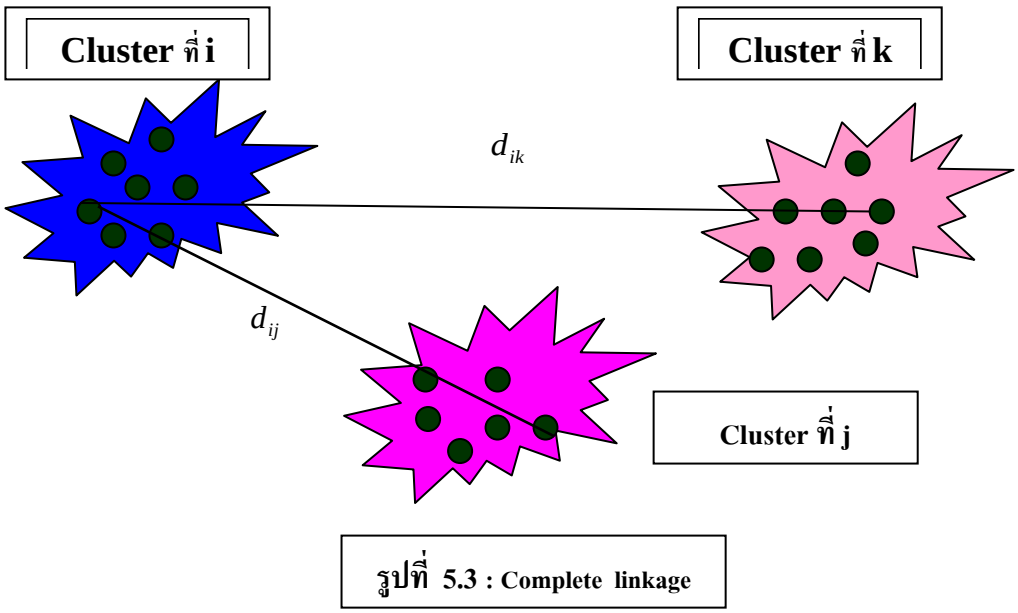
3. Nearest Neighbor หรือเรียกว่า Single Linkage



รูปที่ 5.2 : Single Linkage

วิธีนี้จะรวม Cluster 2 Cluster เข้าด้วยกันโดยพิจารณาจากระยะห่างที่สั้นที่สุด โดยที่ d_{ik} เป็นระยะห่างที่สั้นที่สุดระหว่าง Cluster i และ j ในรูปที่ 5.2 จะรวม Cluster i และ j เข้าด้วยกันเพราะ $d_{ij} < d_{ik}$

4. Furthest Neighbor Technique หรือเรียกว่า Complete Linkage



รูปที่ 5.3 : Complete linkage

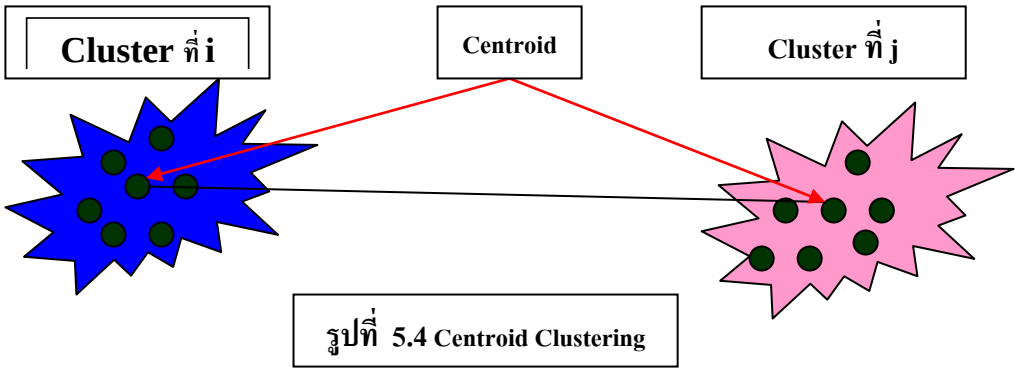
วิธีนี้จะรวม Cluster 2 Cluster เข้าด้วยกันโดยพิจารณาจากระยะห่างที่ยาวที่สุด

d_{ik} = ระยะห่างที่ยาวที่สุดของ Cluster ที่ i และ k

d_{ij} = ระยะห่างที่ยาวที่สุดของ Cluster ที่ i และ j

ในที่นี้ $d_{ij} < d_{ik}$ จึงรวม Cluster ที่ i และ j เข้าเป็น Cluster เดียวกัน

5. Centroid Clustering



รูปที่ 5.4 Centroid Clustering

วิธีนี้จะรวม Cluster 2 Cluster เข้าด้วยกันโดยพิจารณาจากระยะห่างของจุดกลางของ Cluster (กลุ่ม)

วิธีการนี้จะคำนวณหาระยะห่างระหว่าง Centroid ของ Cluster ทีละคู่ ในที่นี้จะเรียกค่าเฉลี่ย หรือค่ากลางของแต่ละ Cluster ว่า Centroid ของ Cluster เนื่องจากการแบ่งกลุ่ม Case จะพิจารณาจากตัวแปรหลาย ๆ ตัวพร้อม ๆ กัน จึงเรียกค่ากลางหรือค่าเฉลี่ยว่า Centroid

ถ้าระยะห่างระหว่าง Centroid ของ Cluster คู่ใดต่ำจะรวม Cluster คู่นั้นเข้าเป็น Cluster เดียวกัน

6. Median Clustering

วิธีนี้จะรวม Cluster 2 Cluster เข้าด้วยกัน โดยให้แต่ละ Cluster สำคัญเท่ากัน (ให้น้ำหนักเท่ากัน) ในขณะที่วิธีของ Centroid Clustering จะให้ความสำคัญแก่ Cluster มีขนาดใหญ่มากกว่า Cluster ที่มีขนาดเล็ก (ให้น้ำหนักไม่เท่ากัน)

Median Clustering จะใช้ค่า Median เป็นค่ากลางของ Centroid ถ้าระยะห่างระหว่างค่า Median ของ Clustering จะใช้ค่า Median เป็นค่ากลางของ Centroid ถ้าระยะห่างระหว่างค่า Median ของ Cluster คู่ใดต่ำจะรวม Cluster คู่นั้นเข้าด้วยกัน

7. Ward's Method

หลักการของวิธีนี้จะพิจารณาจากค่า Sum of the squared within-cluster distance โดยจะรวม Cluster ที่ทำให้ค่า Sum of square within-cluster distance เพิ่มขึ้นน้อยที่สุด โดยค่า Square within-cluster distance คือค่า Square Euclidean distance ของแต่ละ Case กับ Cluster Mean

5.4 ขั้นตอนการใช้ SPSS ในการจัดกลุ่ม Cases ด้วยเทคนิค Hierarchical Cluster

ขั้นที่ 1 : สร้างแฟ้มข้อมูล ซึ่งอาจจะสร้างโดย

ก) ใช้ข้อมูลจริงที่มี ซึ่งจะมีตัวแปรหลาย ๆ ตัวที่จะนำมาใช้ในการแบ่ง Case หรือแบ่งกลุ่มตัวแปรโดยให้คำนวณหาค่าระยะห่าง หรือค่าความคล้ายของ Case แต่ละคู่ ถ้าหน่วยของตัวแปรต่างกัน อาจจะมีผลต่อค่าระยะห่าง และค่าความคล้าย ซึ่งทำให้เกิดผลต่อการจัดกลุ่มด้วย ตัวแปรที่มีค่ามากจะมีอิทธิพล ต่อค่าระยะห่างมากกว่าตัวแปรที่มีค่าน้อย (เนื่องจากหน่วยต่างกัน) เช่น ถ้าวัดความคล้ายของนางกัลยา และนายชาตรีโดยตัวแปรที่วัดคือ อายุ (ปี) และรายได้ (หน่วย : 10,000 บาท)

ตารางที่ 5.1 : ข้อมูลดิบ

	อายุ (ปี)	รายได้ (10,000 บาท)
กัลยา	45	2
ชาตรี	60	7

ตารางที่ 5.2 ข้อมูลที่ Standardized แล้ว

	อายุ (ปี)	รายได้
กัลยา	.707	-.707
ชาตรี	-.707	.707

ถ้าในที่นี้ใช้ Euclidean Distance ในการหาระยะห่างระหว่างนางกัลป์ยา และนายชาตรี โดยใช้ข้อมูลในตารางที่ 5.1 ได้ระยะห่างของอายุและรายได้ = $(45 - 60)^2 + (2 - 7)^2 = 225 + 25 = 250$ นั่นคือ ระยะห่าง 250 นั้นเป็นอิทธิพลของตัวแปรอายุ = (255 / 250) x 100 = 90% อีก 10% เป็นอิทธิพลของตัวแปรรายได้

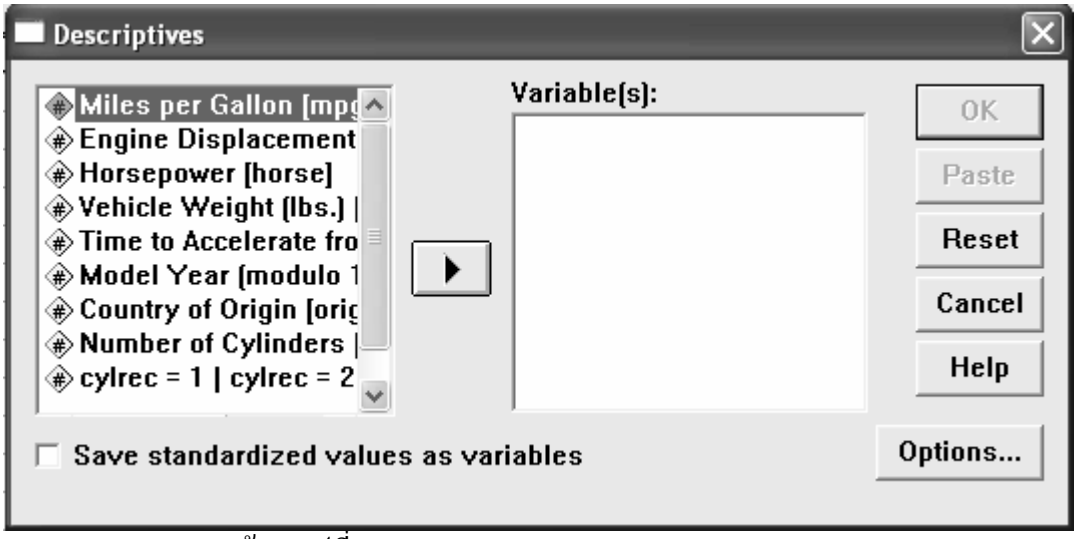
แต่ถ้าใช้ข้อมูลที่ทำ Standardized แล้ว ในที่นี้คือ การทำ Z-score จากตารางที่ 5.2 ได้ค่าระยะห่างของ Euclidean distance ในรูป Z-score เป็น $(-.707 - (-.707)^2 + (-.707 - .707)^2 = .999$ ซึ่งเป็นผลจากอายุ และรายได้เท่า ๆ กัน คือ อย่างละ 50% จึงควรทำการเปลี่ยนแปลงข้อมูลดิบของตัวแปรต่าง ๆ เพื่อกำจัดอิทธิพลของหน่วยที่ต่างกันออกไป

ข) ใช้ข้อมูลที่เปลี่ยนแปลงแล้ว เช่น ข้อมูลที่ Standardized แล้ว หรือเปลี่ยนแปลงข้อมูลของทุกตัวแปรให้มีค่าต่ำสุดเป็น 0 และค่าสูงสุดเป็น 1 ในคาสั่งย่อยของ Hierarchical Cluster จะมีการให้เลือกวิธีการ Standardized หลายวิธี ซึ่งจะกล่าวถึงในตัวอย่างที่ 5.1

ในกรณีที่ไม่ต้องการใช้คำสั่งย่อยของคำสั่ง Hierarchical Cluster เพื่อคำนวณค่า Z-score ของตัวแปรทุกตัวที่ต้องการนำมาใช้ในการจัดกลุ่ม แต่ต้องการทำ Standardized ข้อมูลเอง หลังจากที่มีการสร้างแฟ้มข้อมูลแล้ว ให้ใช้คำสั่ง ดังนี้

Analyze → Descriptive statistics → Descriptive … จะได้หน้าขอ ดังแสดงในรูปที่ 5.5

รูปที่ 5.5 : Descriptive statistics box



จากหน้าจอรูปที่ 5.5

- ให้เลือกตัวแปรอย่างน้อย 1 ตัว ใส่ใน box ของ variable (s) สำหรับเทคนิค Cluster จะต้องเลือกตัวแปรทุกตัวที่จะใช้แบ่งกลุ่ม Case แล้วเลือก
- ☑ Save Standardized values as variables.

ในกรณีนี้จะได้ตัวแปรใหม่อยู่ในรูป Z-score โดยตัวแปรใหม่ทุกตัวจะอยู่ในแฟ้มข้อมูลเดิมต่อท้ายจากตัวแปรที่มีในแฟ้มเดิม และตัวแปรใหม่ทุกตัวจะมีชื่อเหมือนตัวแปรเดิม แต่นำหน้าด้วยตัว Z ซึ่งหมายถึงตัวแปรเดิมที่คำนวณให้อยู่ในรูป Z-score ดังแสดงในรูปที่ 5.6

Cars.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

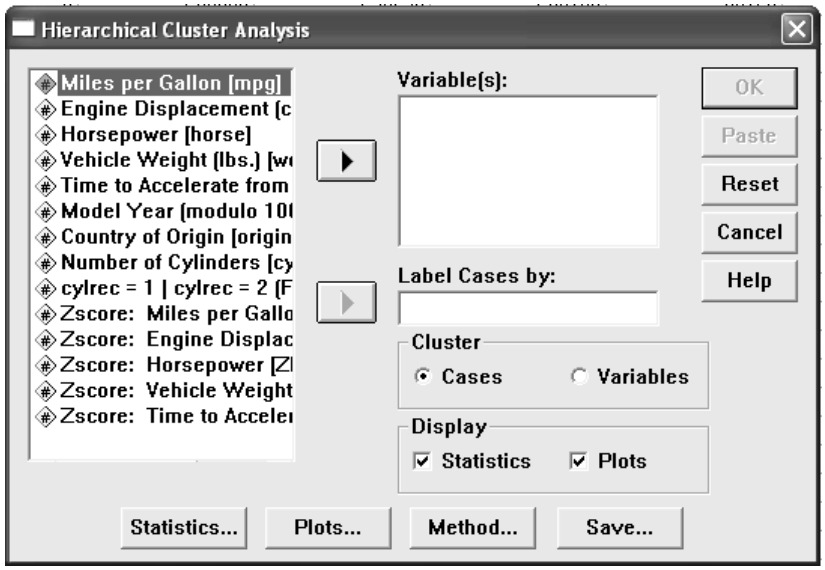
1 : mpg 18

	origin	cylinder	filter_$	Zmpg	Zengine	Zhorse	Zweight	Zaccel
1	1	8	0	-.70555	1.07368	.65333	.62888	-1.23896
2	1	8	0	-1.08938	1.48240	1.56190	.85128	-1.41620
3	1	8	0	-.70555	1.17824	1.17251	.54886	-1.59344
4	1	8	0	-.96144	1.04517	1.17251	.54533	-1.23896
5	1	8	0	-.83349	1.02616	.91292	.56416	-1.77068
6	1	8	0	-1.08938	2.23330	2.41855	1.61379	-1.94793
7	1	8	0	-1.21732	2.47092	2.98965	1.62908	-2.30241
8	1	8	0	-1.21732	2.33785	2.85985	1.57966	-2.47966
9	1	8	0	-1.21732	2.48043	3.11945	1.71263	-1.94793
10	1	8	0	-1.08938	1.86260	2.21088	1.03602	-2.47966
11	2	4	1	.	-.58019	.26394	.14172	.71072
12	1	8	0	.	1.48240	1.56190	1.37962	-1.41620
13	1	8	0	.	1.49190	1.25039	1.25254	-1.59344

รูปที่ 5.6 Z-score

ขั้นที่ 2 : ใช้คำสั่งการจัดกลุ่มใน ดังนี้

Analyze → Classify → Hierarchical Cluster … จะได้หน้ารูปที่ 5.7



รูปที่ 5.7 : Hierarchical Cluster Dialog box

จากรูปที่ 5.7 อธิบายได้ดังนี้

ส่วนที่ 1 : Variable (s) box ถ้าต้องการจัดกลุ่ม Case จะต้องเลือกตัวแปรที่มีค่าเป็นตัวเลข (Numeric variable) อย่างน้อย 1 ตัว แต่ถ้าต้องการจัดกลุ่มตัวแปร จะต้องเลือกตัวแปรที่มีค่าเป็นตัวเลขอย่างน้อย 3 ตัว

ส่วนที่ 2 : Label Case By เป็นการระบุชื่อ Case หรือความหมายของ Case เช่น ถ้าแบ่งกลุ่มจังหวัด กรณีนี้ 1 Case คือ 1 จังหวัด ถ้าสร้างตัวแปร Province ที่ระบุชื่อจังหวัด จะเลือกตัวแปร Province มาใส่ในนี้ โดยที่ตัวแปรที่จะอยู่ใน box ของ Label Cases by จะต้องเป็นตัวแปร Nominal และเป็นชนิด String ถ้าไม่เลือกตัวแปรใส่ใน Box ของ Label Cases by ผลลัพธ์จะให้หมายเลข Case

ส่วนที่ 3 : Cluster ผู้วิเคราะห์ต้องเลือกว่าต้องการจัดกลุ่ม Case หรือจัดกลุ่มตัวแปรอย่างใดอย่างหนึ่งเพียงอย่างเดียว

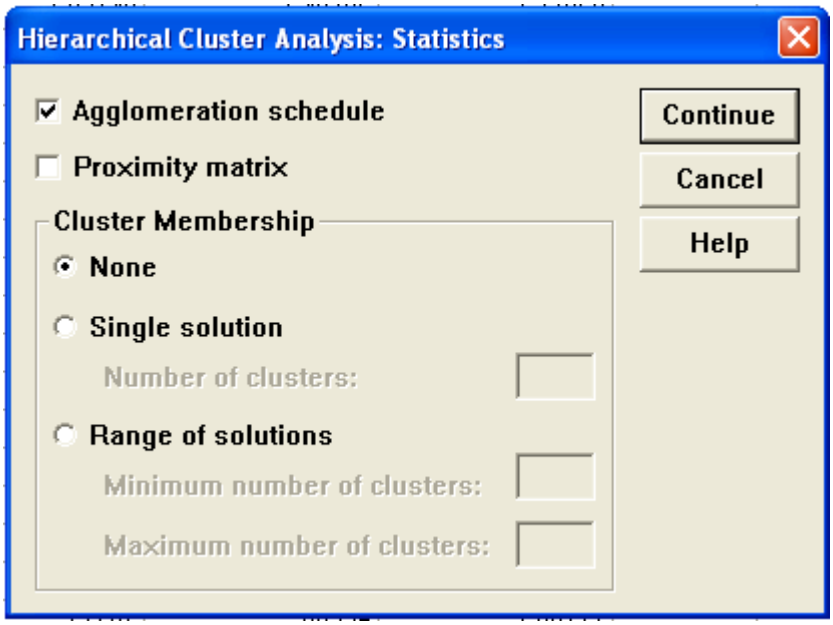
- Cases เลือกทางเลือกนี้ ถ้าต้องการจัดกลุ่ม Case

- Variables เลือกทางเลือกนี้ ถ้าต้องการจัดกลุ่มตัวแปร

ส่วนที่ 4 : Display ผู้ใช้สามารถเลือกให้ผลลัพธ์แสดงทั้งค่าสถิติ และกราฟ หรืออาจเลือกทางเลือกใดทางเลือกหนึ่งก็ได้

- Statistics แสดงค่าสถิติในผลลัพธ์
- Plots แสดงกราฟในผลลัพธ์

จากรูปที่ 5.7 เลือก Statistics... จะได้หน้าจอดังในรูปที่ 5.8



รูปที่ 5.8 : Hierarchical Cluster Analysis : Statistics

ในรูปที่ 5.8 แบ่งเป็น 2 ส่วนดังนี้

ส่วนที่ 1 : ส่วนนี้มี 2 ทางเลือก ผู้ใช้สามารถเลือกทางเลือกใดทางเลือกหนึ่ง หรือ 2 ทางเลือกก็ได้ ดังนี้

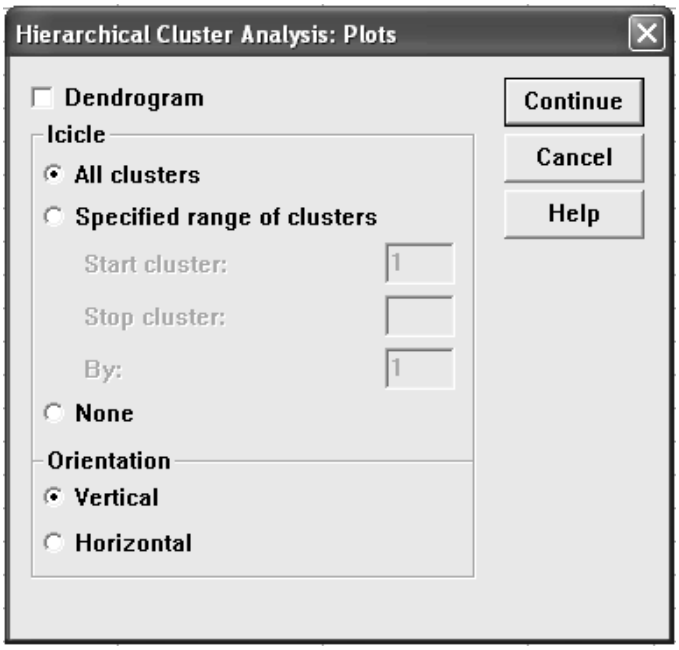
- Agglomeration schedule จะแสดงขั้นตอนการรวมกลุ่ม Case ดังได้อธิบายในหัวข้อ 3.3 และเป็น Default
- Proximity matrix จะแสดง Matrix ของระยะห่างระหว่าง Case แต่ละคู่

ส่วนที่ 2 : Cluster Membership จะแสดงว่าแต่ละ Case เป็นสมาชิกกลุ่มใด หรือ Cluster ใด ผู้ใช้สามารถเลือกใดทางเลือดหนึ่งจากต่อไปนี้

- None ไม่แสดงกาหรเป็นสมาชิกของ Case ทางเลือกนี้เป็น Default

- Single solutions จะแสดงสมาชิกของ cluster โดยกำหนดจำนวน Cluster (กลุ่ม) ที่ต้องการโดยต้องใส่เลขจำนวนเต็มที่มีค่าตั้งแต่ 1 ขึ้นไป เช่น ถ้าต้องการสมาชิกของกลุ่ม 3 กลุ่ม ใส่หมายเลข 3 ลงใน
- Range of solutions จะแสดงสมาชิกของ Cluster โดยกำหนดช่วงของจำนวนกลุ่ม โดยต้องระบุจำนวนกลุ่มต่ำสุด และสูงสุด โดยเลขที่ใส่ใน ทั้งสองจะต้องเป็นเลขจำนวนเต็ม มีค่าตั้งแต่ 2 ขึ้นไป และค่าแรกต้องน้อยกว่าค่าที่สองเสมอ

จากหน้าจอที่ 5.7 เลือก Plots... จะได้หน้าจอดังรูปที่ 5.9



รูปที่ 5.9 :Hierarchical Cluster Analysis :Plots

ในรูปที่ 5.9 แบ่งออกเป็น 3 ส่วน ดังนี้

ส่วนที่ 1 : Dendrogram จะให้กราฟ ซึ่งแสดงถึงการรวมกันของ Cluster และให้ค่าระยะห่างในแต่ละขั้นตอนด้วย โดยจะเปลี่ยนหน่วยระยะห่างของข้อมูลเดิม เป็นระยะห่างมีค่าในช่วง 1 ถึง 25

ส่วนที่ 2 : Icicle หมายถึง Icicle Plots ซึ่งมี 3 ทางเลือก ให้ผู้ใช้เลือกทางเลือกใดทางเลือกหนึ่ง

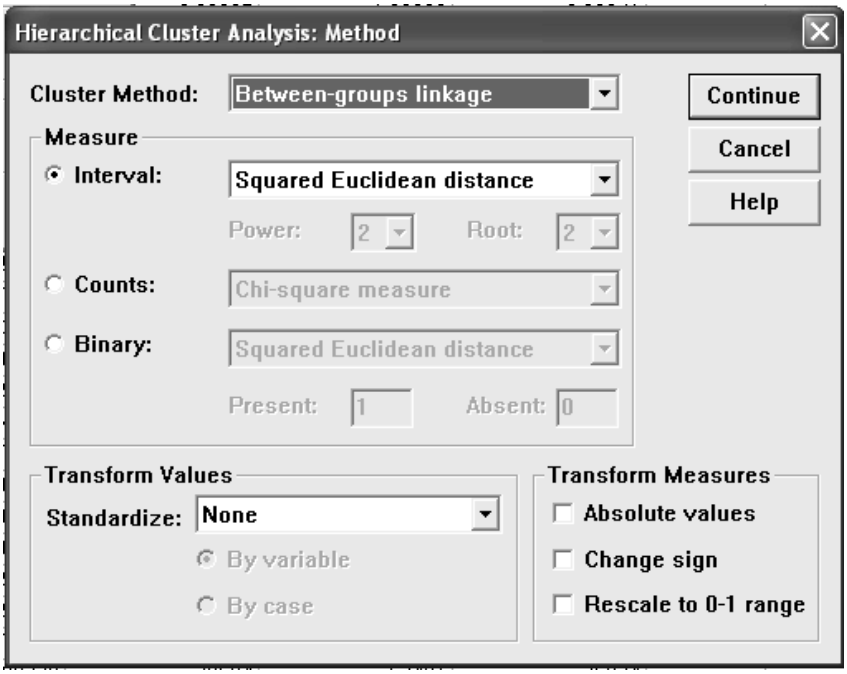
- All Clusters แสดง Icicle Plot ของทุก Cluster

- Specified range of clusters แสดง Icicle Plot ตามช่วงของจำนวน Cluster ที่กำหนดโดยใส่เลขจำนวนเต็มบวกในช่อง Start, Stop และ By โดย Start น้อยกว่า Stop ส่วน By หมายถึง การเพิ่มขึ้นครั้งละ เช่น ใส่เลข 3, 7 และ 2 จะทำให้ Icicle Plot แสดง 3, 5, 7 กลุ่มหรือ Cluster เป็นต้น
- None ไม่แสดง Icicle Plot

ส่วนที่ 3 : Orientation มีทางเลือกดังนี้

- Vertical แสดง Icicle Plot ในแนวตั้ง
- Horizontal แสดง Icicle Plot ในแนวนอน

จากหน้าจอที่ 5.7 เลือก **Method...** จะได้หน้าจอดังรูปที่ 5.10



รูปที่ 5.10:Hierarchical Cluster Analysis :Method

ในรูปที่ 5.10 แบ่งออกเป็น 4 ส่วน

ส่วนที่ 1 : Cluster Method เลือกวิธีการรวมกลุ่ม Cluster ที่อธิบายไว้แล้วในหัวข้อ 3.3 ผู้ใช้สามารถคลิกเครื่องหมาย ซึ่งมีวิธีในการรวมกลุ่ม Cluster

- Between-group linkage : Average linkage between groups (UPGMA)
- Within-group linkage : Average linkage within groups
- Nearest neighbor : Single linkage

- Furthest neighbor : Complete linkage
- Centroid clustering
- Medain clustering
- Ward's method

ส่วนที่ 2 : Measure วิธีการวัดระยะห่างและความคล้าย ซึ่งการเลือกวิธีการวัดระยะห่างหรือความคล้ายจะขึ้นกับชนิดของข้อมูลที่แบ่งเป็น 3 ประเภท ดังนี้

- Interval หมายถึง ข้อมูลชนิด Interval หรือ Radio scale จะคำนวณหาระยะห่างและความคล้ายโดยผู้ใช้ต้องเลือกวิธีการโดยการคลิ  จะได้

เพิ่มสูตร

- Count ใช้กับข้อมูลที่อยู่ในรูปความถี่ โดยวัดความแตกต่างหรือระยะห่าง โดยเลือกวิธีการทางสถิติ ดังนี้

เพิ่มสูตร

- Binary ใช้กับข้อมูลที่มีค่าได้เพียง 2 ค่า โดย SPSS จะสร้างตาราง 2 X 2 ของ case ให้

  A, b, c, d คือความถี่

  วิธีการคำนวณระยะห่างมีหลายวิธีดังนี้

เพิ่มสูตร

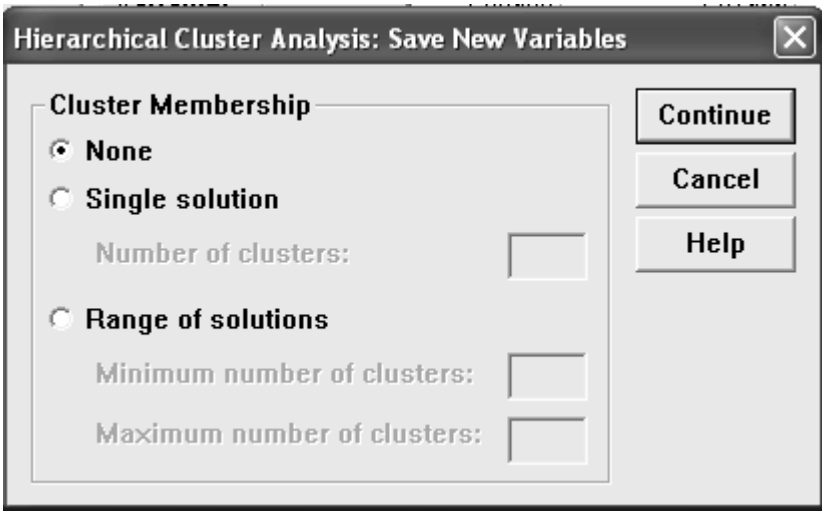
ส่วนที่ 3 : Transform Value เมื่อต้องการเปลี่ยนแปลงค่าของ case หรือตัวแปรเพื่อทำให้ตัวแปรมีความสำคัญเท่ากัน เมื่อข้อมูลเดิมมีสเกลต่างกัน โดยจะทำการ Standardize ข้อมูล

- Standardize ก่อนจะทำการคำนวณค่าระยะห่าง หรือความคล้าย สำหรับข้อมูลชนิด Interval หรือ Count เท่านั้น โดยผู้ใช้ต้องเลือก 1 ทางเลือก ดังต่อไปนี้
  - None ไม่ทำการ Standardize แต่ให้ใช้ข้อมูลเดิม
  - Z score ทำการ Standardize ข้อมูลให้เป็น Z score ที่มีค่าเฉลี่ย 0 ค่าเบี่ยงเบนมาตรฐาน 1
  - Range – 1 to 1 ทำ Standardize ข้อมูลให้มีค่าระหว่าง – 1 ถึง 1
  - Range 0 to 1 ทำ Standardize ข้อมูลให้มีค่าระหว่าง 0 ถึง 1

ส่วนที่ 4 : Transform Measure ใช้เฉพาะข้อมูลชนิด Interval หรือ Count เท่านั้น ใช้ในการ Standardize ข้อมูลสำหรับ Case หรือค่าของข้อมูลก่อน ที่จะคำนวณค่า proximity โดยมีทางเลือกดังนี้

- Absolute values จะคำนวณค่าสัมบูรณ์ของระยะห่าง
- Change sign เป็นการเปลี่ยนความคล้ายให้เป็นความไม่คล้าย (ความห่าง) หรือเปลี่ยนความไม่คล้ายให้เป็นความคล้าย
- Rescale to 0 – 1 range เป็นการเปลี่ยนระยะห่างให้มีค่าในช่วง 0 ถึง 1 ซึ่งถือเป็นการทำ Standardize อย่างหนึ่ง โดยการนำค่าระยะห่างที่สั้นที่สุดไปลบจากระยะห่างต่าง ๆ แล้วหารด้วยค่าพิสัยระยะห่าง

จากหน้าจอรูปที่ 5.7 คลิกปุ่ม Save... **จะได้หน้าจอรูปที่ 5.11**



รูปที่ 5.11: Save

ในหน้าจอรูปที่ 5.11 เป็นการให้ระบุกลุ่มที่ Case หรือตัวแปรเป็นสมาชิกอยู่ ในตาราง Cluster Membership ในผลลัพธ์ ซึ่งมีทางเลือกดังนี้

- None ไม่ต้องการบันทึกเลขที่กลุ่ม
- Single solution บันทึกเลขที่กลุ่มโดยที่ระบุจำนวนกลุ่มที่แน่นอนเพียงค่าเดียว
- Range of solutions ให้บันทึกเลขที่กลุ่มกรณีที่กำหนดว่าจำนวนกลุ่มหลาย ๆ แบบ เช่น จำนวนบันทึกเลขที่กลุ่มของแต่ละ case เมื่อแบ่งเป็น 2, 3, 4, 5 หมายถึงใส่ from เป็น 2 และ through เป็น 5 โดยที่ค่าที่ใส่ใน box ต้องเป็นเลขจำนวนเต็มบวกที่มากกว่า 1 และเลขใน box ที่สองต้องมีค่ามากกว่า box แรก

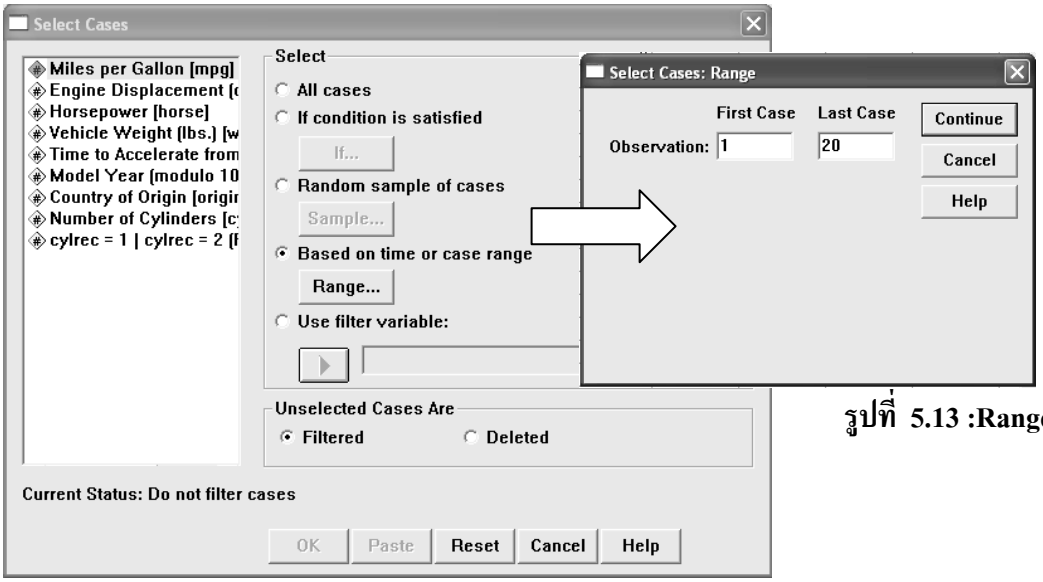
5.5 ตัวอย่างการใช้เทคนิค Hierarchical Cluster Analysis

<u>ตัวอย่างที่ 5.1</u> สำหรับตัวอย่างที่ 5.1 จะใช้เทคนิค Hieratchical Cluster แบ่งกลุ่มCase โดยไม่จำเป็นต้องทราบจำนวนกลุ่มที่แน่นอน และไม่ต้องทราบว่าแต่ละCaseอยู่กลุ่มใดบ้าง สำหรับตัวอย่างนี้จะใช้ข้อมูลแค่ 20 Caseแรกในการจัดกลุ่ม เนื่องจากไม่ต้องการให้ผลลัพธ์ที่ได้ยาวเกินไป จนทำให้ไม่สะดวกในการอธิบายความหมาย โดยมีขั้นตอนดังนี้

ขั้นที่ 1 : สร้างแฟ้มข้อมูล ซึ่งจะมี case หรือตัวแปรหลายๆตัว ที่จะนำมาใช้ในการแบ่ง case หรือแบ่งกลุ่มตัวแปร ซึ่งในที่นี้จะใช้แฟ้มข้อมูล cars ซึ่งมีอยู่ในโปรแกรม SPSS โดยใช้ข้อมูล แค่ 20 case แรกในการจัดกลุ่ม เนื่องจากไม่ต้องการให้ผลลัพธ์ที่ได้ยาวเกินไป

ขั้นที่ 2 : เลือก Case ที่ 1 – 20 เพื่อใช้ในการวิเคราะห์ โดยใช้คำสั่ง

Data → Select Cases … จะได้หน้าจอรูปที่ 5.12



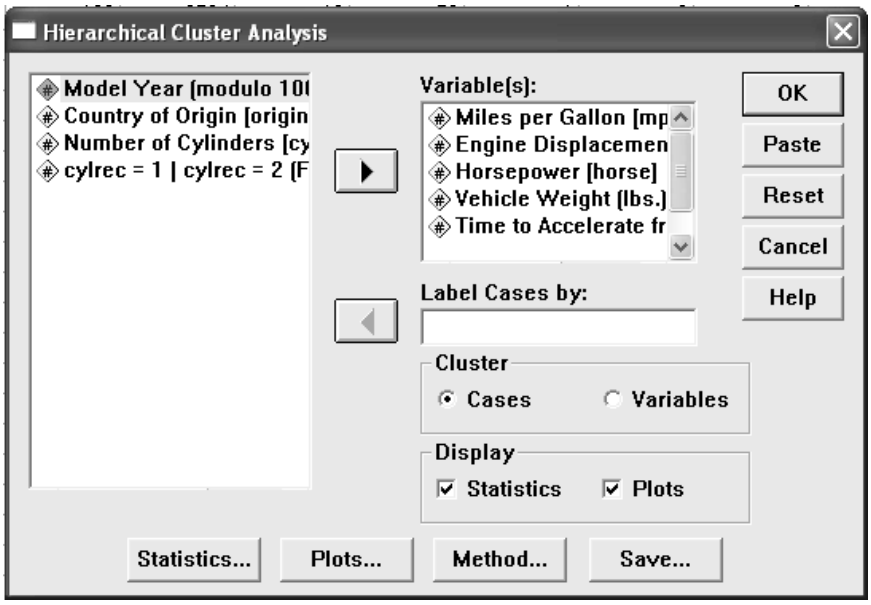
รูปที่ 5.13 :Range

รูปที่ 5.12 : Select Cases

- ในหน้าจอรูปที่ 5.12 เลือก ⊙ Based on time or case range
- คลิก Range... จะได้หน้าจอรูปที่ 5.13
- ใส่ 1 ใน First case และ 20 ใน box ของ Last case
- คลิก Continue และคลิก OK

ขั้นที่ 3 : ทำการแบ่งกลุ่มด้วยเทคนิค Hierarchical Cluster โดยใช้คำสั่ง

Analyze → Classify → Hierarchical Cluster … จะได้หน้าจอรูปที่ 5.14



รูปที่ 5.14 : Hierarchical Cluster

จากหน้าจอรูปที่ 5.14

- เลือกตัวแปรที่คาดว่าจะทำให้มีความแตกต่างระหว่างกลุ่มแตกต่างกัน จึงเลือกตัวแปร 4 ตัวดังกล่าวใส่ใน box ของ Variables (s)
- ในส่วนของ Cluster เลือก ⊙ Cases เนื่องจากต้องการจัดกลุ่ม (Case)
- ในส่วน Display เลือก
  - ☑ Stratistics
  - ☑ Plots

จากหน้าจอที่ 5.14 คลิก **Statistics...** จะได้หน้าจอรูปที่ 5.8 เลือก

- ☑ Agglomeration schedule
- ☑ Proximity matrix
- ⊙ Range of solutions แล้วป้อนค่า
  - From 2 though 4 clusters
  - คลิก **Continue** กลับไปหน้าจอรูปที่ 5.14

จากหน้าจอรูปที่ 5.14 คลิก **Plots...** จะได้หน้าจอรูปที่ 5.9

- เลือก ☑ Dendogram
- ในส่วนของ Icicle เลือก ⊙ All Clusters

- คลิก Continue จะกลับไปหน้าจอรูปที่ 5.14

จากหน้าจอรูปที่ 5.14 คลิก Method... จะได้หน้าจอรูปที่ 5.10

- ในส่วน Cluster Method เลือก Between – groups Linkage
- ในส่วนของ Measure เลือก ⊙ Interval เนื่องจากตัวแปรทั้ง 5 ตัวที่เลือกเป็นข้อมูล Ratio scale และเลือก Square Euclidean distance
- ในส่วนของ Transform Values เลือก Z scores เนื่องจากตัวแปรทั้ง 4 ตัวข้างต้นมีหน่วยที่แตกต่างกัน และ ⊙ By Variable
- คลิก Continue จะกลับไปหน้าจอรูปที่ 5.14

จากหน้าจอรูปที่ 5.14 คลิก Save... จะได้หน้าจอรูปที่ 5.11 ในหน้าจอรูปที่ 5.11

- เลือก ⊙ Range of solution :
  - From 2 through 4 cluster
- คลิก Continue และ OK จะได้ผลลัพธ์ดังแสดงในตารางที่ 5.3 – 5.7 และรูปที่ 5.15

ตารางที่ 5.3

Case Processing Summary [a]

Cases					
Valid		Missing		Total	
N	Percent	N	Percent	N	Percent
14	70.0%	6	30.0%	20	100.0%

a. Squared Euclidean Distance used

ตารางที่ 5.2 ระบุว่าจากข้อมูล 20 Case มีค่า Missing อยู่ 6 จึงมีจำนวนCaseนำมาวิเคราะห์เพียง 14 หรือคิดเป็น 70% (14/20)

Proximity Matrix

	Squared Euclidean Distance						
Case	1:Case 1	2:Case 2	3:Case 3	.....	9:Case 9	.....	14:Case 20
1:Case 1	.000	6.302	1.024	-	28.953	-	25.208
2:Case 2	6.302	.000	5.360	-	11.079	-	10.148
3:Case 3	1.024	5.360	.000	-	23.971	-	19.213
4:Case 4	2.319	1.800	2.603	-	21.153	-	16.350
5:Case 5	1.974	4.071	.797	-	22.681	-	18.117
6:Case 6	18.914	6.407	14.979	-	1.354	-	10.073
7:Case 7	30.418	12.413	24.340	-	.663	-	9.576
8:Case 8	30.160	12.580	23.700	-	1.599	-	9.887
9:Case 9	28.953	11.079	23.971	-	.000	-	9.970
10:Case 10	17.610	6.643	11.987	-	5.954	-	7.362
11:Case 16	9.841	1.796	6.585	-	8.723	-	5.856
12:Case 17	18.698	8.074	13.644	-	13.552	-	11.370
13:Case 19	11.307	3.368	8.191	-	8.887	-	8.969
14:Case 20	25.208	10.148	19.213	-	9.970	-	.000

This is a dissimilarity matrix

ความหมายของผลลัพธ์ตารางที่ 5.4 : Proximity Matrix

ค่าต่าง ๆ ในตารางที่ 5.4 เป็นระยะห่างของ Case แต่ละคู่โดยระยะห่างที่ใช้คือ ค่า Squared Euclidean Distance เช่น case 1 และ case 9 ห่างกัน 28.593 ขณะที่ case 1 และ case 3 ห่างกันเพียง 1.024 ดังนั้น ควรจัดCase case 1 และ case 3 ให้อยู่ในกลุ่มเดียวกัน นั่นคือ case 1 และ case 3 มีค่าตัวแปร 5 ตัว ดังกล่าวคล้ายกัน ในขณะเดียวกัน ควรจัด case 1 และ case 9 อยู่ต่างกลุ่มกัน หรือ case 1 และ case 3 มีความแตกต่างกันในตัวแปรทั้ง 5 ตัว

ตารางที่ 5.5

Agglomeration Schedule

	Cluster Combined			Stage Cluster First Appears		
Stage	Cluster 1	Cluster 2	Coefficients	Cluster 1	Cluster 2	Next Stage
1	7	8	.239	0	0	4
2	3	5	.797	0	0	5
3	11	13	.804	0	0	8
4	7	9	1.131	1	0	6
5	1	3	1.499	0	2	10
6	6	7	1.735	0	4	11
7	2	4	1.800	0	0	10
8	10	11	2.148	0	3	9
9	10	12	3.025	8	0	11
10	1	2	3.768	5	7	13
11	6	10	7.180	6	9	12
12	6	14	9.133	11	0	13
13	1	6	14.933	10	12	0

ความหมายของผลลัพธ์ตารางที

ผลลัพธ์ตารางที่ 5.5 เป็นผลจากการใช้วิธี Between – groups linkage ในหน้าจอรูปที่ 5.10 (หน้าจอ Method) ในการรวมกลุ่ม Case นั่นคือ ในแต่ละ Stage จะบอกว่ามีการรวม Case คู่ใดบ้าง ให้อยู่ในกลุ่มเดียวกัน เช่น

Stage 1 : จะจัดที่ 7 และ Case ที่ 8 อยู่ในกลุ่มเดียวกัน เนื่องจากCaseที่ 7 และ 8 มีระยะห่างกันสั้นที่สุด (จากตารางที่) ซึ่งระยะห่าง (ค่า Squared Euclidean Distance) คือค่าใน Column ของ Coefficients ซึ่งเท่ากับ .239 และค่า Next Stage ใน Column สุดท้าย = 4 หมายถึง กลุ่มหรือ Cluster ที่มีCaseที่ 7 และ 10 จะรวมกับCaseอื่นต่อไปใน stage ที่ 4

Stage 2 : มีการจัดให้ Caseที่ 3 และ Caseที่ 5 ให้อยู่ในกลุ่มหรือ Cluster เดียวกัน ซึ่ง Caseที่ 3 และ 5 มีระยะห่าง = .797 และกลุ่มที่มีCaseที่ 3 และ 5 อยู่จะรวมกับCaseอื่นอีกใน Stage ที่ 5 (Next Stage = 5)

Stage 4 : มีการจัด Caseที่ 7 และ 9 ให้อยู่ในกลุ่มเดียวกัน แต่Caseที่ 7 อยู่กลุ่มเดียวกับCaseที่ 8 ในขั้นที่ 1 แล้ว โดยพิจารณา Column ของ Stage Cluster First Appears ในส่วนของ Cluster 1 = 1 เป็นการระบุว่า Caseที่ 7 ถูกรวมกับCaseที่ 8 ใน Stage ที่ 1 แล้ว ดังนั้น Caseที่ 7 และ 8 และ 9 จะรวมอยู่ในกลุ่มเดียวกัน และจาก Column ของ Next Stage = 6 แสดงว่าจะมีCaseใหม่อีก 1 Case มารวมกับกลุ่มนี้ใน Stage ที่ 6

สำหรับการรวม Caseที่ 9 เข้าในกลุ่มเดิมที่มีอยู่แล้ว (กลุ่มที่มี Case 7 และ 8) จะใช้วิธี Between-groups linkage (Average Linkage) นั่นคือ ใช้ค่าเฉลี่ยของระยะห่างระหว่างCase 9กับ Case 7 และระยะห่างระหว่างCase 9 และ 8 (จากตารางที่ 5.4)

-
-
-
-

⇩

Stage 10 : จะมีการนำCaseที่ 2 มารวมกับกลุ่มที่มีCaseที่ 2 และ 4 อยู่แต่Caseที่ 1 นี้ถูกรวมอยู่ในกลุ่มที่มีCase ที่ 3 Stage ที่ 51 และเป็นเช่นนี้ไปเรื่อย ๆ จนถึง Stage ที่ 13 จะเป็นการรวมทุก Case อยู่ในกลุ่มเดียวกันซึ่งจะแสดงด้วยกราฟในรูปที่ : Dendogram

ดังที่ได้กล่าวแล้วว่าเทคนิค Cluster ในขั้นแรกจะให้ จำนวนกลุ่ม = จำนวน Case นั่นคือ ในตัวอย่างนี้มี 14 Case (เนื่องจากมีการ Missing 6 Case) จึงเริ่มต้นมี 14 กลุ่ม ๆ ละ 1 Case แล้วจึงค่อย ๆ รวม Case ทีละคู่ ดังในตารางที่ จนในที่สุดเหลือกลุ่มเดียว ดังนั้น การพิจารณาว่าควรแบ่งเป็นกี่กลุ่มย่อยจึงอยู่ที่การพิจารณาของผู้วิเคราะห์ โดยจะพิจารณาจากระยะห่างหรือความคล้าย

ตารางที่ 5.6

Cluster Membership

Case	4 Clusters	3 Clusters	2 Clusters
1:Case 1	1	1	1
2:Case 2	1	1	1
3:Case 3	1	1	1
4:Case 4	1	1	1
5:Case 5	1	1	1
6:Case 6	2	2	2
7:Case 7	2	2	2
8:Case 8	2	2	2
9:Case 9	2	2	2
10:Case 10	3	2	2
11:Case 16	3	2	2
12:Case 17	3	2	2
13:Case 19	3	2	2
14:Case 20	4	3	2

ความหมายของผลลัพธ์ตารางที่ 5.6

ตารางที่ เป็นการระบุว่าแต่ละ Case อยู่กลุ่มใดโดยแบ่งเป็น 3 รูปแบบ (ตามที่ระบุในหน้าจอรูปที่ 5.11) ดังนี้

1) กรณีที่มี 4 กลุ่ม (4 Clusters) เมื่อมี 4 กลุ่ม คือ

กลุ่มที่ 1 : มี 5 Case คือ 1, 2, 3, 4 และ 5

กลุ่มที่ 2 : มี 4 Case คือ 6, 7, 8, และ 9

กลุ่มที่ 3 : มี 4 Case คือ 10, 16, 17 และ 19

กลุ่มที่ 4 : มี 1 Case คือ 20

2) กรณีที่มี 3 กลุ่ม (3 Clusters) เมื่อมี 3 กลุ่ม

กลุ่มที่ 1 : มี 5 Case คือ 1, 2, 3, 4 และ 5

กลุ่มที่ 2 : มี 8 Case คือ 6, 7, 8, 9, 10, 16, 17 และ 19

กลุ่มที่ 3 : มี 1 Case คือ 20

จะพบว่าทั้งกรณีที่มี 4 กลุ่มและ 3 กลุ่ม Case 20 จะเป็นCase เดียวที่ไม่สามารถรวมกลุ่มกับCaseอื่น ๆ ได้ เนื่องจากแตกต่างจากCaseอื่น ๆ มาก (ระยะห่างมาก)

3) กรณีที่มี 2 กลุ่ม (2 Clusters)

กลุ่มที่ 1 : มี Case คือ 1, 2, 3, 4 และ 5

กลุ่มที่ 2 : มี 13 Case คือ 6, 7, 8, 9, 10, 16, 17,19และ 20

จะพบว่ากรณีที่มี 3 กลุ่มและ 2 กลุ่ม กลุ่มที่ 2 จะเหมือนกัน กลุ่มที่ 1 ก็จะเหมือนกันยกเว้นCase 20 ที่ถ้าแบ่งเป็น 2 กลุ่มจะถูกจัดให้อยู่ในกลุ่มที่ 1 ทั้งนี้เนื่องจากระยะห่างเฉลี่ยของCase 20 กับCaseอื่น ๆ ในกลุ่มที่ 1 สั้นกว่ากลุ่มที่ 2 หรือมีความคล้ายCase อื่น ๆ ในกลุ่มที่ 1 มากกว่า

การพิจารณาว่าควรจัดแบ่งCaseเป็นกี่กลุ่มจึงจะเหมาะสม จะพิจารณาจากระยะห่างในตารางที่ 5.4, ค่า Coefficient ในตารางที่ 5.5, ตารางที่ 5.6 และรูปที่ 5.15 – 5.16

ตารางที่ 5.7

Vertical Icicle

Number of clusters	Case																													
	14:Case 20		12:Case 17		13:Case 19		11:Case 16		10:Case 10		9:Case 9		8:Case 8		7:Case 7		6:Case 6		4:Case 4		2:Case 2		5:Case 5		3:Case 3		1:Case 1	
1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
2	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X
3	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X
4	X		X	X	X	X	X	X	X		X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X
5	X		X	X	X	X	X	X	X		X	X	X	X	X	X	X		X	X	X		X	X	X	X	X	X	X	X
6	X		X		X	X	X	X	X		X	X	X	X	X	X	X		X	X	X		X	X	X	X	X	X	X	X
7	X		X		X	X	X		X		X	X	X	X	X	X	X		X	X	X		X	X	X	X	X	X	X	X
8	X		X		X	X	X		X		X	X	X	X	X	X	X		X		X		X	X	X	X	X	X	X	X
9	X		X		X	X	X		X		X	X	X	X	X		X		X		X		X	X	X	X	X	X	X	X
10	X		X		X	X	X		X		X	X	X	X	X		X		X		X		X	X	X		X	X	X	X
11	X		X		X	X	X		X		X		X	X	X		X		X		X		X	X	X		X	X	X	X
12	X		X		X		X		X		X		X	X	X		X		X		X		X	X	X		X	X	X	X
13	X		X		X		X		X		X		X	X	X		X		X		X		X		X		X	X	X	X

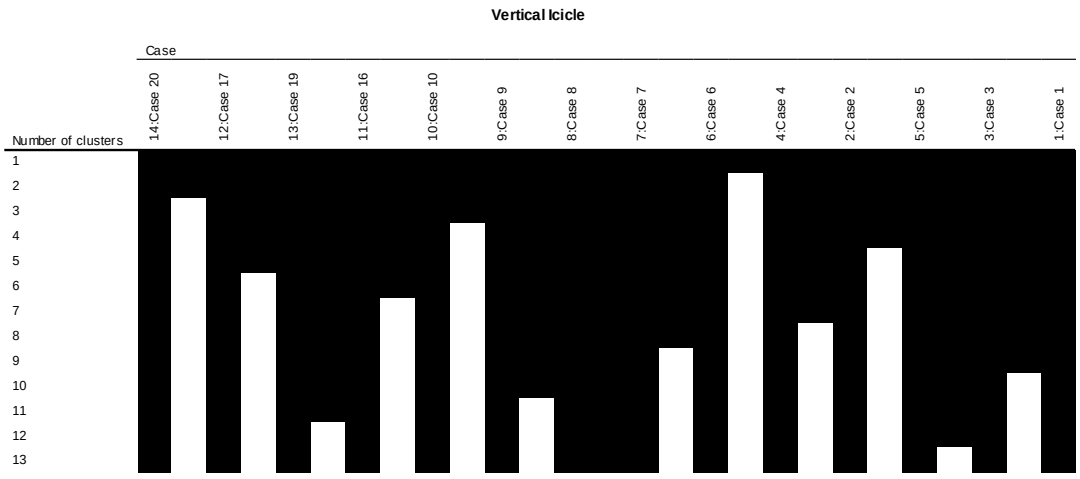
ความหมายของผลลัพธ์ตารางที่ 5.7

- ตารางที่ 5.7 เป็นผลจากการเลือก ⊙ All Clusters ในส่วน Icicle ของหน้าจอ Plots รูปที่ 5.9
- จะพบว่าถ้าในขั้นตอนใดมีการรวมCaseก็จะเชื่อมด้วยเครื่องหมาย X
- ซึ่งจะพบว่าจะพิจารณาค่อนข้างยาก ดังนั้น จึงจะปรับตารางที่ 5.7 เป็นรูปที่ 5.15 ทำให้พิจารณาง่ายขึ้นกว่าในแต่ละขั้นมีการรวม Case ใดบ้าง

➢ การสร้างรูปที่ 5.15 มีขั้นตอนดังนี้

1. เมื่ออยู่ที่หน้าจอผลลัพธ์ เลือก Edit ⇨ Options
2. เลือก Scripts tab
3. ในส่วนของ Autoscripts เลือก ☑ Enable Autoscripts
4. เลือก ☑ Cluster_Table_Icicle_Create แล้วคลิก
5. ใช้คำสั่ง Hierarchical Cluster.. ใหม่อีกครั้ง จะได้ผลลัพธ์ใหม่และตารางที่ 5.7 จะแสดงอยู่ในรูปกรอบของรูปที่ 5.15

รูปที่ 5.15



ความหมายของรูปที่ 5.15

- Block bar ที่อยู่ส่วนบนของตาราง หมายถึง แต่ละCase

- ในแถวที่ 1 หรือเมื่อมี 1 กลุ่ม หรือ 1 Cluster จะเป็นสีดำหมด หมายถึงทุก Case เชื่อมกันหรือรวมอยู่ในกลุ่มเดียวกัน
- ในแถวสุดท้ายหรือเมื่อมี 13 กลุ่มหรือ 18 Clusters จะพบว่าCaseที่ 7 กับ 8 จะรวมอยู่ในกลุ่มเดียวกัน เนื่องจากมีการระบายสีดำเชื่อมCase 7 และ 8
- ในแถวที่ 12 หรือเมื่อมี 17 Clusters จะรวมCaseที่ 3 และ 5 หรือCase 3และ 5 เข้าอยู่ในกลุ่มเดียวกัน เนื่องจากมีการระบายสีดำเชื่อมCase 3 และ 5
- ในแถวที่ 11 หรือเมื่อมี 11 Clusters จะรวมCaseที่ 19 หรือCase 16

5.6 การพิจารณาเลือกจำนวนกลุ่มที่เหมาะสม

ดังได้กล่าวมาแล้วว่า ผลลัพธ์ของเทคนิค Cluster ไม่ได้ให้ค่าสถิติ หรือผลการทดสอบสมมติฐานเพื่อให้ตัดสินใจหาจำนวนกลุ่มที่เหมาะสม ผู้วิเคราะห์จะต้องพิจารณาความเหมาะสมเอง โดยอาจใช้ระยะห่าง หรือความคล้าย โดยใช้ dendogram ซึ่งผู้วิเคราะห์จะสามารถพิจารณาจำนวนกลุ่มจาก dengdogram โดยการกำหนดตัวเลขระหว่าง หรือความคล้ายเป็นเกณฑ์ในการตัดสินใจ

1) การใช้ Dendogram

สำหรับ Dendogram ถ้ากำหนดระยะห่างระหว่างกลุ่ม เป็นหน่วยที่แตกต่างกันไปก็จะได้จำนวน Cluster ที่แตกต่างกันไป คือยิ่งระห่างยิ่งมาก จำนวน Cluster ก็จะเพิ่มขึ้น

2) การพิจารณาลักษณะ (Profile) ของแต่ละกลุ่มย่อย

จากการใช้คำสั่ง Save หน้าจอรูปที่ 5.11 เมื่อเลือก ⊙ Rang of solutions และใส่จำนวน Cluster เป็น 2 – 4 จะทำให้โปรแกรม SPSS สร้างตัวแปรใหม่ในแฟ้มข้อมูลอีก 3 ตัว คือ clu4_1, clu3_1 และ clu2_1 โดยที่

clu4_1 หมายถึงตัวแปรที่แสดงเลขที่กลุ่มขอแต่ละ case ส่วนเลข 4 หมายถึงมี 4 กลุ่ม หรือ 4 clusters และ 1 หมายถึงการวิเคราะห์ครั้งที่ 1

clu2_1 เป็นตัวแปรที่แสดงเลขที่ Cluster ของแต่ละ case กรณีที่มี 2 clusters และเป็นการวิเคราะห์ครั้งที่ 1

origin	cylinder	filter_$	CLU4_1	CLU3_1	CLU2_1
1	8	0	1	1	1
1	8	0	1	1	1
1	8	0	1	1	1
1	8	0	1	1	1
1	8	0	1	1	1
1	8	0	2	2	2
1	8	0	2	2	2
1	8	0	2	2	2
1	8	0	2	2	2
1	8	0	3	2	2
2	4	1	.	.	.
1	8	0	.	.	.
1	8	0	.	.	.
1	8	0	.	.	.
1	8	0	.	.	.
1	8	0	3	2	2
1	8	0	3	2	2
1	8	0	.	.	.
1	8	0	3	2	2
1	8	0	4	3	2

รูปที่ 16

แสดงค่าของตัวแปร clu4_1, clu3_1 และ clu2_1

ถ้าในหน้าจอ Hierachical Cluster Analysis รูปที่ 12.15 เลือกตัวแปร Company ใส่ใน Label cases by โปรแกรมจะไม่มีการ Save ตัวแปร clu4_1, clu3_1 และ clu2_1 ให้ แต่จะมี Warning ดังนี้

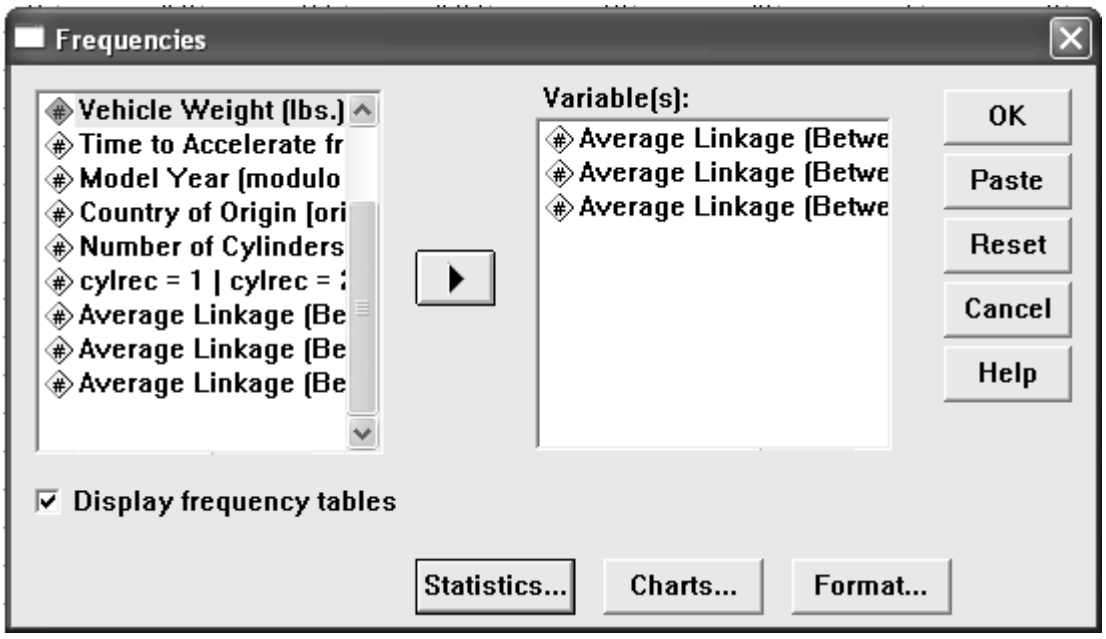
Warnings

SAVE will not be performed, since original case numbers unknown	!! Warnings ระบุว่าจะไม่มีการ save ให้ตามที่เลือกในหน้าจอ Save

ดังนั้นในหน้าจอ Hierarchical Clusters จะต้องไม่เลือกตัวแปรใส่ใน box ของ Label Cases by โปรแกรม SPSS จึงจะ Save ตัวแปร clu4_1, clu3_1 และ clu2_1 ให้ในแฟ้มข้อมูล ซึ่งถือว่าตัวแปร clu4_1, clu3_1 clu2_1 ในรูปที่ 5.16 เป็นตัวแปรใหม่ และสามารถนำตัวแปรเหล่านี้มาวิเคราะห์ต่อไปนี้ โดยมีขั้นตอนดังนี้

ขั้นที่ 1 : หาจำนวนCase หรือ Cases ในแต่ละ Cluster โดยใช้คำสั่งดังนี้

Analyze ⇨ Descriptive Statistics ⇨ Frequencies … จะได้หน้าจอรูปที่ 5.17



รูปที่ 17: Frequencies

- เลือกตัวแปร clu2_1, clu3_1 และ clu3_1 ใส่ใน box ของ Variable (s)
- เลือก ☑ Display frequency tables
- คลิก OK จะได้ผลลัพธ์ดังตารางในรูปที่ 5.19
-

Average Linkage (Between Groups)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	5	1.2	35.7	35.7
	2	9	2.2	64.3	100.0
	Total	14	3.4	100.0	
Missing	System	392	96.6		
Total		406	100.0		

Average Linkage (Between Groups)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	5	1.2	35.7	35.7
	2	8	2.0	57.1	92.9
	3	1	.2	7.1	100.0
	Total	14	3.4	100.0	
Missing	System	392	96.6		
Total		406	100.0		

Average Linkage (Between Groups)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	5	1.2	35.7	35.7
	2	4	1.0	28.6	64.3
	3	4	1.0	28.6	92.9
	4	1	.2	7.1	100.0
	Total	14	3.4	100.0	
Missing	System	392	96.6		
Total		406	100.0		

รูปที 5.18

ความหมายของผลลัพธ์รูปที่ 5.18

1. แสดงจำนวนและเปอร์เซ็นต์ของแต่ละ Cluster เมื่อแบ่งเป็น 2 Clusters

Cluster ที่ 1 มี 5 Case หรือร้อยละ 35.7

Cluster ที่ 2 มี 9 Case คิดเป็นร้อยละ 64.3

2. ใช้เมื่อแบ่งเป็น 3 Clusters จะพบว่ามีการแบ่ง Cluster ที่ 1 มี 5 Case เหมือนเดิม Cluster ที่ 2 มี 8 Case จากเดิม เป็น 9

3. แสดงกรณีที่แบ่งเป็น 4 Clusters จะพบว่ามีการแบ่ง Cluster ที่ 2 มี 4 Case จากเดิมเป็น 8 Case และ Cluster ที่ 3 มี 4 Case และ Cluster ที่ 4 มี 1 Case

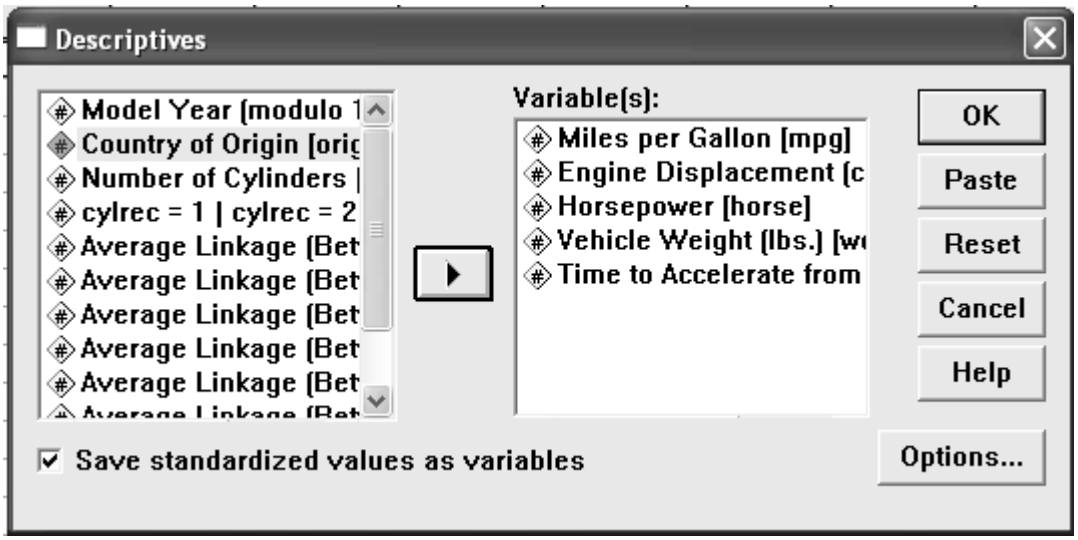
การพิจารณาว่าจำนวน Cluster ควรเป็น 2 หรือ 3 หรือ 4 นอกจากจะใช้ Dendogram ดังที่ได้กล่าวมาแล้ว ยังอาจจะพิจารณาจากจำนวน ดังแสดงในตารางที่ 12.20 และกราฟ

ขั้นที่ 2 : การสร้างกราฟแสดงค่าเฉลี่ยของตัวแปรที่ใช้แบ่งกลุ่ม

2.1) ปรับค่าตัวแปร mpq, engine, horse ,weight และ accel ให้อยู่ในรูป Standardized เพื่อกำจัดความแตกต่างของหน่วย โดยทำดังนี้

Analyze ⇨ Descriptive Statistics ⇨ Descriptives … จะได้หน้าจอรูปที่ 5.19

รูปที่ 5.19

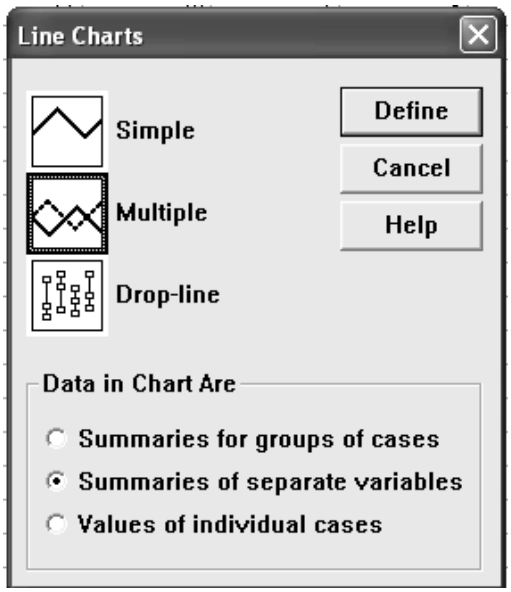


ในหน้าจอรูปที่ 12.21 เลือกตัวแปร Miles per Gallon, Engine Displacement, Horsepower, Vehicle Weight และ Time to Accelerate ใส่ใน Variable (s) box

- เลือก ☑ Save standardized values as variables
- คลิกปุ่ม OK จะได้ผลลัพธ์เป็นค่าตัวแปร zsize, zrevenue, zyears และ zproduct อยู่ในแฟ้มข้อมูล ซึ่งเป็นตัวแปรที่ Standardized แล้ว

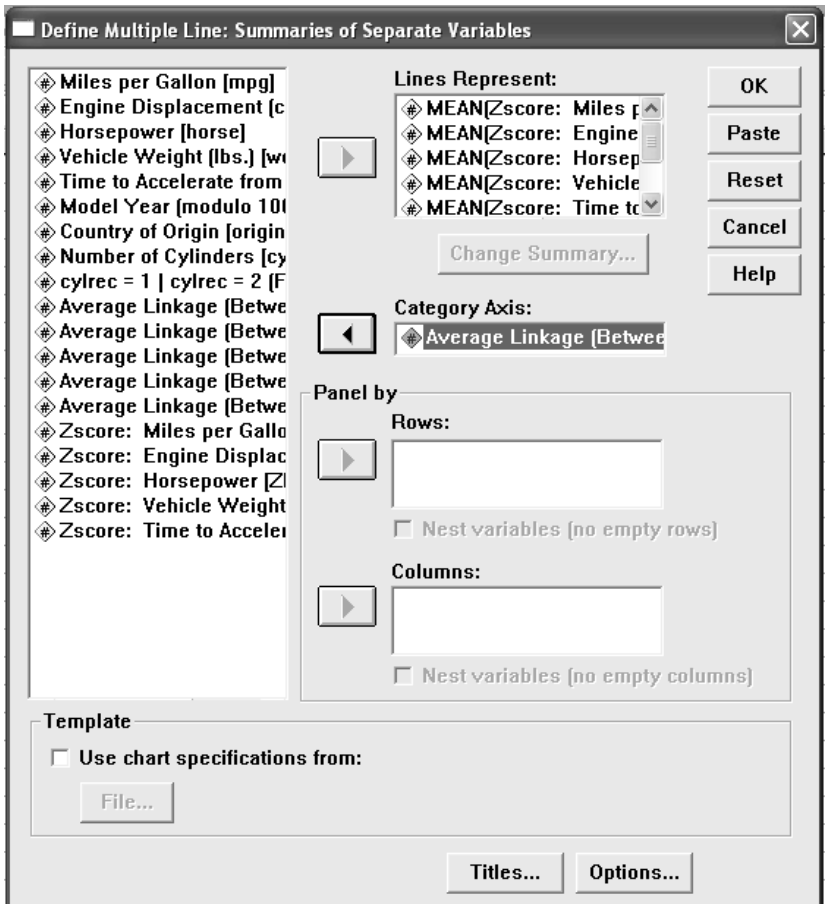
2.2) ใช้คำสั่ง Graphs ⇨ Line จะได้หน้าจอรูปที่ 5.20

รูปที่ 5.20: Line Charts



- เลือก Multiple
- ส่วนของ Data In Chart Are เลือก ⊙ Summaries of separate variables
- คลิกปุ่ม **Define** จะได้หน้าจอรูปที่ 5.21

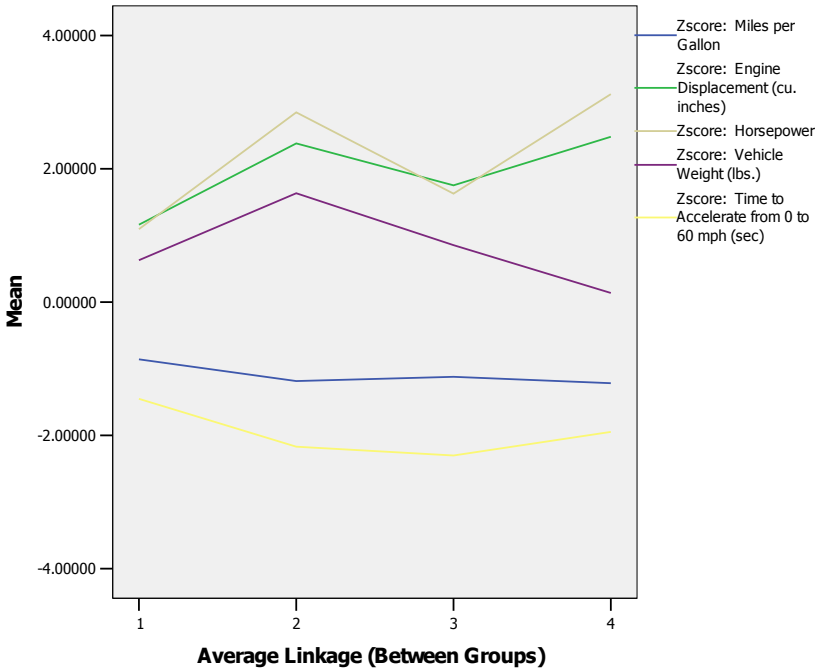
รูปที่ 5.21 : Multiple Line Charts



จากหน้าจอรูปที่ 5.21

- เลือกตัวแปร ใส่ใน box ของ Lines Represent ดังรูป
- เลือกตัวแปร clu4_1 ใส่ใน box ของ Category Axis
- คลิกปุ่ม **OK** จะได้ผลลัพธ์ดังรูปที่ 5.22

รูปที่ 5.22 : Line Chart



6. K-Means Clustering

ดังที่กล่าวแล้วว่า เทคนิคย่อยของ Cluster Analysis มี 2 เทคนิค คือ Hierarchical Cluster Analysis และ K-Means Cluster Analysis โดยในหัวข้อ 5 ได้กล่าวถึงเทคนิค Hierarchical โดยละเอียดแล้ว สำหรับหัวข้อ 6 นี้จึงจะกล่าวถึงเทคนิค K-Means Clustering

6.1 เทคนิค K-Means Clustering

เป็นเทคนิคการจำแนก Case ออกเป็นกลุ่มย่อย จะใช้เมื่อมีจำนวน Case มาก โดยจะต้องกำหนดจำนวนกลุ่ม หรือจำนวน Cluster ที่ต้องการ เช่นกำหนดให้มี k กลุ่ม เทคนิค K-Means จะมีการทำงานหลาย ๆ รอบ (Iteration) โดยในแต่ละรอบจะมีการรวม Cases ให้ไปอยู่ในกลุ่มใดกลุ่มหนึ่ง โดยเลือกกลุ่มที่ Case นั้นมีระยะห่างจากค่ากลางของกลุ่มน้อยที่สุด แล้วคำนวณค่ากลางของกลุ่มใหม่ จะทำเช่นนี้จนกระทั่งค่ากลางของกลุ่มไม่เปลี่ยนแปลง หรือครบจำนวนรอบที่กำหนดไว้

6.2 ชนิดของตัวแปรที่ใช้ในเทคนิค K-Means Clustering

ตัวแปรที่ใช้ในเทคนิค K-Means Clustering จะต้องเป็นตัวแปรเชิงปริมาณ คือเป็นสเกลอันตรภาค (Interval Scale) หรือสเกลอัตราส่วน (Ration Scale) โดยไม่สามารถใช้กับข้อมูลที่อยู่ในรูปความถี่ หรือ Binary เหมือนเทคนิค Hierarchical

6.3 ข้อแตกต่างระหว่างเทคนิค Hierarchical กับวิธี K-Means

1. เทคนิค K-Means ใช้เมื่อมีจำนวน Case หรือจำนวนข้อมูลมาก โดยทั่วไปนิยมใช้เมื่อ $n \geq 200$ เพราะเมื่อ n มาก เทคนิค K-Means จะง่ายกว่า และใช้ระยะเวลาในการคำนวณน้อยกว่าการใช้เทคนิค Hierarchical หรือกล่าวได้ว่าเมื่อมีจำนวน Case ไม่มากรควรใช้เทคนิค Hierarchical

2. เทคนิค K-Means นั้น ผู้ใช้จะต้องกำหนดจำนวนกลุ่มที่แน่นอนไว้ล่วงหน้า กรณีที่ผู้วิเคราะห์ยังไม่แน่ใจว่าควรมีกี่กลุ่มจึงจะเหมาะสม ผู้วิเคราะห์อาจจะใช้วิธีใดวิธีหนึ่งดังต่อไปนี้

- ทำการวิเคราะห์ด้วยวิธี K-Means หลาย ๆ ครั้ง แต่ละครั้งกำหนดจำนวนกลุ่มแตกต่างกันไป เช่น เป็น 3, 4 หรือ 5 กลุ่ม แล้วพิจารณาหาจำนวนกลุ่มที่เหมาะสม แต่เมื่อมีข้อมูลมากวิธีนี้จะทำให้เสียเวลามาก
- ใช้ข้อมูลบางส่วนทำการวิเคราะห์โดยวิธี Hierarchical เพื่อหาจำนวนกลุ่มที่ควรจะเป็นจากนั้นจึงใช้เทคนิค K-Means กับข้อมูลทั้งหมดที่มี

3. เทคนิค Hierarchical นั้น ผู้วิเคราะห์จะ Standardized ข้อมูลหรือไม่ก็ได้ แต่โดยวิธี K-Means จะต้องทำการ Standardized ข้อมูลก่อนเสมอ

4. วิธี K-Means จะหาระยะห่างโดยวิธี Euclidean distance โดยอัตโนมัติ ขณะที่ Hierarchical ผู้วิเคราะห์มีสิทธิ์ที่จะเลือกวิธีการคำนวณระยะห่าง หรือความคล้ายได้

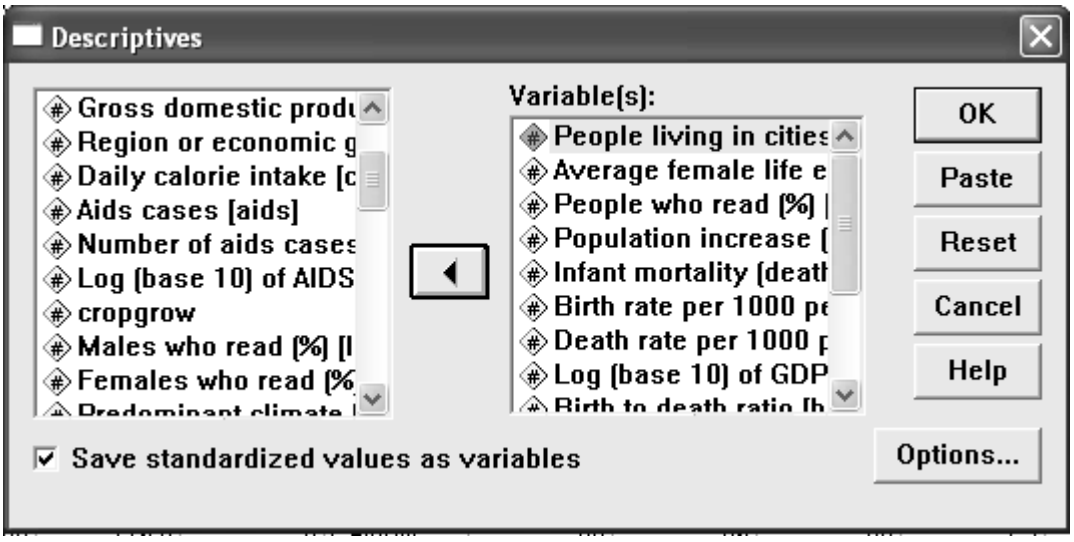
<u>ตัวอย่างที่ 6.1</u> ในตัวอย่างนี้จะใช้แฟ้มข้อมูลที่มีอยู่ในโปรแกรม SPSS คือแฟ้มข้อมูล World 95 for Missing Values มีใน SPSS Version 8 – 10 สำหรับ Version 7 ให้ใช้แฟ้ม World 95 แทน ซึ่งผลลัพธ์จะคล้าย ๆ กัน ถึงแม้แฟ้ม World 95 for Missing Values จะมีจำนวน case น้อยกว่า 200 แต่ก็มากพอที่จะใช้วิธี K-Means ได้

แฟ้ม World 95 for Missing Values เป็นแฟ้มแสดงตัวแปรต่าง ๆ ของแต่ละประเทศ จำนวน 109 ประเทศ

ขั้นที่ 1 : ทำการ Standardized ตัวแปรที่นำมาวิเคราะห์

Analyze ⇨ Descriptive Statistics ⇨ Descriptives … จะได้หน้าจอดังรูปที่ 6.1

รูปที่ 6.1 :Descriptives

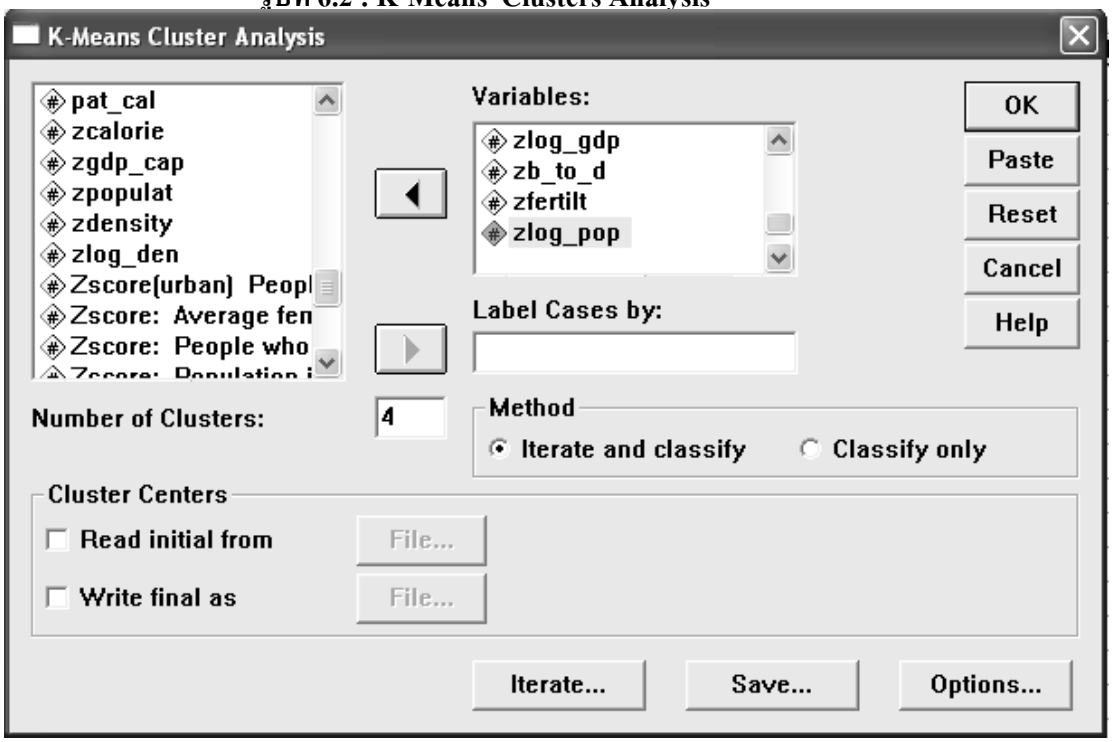


- ในหน้าจอรูปที่ 6.1 เลือกตัวแปร 11 ตัว คือ urban, lifeexpf, literacy, pop_incr, babymort, birth_rt, death_rt, log_gdp, b_to_d, fertility และ log_pop ใส่ใน box ของ Variable (s) (ตัวแปรทั้ง 11 ตัว เป็นตัวแปรชนิดตัวเลข)
- เลือก ☑ Save standardized values as variables
- จะได้ตัวแปรใหม่ 11 ตัวที่มีชื่อเดิมแต่มี Z นำหน้าต่อจากตัวแปรสุดท้ายในแฟ้มข้อมูล

ขั้นที่ 2 : การจำแนกกลุ่มด้วยเทคนิค K-Means โดยใช้คำสั่ง

Analyze ⇨ Classify ⇨ K-Means Clusters … จะได้หน้าจอรูปที่ 6.2

รูปที่ 6.2 : K-Means Clusters Analysis

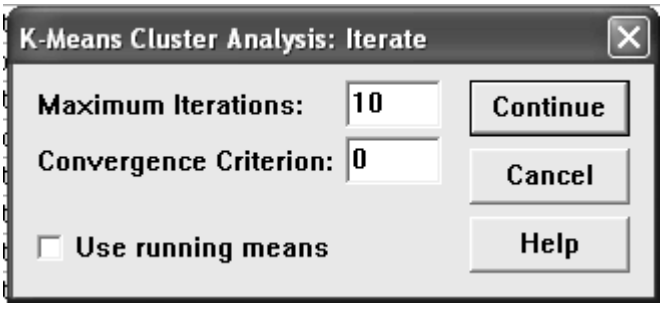


จากหน้าจอรูปที่ 6.2

- เลือกตัวแปร zurban, zlifeezp, zliterac, zpop_inc, zbabymor, zbirth_r, zdeath_r, zlog_gdp, zb_to_d, zfertilt และ zlog_pop ใส่ใน box ของ Variables
- เลือกตัวแปร county ซึ่งเป็นตัวแปรชนิด String ใส่ใน box ของ Label Cases by
- ในส่วนของ number of Clusters ใส่ 4 หมายถึงต้องการแบ่งประเทศออกเป็น 4 กลุ่ม
- ในส่วนของ Method เลือก ◉ Iterate and classify

Iterate... ในหน้าจอ 5024 จะใช้ได้ต่อเมื่อผู้วิเคราะห์เลือกวิธี ◉ Iterate and classify เท่านั้น คลิกปุ่ม Iterate... จะได้หน้าจอรูปที่ 6.3

รูปที่ 6.3 : Iterate



หน้าจอรูปที่ 6.3 ประกอบด้วย

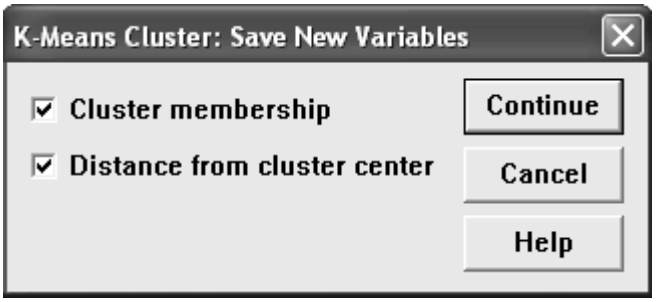
ส่วนที่ 1 : Maximum Iteration เป็นการกำหนดจำนวนรอบ (Iteration) ในการคำนวณ ซึ่งตัวเลขที่ใส่ใน box ต้องมีค่าตั้งแต่ 1 ถึง 999 โดยโปรแกรมจะคำนวณไม่เกินจำนวนรอบที่กำหนด

ส่วนที่ 2 : Convergence Criterion เป็นการกำหนดการหยุดการคำนวณ โดยการกำหนดสัดส่วนของระยะห่างที่สั้นที่สุด ระหว่างค่ากลางของ Cluster ในตอนเริ่มแรก โดยค่าที่กำหนดใน box จะต้องมากกว่า 0 แต่ไม่เกิน 1

ส่วนที่ 3 : ☑ Use running means ถ้าเลือกทางเลือกนี้หมายถึงจะให้หาค่ากลางของ Cluster ทุกครั้งที่มีการกำหนด Case ให้แก่ Cluster ถ้าไม่เลือกจะมีการคำนวณค่ากลางใหม่ต่อเมื่อได้กำหนด Cluster ให้แก่ทุก Case แล้ว

Save... ในหน้าจอรูปที่ 6.2 คลิกปุ่ม Save... จะได้หน้าจอรูปที่ 6.4

รูปที่ 6.4 : Save



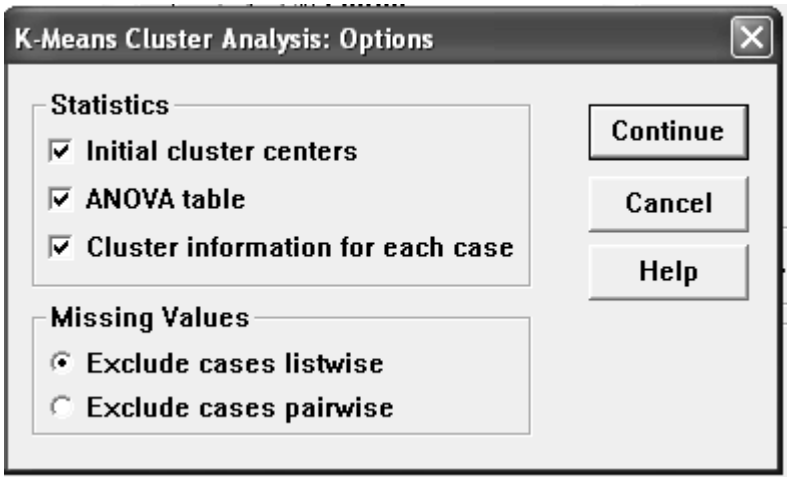
ในหน้าจอที่ 5.26 มีทางเลือก 2 ทางคือ

- Cluster membership จะสร้างค่าตัวแปรใหม่ซึ่งเป็นตัวแปรที่ระบุกลุ่มคือ Cluster ที่แต่ละ case เป็นสมาชิกอยู่
- Distance from cluster center จะสร้างตัวแปรใหม่ โดยตัวแปรใหม่นี้จะระบุค่า Euclidean distance จากแต่ละ case ไปยังค่ากลางของกลุ่ม

ในตัวอย่างนี้เลือกทั้ง 2 ส่วนคือ ☑ Cluster membership และ ☑ Distance from cluster center

Options... จากหน้าจอรูปที่ 6.2 คลิกปุ่ม Options... จะได้หน้าจอรูปที่ 6.5

รูปที่ 6.5 : Options



หน้าจอรูปที่ 6.5 ประกอบด้วย 2 ส่วนคือ

ส่วนที่ 1 : Statistics มี 3 ทางเลือกคือ

☐ Initial cluster centers เป็นการให้แสดงค่ากลางของแต่ละกลุ่มในตอนเริ่มแรก

☐ ANOVA Table ให้ค่าสถิติ F เพื่อแสดงความแตกต่างระหว่างกลุ่มของตัวแปรแต่ละตัวเมื่ออยู่ต่างกลุ่มกัน

☐ Cluster information for each case จะแสดงรายละเอียดของ Cluster ให้สำหรับแต่ละ Case

ในตัวอย่างนี้เลือกทั้ง 3 ทางเลือก

ส่วนที่ 2 : Missing Values มีทางเลือกสำหรับค่า Missing คือ

⦿ Exclude cases listwise

⦿ Exclude cases listwise

สำหรับความหมายได้อธิบายแล้วในหนังสือ “การใช้ SPSS for Windows ในการวิเคราะห์ข้อมูล” ในตัวอย่างนี้เลือก ⦿ Exclude cases listwise

ตารางที่ 6.1 : Initial Cluster

Initial Cluster Centers

	Cluster			
	1	2	3	4
zurban	-1.59	-1.26	1.63	1.80
zlifeexp	-2.47	-1.06	.74	.84
zliterac	-2.16	-1.15	-23	.42
zpop_inc	.93	.18	2.97	-.40
zbabymor	3.30	.96	-.78	-.96
zbirth_r	2.19	.25	.17	-.80
Zscore: Death rate per 1000 people	2.92547	.10408	-1.77684	-.83638
zlog_gdp	-1.79	-1.58	.66	1.22
zb_to_d	-37	-.14	5.08	-25
zfertilt	1.75	.48	.23	-.88
zlog_pop	.30	2.82	-1.31	-1.00

ค่าต่าง ๆในตารางที่ 6.1 แสดงค่าเฉลี่ยของตัวแปรแต่ละตัวที่ Standardized ใน Cluster ต่าง ๆ หรือถ้าเป็นค่ากลางของ Cluster ในตอนเริ่มต้นนั่นเอง ในที่นี้มี 4 กลุ่มหรือ 4 Clusters เนื่องจากได้กำหนดไว้ในหน้าจอรูปที่ 6.2

ตารางที่ 6.2 : Iteration History [a]

Iteration History [a]

	Change in Cluster Centers			
Iteration	1	2	3	4
1	1.958	2.544	2.072	1.846
2	.000	.442	.941	.457
3	.193	.271	.408	.187
4	.269	.363	.357	.219
5	.321	.336	.000	.114
6	.000	.084	.337	.000
7	.000	.184	.501	.000
8	.000	.182	.397	.044
9	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 9. The minimum distance between initial centers is 5.381.

ความหมายของผลลัพธ์ตารางที่ 6.2

เป็นการแสดงค่าเฉลี่ย หรือค่ากลางของแต่ละ Cluster ที่เปลี่ยนไปในแต่ละรอบของการคำนวณจะพบว่าในตัวอย่างนี้กำหนดให้มีจำนวนรอบสูงสุด = 10 รอบ แต่ในตารางแสดงแค่ 9 รอบ (Iteration) เนื่องจากในรอบที่ 9 ไม่มีการเปลี่ยนแปลงของค่ากลางเมื่อเทียบกับค่ากลางของรอบที่ 8 (ใน Iteration ที่ 9 ค่าที่เปลี่ยนไปของค่ากลางเป็นศูนย์หมด)

ตารางที่ 6.3 : Cluster Membership

Cluster Membership

Case Number	Cluster	Distance
1	1	2.468
2	4	1.573
3	4	2.051
4	4	1.091
5	4	1.037
6	2	1.725
7	3	2.053
8	1	2.345
9	4	2.806
10	4	.707
11	4	1.210
12	2	1.507
13	.	.
14	2	2.511
15	2	2.197
16	4	1.184
17	1	1.295
18	1	1.734

ความหมายของผลลัพธ์ตารางที่ 6.3

ตารางที่ 6.3 เป็นเพียงบางส่วนของข้อมูลทั้งหมด เนื่องจากเป็นตารางที่แสดงถึง Cluster ที่แต่ละ Case อยู่เช่น Case ที่ 1 คือ ประเทศ Afghanistan อยู่ใน Cluster ที่ 1 และมีระยะห่างจากค่ากลางของ Cluster ที่ 1 มากที่สุดคือ 2.468 เนื่องจากมีทั้งหมด 109 ประเทศ ตารางนี้แสดงเฉพาะประเทศที่ 1 – 18

ตารางที่ 6.4 : Final Cluster

Final Cluster Centers

	Cluster			
	1	2	3	4
zurban	-1.31	-.25	.57	.69
zlifeexp	-1.80	-.08	.22	.77
zliterac	-1.62	-.05	-.23	.80
zpop_inc	.91	.31	1.52	-.95
zbabymor	1.72	.17	-.18	-.80
zbirth_r	1.50	.18	.77	-.95
Zscore: Death rate per 1000 people	1.53829	-.55197	-1.08325	-.03111
zlog_gdp	-1.38	-.49	.24	.87
zb_to_d	-.13	.41	2.15	-.74
zfertilt	1.49	-.01	.86	-.88
zlog_pop	.04	.41	-.67	-.11

ความหมายของผลลัพธ์ตารางที่ 6.4

ค่าในตารางที่ 6.4 เป็นค่าเฉลี่ยตัวแปรที่ Standardized แล้ว ค่าเฉลี่ยเหล่านี้คือ ค่ากลางของแต่ละ Cluster จะพบว่าค่าเฉลี่ยของตัวแปร babymort จะแตกต่างกันเมื่ออยู่ Cluster ที่ต่างกัน และแตกต่างกันมากเมื่อเทียบกับตัวแปรอื่นๆ นั่นคือค่าเฉลี่ยของ babymort ใน Cluster ที่ 1=1.71845 หรือ มากกว่าค่าเฉลี่ยรวม 1.71845 เท่าของค่าเบี่ยงเบนมาตรฐาน ขณะที่ของ Cluster ที่ 4 เป็น -.80052 หรือน้อยกว่าค่าเฉลี่ยรวมถึง .8 เท่าของค่าเบี่ยงเบนมาตรฐาน ในทำนองเดียวกับตัวแปร lifeexp, birth_literac ก็มีค่าเฉลี่ยแตกต่างกันมากเมื่ออยู่ต่าง Cluster กัน

ตารางที่ 6.5 : Distances between Final Cluster Canters

Distances between Final Cluster Centers

Cluster	1	2	3	4
1		4.346	5.448	6.767
2	4.346		2.918	3.257
3	5.448	2.918		4.897
4	6.767	3.257	4.897	

ความหมายของผลลัพธ์ตารางที่ 6.5

ค่าในตารางที่ 6.5 เป็นระยะห่างระหว่างค่ากลางของทั้ง 4 Cluster จะพบว่า Cluster ที่ 1 มีระยะห่างจาก Cluster ที่ 4 มากที่สุด คือ 6.767 และใกล้ Cluster 2 มากที่สุด คือ 4.346 และ Cluster 3 ก็ใกล้ Cluster 2 มากที่สุดเช่นกัน

ตารางที่ 6.6 : ANOVA

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
zurban	20.186	3	.414	101	48.794	.000
zlifeexp	30.553	3	.146	101	208.801	.000
zliterac	27.062	3	.241	101	112.070	.000
zpop_inc	27.295	3	.211	101	129.369	.000
zbabymor	29.485	3	.170	101	173.760	.000
zbirth_r	30.477	3	.124	101	245.816	.000
Zscore: Death rate per 1000 people	22.843	3	.363	101	63.005	.000
zlog_gdp	26.541	3	.265	101	100.007	.000
zb_to_d	25.358	3	.244	101	103.766	.000
zfertilt	28.589	3	.168	101	170.097	.000
zlog_pop	3.362	3	.947	101	3.550	.017

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.

ความหมายของผลลัพธ์ตารางที่ 6.6 : ANOVA (1-Way ANOVA)

เป็นการแสดงค่า Mean Square ระหว่าง Cluster (Between – cluster Mean Square) และ Mean Square Error หรือ Within – Cluster Mean Square และให้ค่าสถิติ F โดยที่จะไม่ใช้ค่าสถิติ F และค่า Significance ใน Column สุดท้ายของตาราง ในการทดสอบค่าความแตกต่างระหว่างค่าเฉลี่ยของแต่ละตัวแปรเมื่ออยู่ต่าง Cluster กัน จะพบว่าค่าเฉลี่ยของตัวแปร birth_rt เมื่อมีต่างกลุ่มกันจะมีความแตกต่างกันมากที่สุด เนื่องจากค่าสถิติ F สูงสุด คือ 245.816 และของตัวแปร lifeexp รองลงมาคือ F = 208.801 ซึ่งอาจจะแตกต่างจากคำอธิบายของตารางที่ 6.4 เล็กน้อย เนื่องจากตารางที่ 6.4 เปรียบเทียบเฉพาะค่าเฉลี่ย ในตารางนี้ใช้ค่า Mean Square มาเปรียบเทียบกัน ส่วนตัวแปร log_pop มีค่าเฉลี่ยแตกต่างกันน้อยที่สุดเมื่ออยู่ต่าง Cluster กัน (F=3.55)

ตารางที่ 6.7 : Number of Cases in each Cluster

Number of Cases in each Cluster

Cluster	1	20.000
	2	31.000
	3	10.000
	4	44.000
Valid		105.000
Missing		4.000

ความหมายของผลลัพธ์ตารางที่ 6.7

จากตารางจะแสดงจำนวน Case หรือ ประเทศที่อยู่ในแต่ละ Cluster จะพบว่าประเทศส่วนใหญ่อยู่ใน Cluster ที่ 4 ส่วน Cluster ที่ 3 จะมีจำนวนประเทศน้อยที่สุด

การประเมินผลของการจำแนกกลุ่ม

เพื่อที่จะให้เข้าใจความหมายของกลุ่ม หรือ Cluster มากขึ้น จึงควรจะบันทึกเลขที่กลุ่มและระยะห่างจากแต่ละ Case ไปยังค่ากลางของกลุ่มที่ Case นั้นอยู่ (ในหน้าจอรูปที่ 6.4 ซึ่งหมายเลข Cluster ที่แต่ละ Case อยู่จะอยู่ในตัวแปรชื่อ qcl_1 และระยะห่างจากแต่ละ Case ไปยังค่ากลางของกลุ่มจะอยู่ในตัวแปรชื่อ qcl_2 ซึ่งอยู่ท้ายแฟ้มข้อมูล

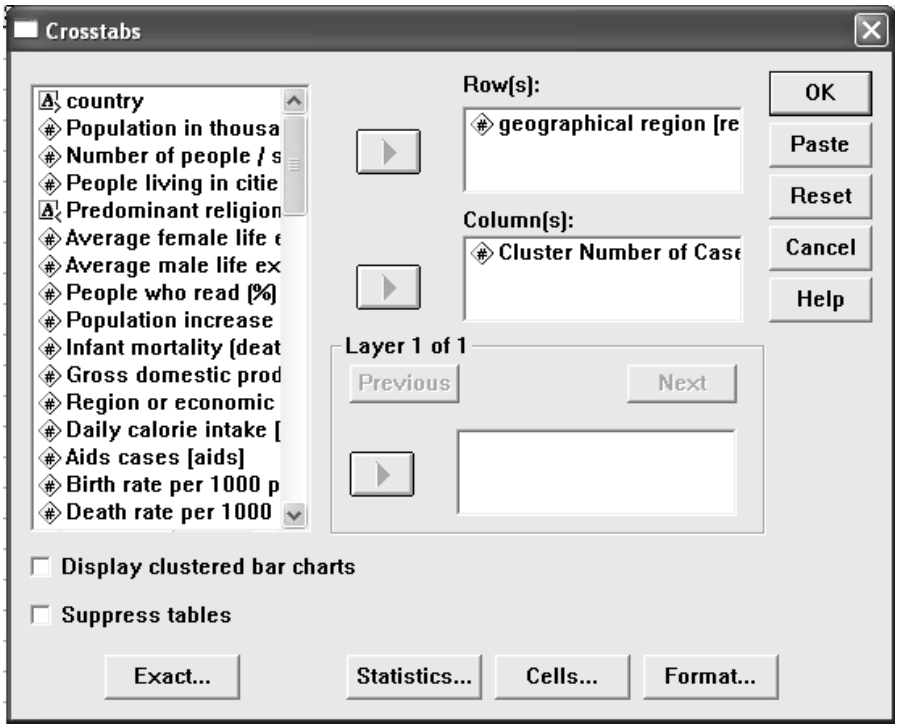
รูปที่ 6.6 : ตัวแปร QCL_1 และ QCl_2

	ZSco02	ZSco03	Zfertilty	ZSco04	QCL_1	QCL_2
1	-1.79026	-.37385	1.75399	.30228	1	2.46832
2	.17851	-.46178	-.40110	.63621	4	1.57348
3	.44699	.29639	-.19610	-.83437	4	2.05122
4	1.29784	-.62519	-.87417	.20852	4	1.09081
5	1.35941	-.99417	-1.08442	-.32243	4	1.03733
6	.08920	.03869	-.40110	-.37419	2	1.72474
7	.76515	1.90426	.20863	-2.04210	3	2.05299
8	-1.80059	-.01020	.59760	1.50253	1	2.34495
9	.67763	-.61118	-.93725	-2.60757	4	2.80606
10	.63075	-.95139	-.88468	-.15466	4	.70690
11	1.34074	-.99417	-.97930	-.16768	4	1.21016
12	-.90071	.27025	.34004	-.33078	2	1.50748
13	.11171	-.47651	.	-.68982	.	.
14	.00941	.37483	.80785	-1.49931	2	2.51082
15	-.08065	-.40950	-.45366	1.65216	2	2.19706

วิธีที่ 1 : ในที่นี้จะวิเคราะห์ตัวแปร qcl_1 โดยใช้คำสั่ง Crosstabs เพื่อแสดงจำนวนและ เปอร์เซ็นต์ของประเทศในทวีปต่าง ๆ ที่ถูกจัดอยู่ใน Cluster ต่างๆ โดยใช้คำสั่ง

Analyze ⇨ Descriptive Statistics ⇨ Crosstabs… จะได้หน้าจอรูปที่ 6.7

รูปที่ 6.7 : Crosstabs



- เลือกตัวแปร qcl_1 ใส่ใน box ของ Row
- เลือกตัวแปร region2 ใส่ใน box ของ Column

ตารางที่ 6.8 : geographical region * Cluster Number of Case Crosstabulation

geographical region * Cluster Number of Case Crosstabulation

Count

		Cluster Number of Case				
		1	2	3	4	Total
geographical region	Europe	0	0	0	17	17
	East Europe	0	0	0	12	12
	Pacific/Asia	4	8	0	6	18
	Africa	15	4	0	0	19
	Middle East	0	6	8	2	16
	Latn America	1	13	2	5	21
Total		20	31	10	42	103

ความหมายของผลลัพธ์ตารางที่ 6.8

ค่าในตารางที่ 6.8 แสดงจำนวนประเทศในแต่ละทวีปที่อยู่ใน Cluster 1-4 จะพบว่าทุกประเทศในยุโรปอยู่ใน Cluster ที่ 4 หมด (17 ประเทศ) และประเทศใน East Europe ก็อยู่ใน Cluster ที่ 4 ทั้งหมดเช่นกัน (12 ประเทศ) ขณะที่ประเทศใน Africa ส่วนใหญ่อยู่ใน Cluster ที่ 1 และประเทศใน Latin America ส่วนใหญ่อยู่ใน Cluster ที่ 2

ตารางที่ 6.9 : geographical region * Cluster Number of Case Crosstabulation

geographical region * Cluster Number of Case Crosstabulation

			Cluster Number of Case				
			1	2	3	4	Total
geographical region	Europe	Count	0	0	0	17	17
		% within geographical region	.0%	.0%	.0%	100.0%	100.0%
	East Europe	Count	0	0	0	12	12
		% within geographical region	.0%	.0%	.0%	100.0%	100.0%
	Pacific/Asia	Count	4	8	0	6	18
		% within geographical region	22.2%	44.4%	.0%	33.3%	100.0%
	Africa	Count	15	4	0	0	19
		% within geographical region	78.9%	21.1%	.0%	.0%	100.0%
	Middle East	Count	0	6	8	2	16
		% within geographical region	.0%	37.5%	50.0%	12.5%	100.0%
	Latn America	Count	1	13	2	5	21
		% within geographical region	4.8%	61.9%	9.5%	23.8%	100.0%
Total		Count	20	31	10	42	103
		% within geographical region	19.4%	30.1%	9.7%	40.8%	100.0%

ความหมายของผลลัพธ์ตารางที่ 6.9

ตารางที่ 6.9 ได้จากการใช้คำสั่ง Crosstabs แล้วคลิกปุ่ม Cells... เลือกเฉพาะ % of Row เป็นการแสดงเปอร์เซ็นต์ของประเทศในทวีปต่างๆ ที่อยู่ใน Cluster 1-4 โดยประเทศในยุโรป และ East Europe อยู่ใน Cluster ที่ 4 ถึง 100 % ในขณะที่ประเทศใน Asia อยู่ใน Cluster 2 เท่ากับ 44.4% ส่วนประเทศใน Africa อยู่ใน Cluster 1 ร้อยละ 78.9 เป็นต้น

สรุป

การที่ประเทศในยุโรปอยู่ใน Cluster ที่ 4 ถึง 100% และประเทศไทยในทวีปอื่นอยู่ใน Cluster ที่ 4 น้อย เนื่องจากประเทศในยุโรปมีค่าตัวแปรต่าง ๆ แตกต่างจากประเทศในทวีปอื่น ๆ ค่อนข้างมาก เมื่อ พิจารณาจากตารางที่ 6.4 : Final Cluster Center จะพบว่า ใน Cluster ที่ 4

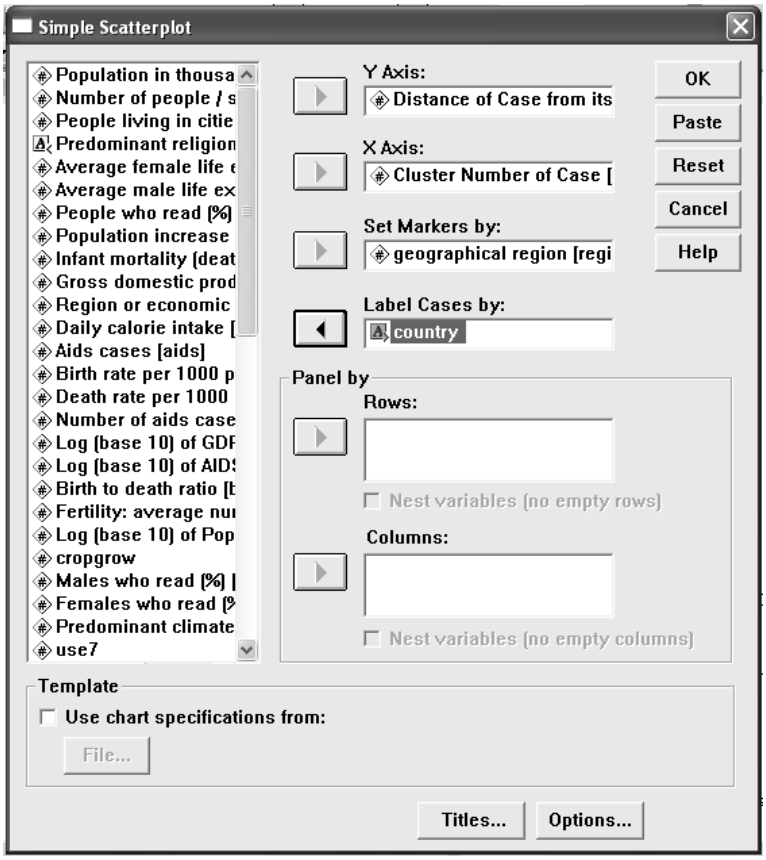
- ตัวแปร urban (สัดส่วนของประชากรที่อาศัยอยู่ในเมือง) มีค่าเฉลี่ยสูงกว่า Cluster อื่น ๆ หมายถึงประเทศที่อยู่ใน Cluster ที่ 4 จะเป็นประเทศที่ประชากรอาศัยในเมืองในสัดส่วนที่สูงกว่าประเทศที่อยู่ใน Cluster 1 – 3
- ตัวแปร Literacy (อัตราการอ่านหนังสือได้ของประชากร) ของ Cluster 4 มีค่าเฉลี่ยเป็นบวก (.800070) ขณะที่ของ Cluster 1-3 เป็นค่าลบ นั่นคือประเทศที่อยู่ใน Cluster ที่ 4 มีอัตราการอ่านหนังสืออกสูงกว่าอัตราเฉลี่ยรวม ในขณะที่อีก 3 Cluster ต่ำกว่าอัตราเฉ,ยรวม
- ตัวแปร pop_inc (อัตราการเพิ่มขึ้นของประชากร) ของ cluster 4 มีค่าเฉลี่ยเป็นลบ (-.94615) ขณะที่ของ Cluster 1 – 3 เป็นค่าบวก นั่นคือ ประเทศที่อยู่ใน Cluster ที่ 4 มี อัตราการเพิ่มขึ้นของประชากร ต่ำกว่าอัตราการเพิ่มขึ้นเฉลี่ยรวม ในขณะที่ของ Cluster 1 – 2 สูงกว่า
- ตัวแปร babymort อัตราการตายของทารก) ของ Cluster 4 มีค่าเฉลี่ยติดลบ = -.8 ขณะที่ของ Cluster 1 – 2 เป็นบวก และของ Cluster 3 เป็นลบ = -.1797 หมายความว่า ประเทศใน Cluster 4 มีอัตราการตายของทารกโดยเฉลี่ย ต่ำกว่าอัตราเฉลี่ยรวม
- ตัวแปร deth_rt (อัตราการตาย) และ birth_rt (อัตราการเกิด) ประเทศใน Cluster ที่ 4 มี อัตราต่ำกว่าประเทศใน Cluster 1 – 3
- ฯลฯ

วิธีที่ 2 : การวิเคราะห์โดยใช้กราฟ

เนื่องจากการวิเคราะห์โดยใช้ K-Mean Clustering ได้สร้างตัวแปรใหม่ 2 ตัว คือ qcl_1 และ qcl_2 จึงนำตัวแปรทั้งสองมาวิเคราะห์ด้วยกราฟ โดยใช้คำสั่ง

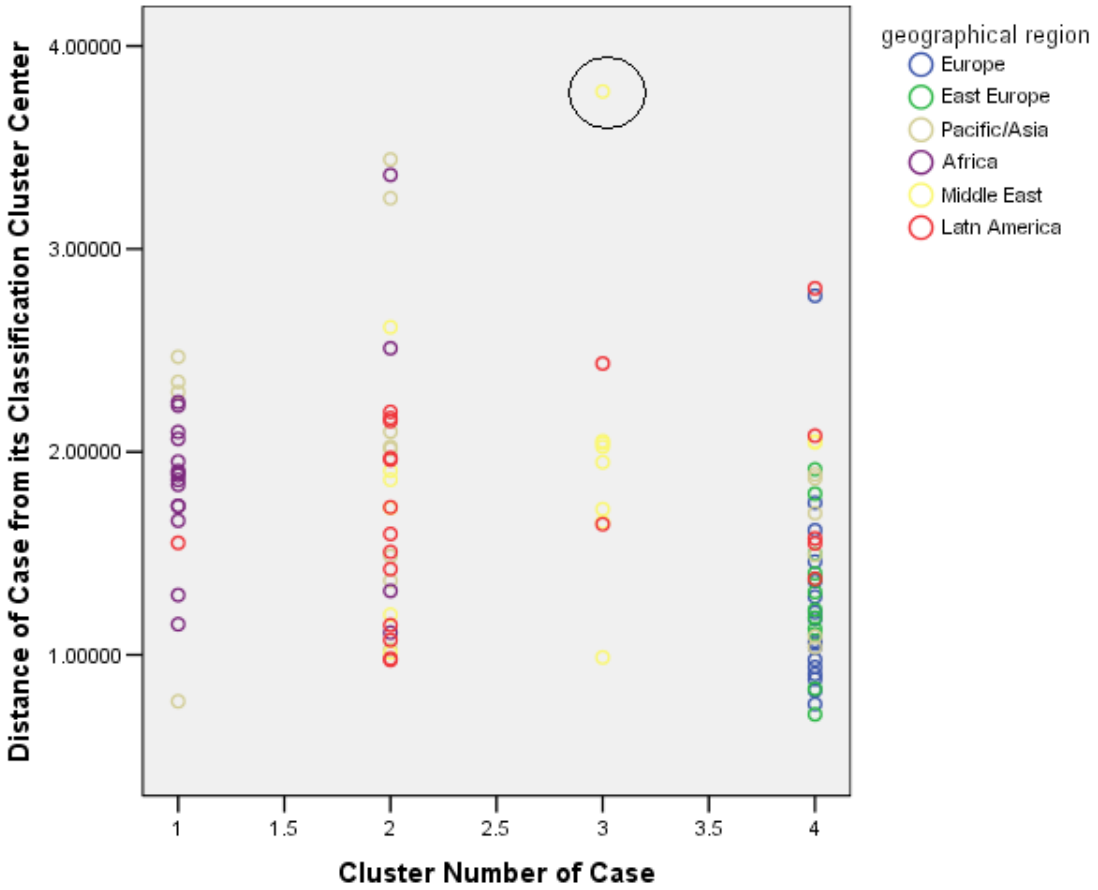
- Graphs ⇨ Scatter …
- เลือก Simple แล้วคลิกปุ่ม Define จะได้หน้าจอรูปที่ 6.8

รูปที่ 6.8 : Simple Scatter plot



- เลือกตัวแปร qcl_2 (ระยะห่างจาก Case ไปยังค่ากลางของ Cluster) ใส่ใน box ของ Y Axis
- เลือกตัวแปร qcl_1 (เลขที่ Cluster ที่ Case อยู่) ใส่ใน box ของ X Axis
- เลือกตัวแปร region 2 (ทวีป) ใส่ใน box ของ Set Markers by
- เลือกตัวแปร countery (ชื่อประเทศ) ใส่ใน box ของ Label cases by จะได้รูปที่ 6.9

รูปที่ 6.9



รูปที่ 5.31 แสดงประเทศในทวีปต่าง ๆ ที่อยู่ใน Cluster 1 – 4 โดยแกนตั้งแสดงระยะห่างของแต่ละ Case จากค่ากลางของ Cluster ที่ Case อยู่ จะพบว่าใน Cluster ที่ 3 มี 1 Case ที่ห่างจากค่ากลางมากแสดงว่าประเทศนี้ต่างประเทศอื่นใน Cluster เดียวกัน