แบบจำลองทำนายราคาขายเฉลี่ยต่อพื้นที่ของโครงการที่อยู่อาศัยด้วยเทคนิคการวิเคราะห์ถดถอยร่วมกับการจัดกลุ่มตามความหนาแน่น พื้นที่กรุงเทพมหานคร

วิทวัส แสงสว่าง

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิศวกรรมศาสตร์มหาบัณฑิต
สาขาวิชาวิศวกรรมข้อมูลขนาดใหญ่
วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์
มหาวิทยาลัยธุรกิจบัณฑิตย์
ปีการศึกษา 2565

PREDICTING AVERAGE RESIDENTIAL PRICE USING REGRESSION WITH DENSITY-BASED CLUSTERING TECHNIQUES: A BANGKOK USE CASE

WITTAWAT SANGSAWANG

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Engineering
Department of Big Data Engineering
College of Innovative Technology and Engineering,
Dhurakij Pundit University
Academic Year 2022

ใบรับรองวิทยานิพนธ์

วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์

วิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมข้อมูลขนาดใหญ่

หัวข้อวิทยานิพนธ์ แบบจำลองทำนายราคาขายเฉลี่ยต่อพื้นที่ของโครงการที่อยู่อาศัยด้วยเทคนิคการวิเคราะห์ถดถอยร่วมกับการจัดกลุ่มตามความหนาแน่นพื้นที่กรุงเทพมหานคร

เสนอโดย วิทวัส แสงสว่าง

สาขาวิชา วิศวกรรมข้อมูลขนาดใหญ่

อาจารย์ที่ปรึกษาวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ ดร.ดวงใจ จิตคงชื่น

ได้พิจารณาเห็นชอบโดยคณะกรรมการสอบวิทยานิพนธ์แล้ว

(ผู้ช่วยศาสตราจารย์ ดร.ณัฐพัชร์ อารีรัชกุลกานต์) ประธานกรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.ดวงใจ จิตคงชื่น) กรรมการที่ปรึกษาวิทยานิพนธ์

(ดร.ธนภัทร มังคะจิตร) กรรมการ

(ดร.เอกสิทธิ์ พัชรวงศ์ศักดา) กรรมการ

วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ รับรองแล้ว

(ดร.ชัยพร เขมะภาตะพันธ์) คณบดีวิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์

วันที่ 31 เดือน กรกฎาคม พ.ศ. 2566

หัวข้อวิทยานิพนธ์	แบบจำลองทำนายราคาขายเฉลี่ยต่อพื้นที่ของโครงการที่อยู่อาศัย ด้วยเทคนิคการวิเคราะห์ถดถอยร่วมกับการจัดกลุ่มตามความหนาแน่นพื้นที่กรุงเทพมหานคร
ชื่อผู้เขียน	วิทวัส แสงสว่าง
อาจารย์ที่ปรึกษาวิทยานิพนธ์	ผู้ช่วยศาสตราจารย์ ดร. ดวงใจ จิตคงชื่น
สาขาวิชา	วิศวกรรมข้อมูลขนาดใหญ่
ปีการศึกษา	2565

บทคัดย่อ

ปัจจุบันการกำหนดราคาขายเฉลี่ยต่อตารางเมตรของโครงการที่อยู่อาศัย ถือว่าเป็นความท้าทายของผู้ประกอบการอสังหาริมทรัพย์ ที่ผ่านมาการกำหนดราคาขายโครงการที่อยู่อาศัย มักจะแบ่งตามขอบเขตความเป็นเมือง ทว่าความเป็นสังคมเมืองและโครงสร้างพื้นฐานต่างๆ มีการพัฒนาอย่างต่อเนื่อง การพิจารณาขอบเขตของพื้นที่แบบเดิมๆ เพียงอย่างเดียว แล้วกำหนดราคาขาย จึงไม่สะท้อนสภาพตลาดของโครงการที่อยู่อาศัยอย่างแท้จริง จากปัญหาดังกล่าวที่เกิดขึ้น งานวิจัยนี้จึงนำอัลกอริทึมการจัดกลุ่มเชิงพื้นที่ตามความหนาแน่น DBSCAN และ HDBSCAN ทดสอบกับข้อมูลโครงการที่อยู่อาศัย จำนวน 2,010 โครงการทั่วกรุงเทพฯ แล้วทำการแปลประเมินด้วยดัชนีวัดคุณภาพกลุ่มจำนวน 4 แบบจำลองพบว่า HDBSCAN (haversine, eps = 2, minpt = 1) มีคุณภาพการจัดกลุ่มที่ดีที่สุด มีค่า Silhouette Coefficient และ DBCV เท่ากับ 0.19 และ 2.45 ตามลำดับ

จากนั้นเลือกกลุ่มที่มีคุณภาพดีตามดัชนีวัดคุณภาพ แล้วนำสร้างแบบจำลองการทำนายราคาเฉลี่ยต่อตารางเมตร พร้อมกับพิจารณาปัจจัยทางด้านโครงการและความเคลื่อนไหวของตลาดร่วมด้วยแล้วทดสอบผลการทำนายราคาเฉลี่ยของโครงการที่อยู่อาศัยด้วยวิธีรากที่สองของค่าเฉลี่ยความผิดพลาดกำลังสอง (RMSE) และค่าเฉลี่ยความผิดพลาดร้อยละสัมบูรณ์ (MAPE) เพื่อให้การกำหนดราคาขายเฉลี่ยต่อตารางเมตรของโครงการที่อยู่อาศัยในอนาคตมีความสมเหตุสมผลมากขึ้น

คำสำคัญ : การจัดกลุ่มตามความหนาแน่น, ดัชนีวัดคุณภาพการจัดกลุ่ม, กลุ่มเสถียร

.....

Thesis Title PREDICTING AVERAGE RESIDENTIAL PRICE USING REGRESSION WITH DENSITY-BASED CLUSTERING TECHNIQUES: A BANGKOK USE CASE

Author WITTAWAT SANGSAWANG

Thesis Advisor Asst.Prof. Duangjai Jitkongchuen, PhD.

Department Big Data Engineering

Academic Year 2022

ABSTRACT

Spatial clustering analysis plays such an important task to divide residential locations into clusters. It can cluster the locations better than traditionally or experiencedly allowed. Rather than just defined those locations according to predetermined areas for a purpose of administration, our study is to apply well-known density-based clustering algorithms i.e., DBSCAN and HDBSCAN on real-world dataset of 2,010 residential locations entire Bangkok metropolitan using 4 cluster validation techniques to quantify quality of resulted clusters. As a result, HDBSCAN (haversine distance, epsilon = 2, minimum point = 1) outperforms than the others with 0.19 silhouette coefficient and 2.45 DBCV.

In addition, we then formulate various regression models for average residential price prediction with each record contains general residential and market movement attributes. Regarding to model evaluation, we provide Root Mean Square Error (RMSE) and Mean Absolute Percent Error (MAPE) to quantify the models which support reasonable residential pricing decision in the future.

Keywords: Density-based clustering algorithm, Cluster validation, Stable cluster

.....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงด้วยการให้ความช่วยเหลือของผู้ช่วยศาสตราจารย์ ดร.ดวงใจ จิตคงชื่น ซึ่งเป็นอาจารย์ที่ปรึกษาวิทยานิพนธ์ กรุณาให้คำแนะนำและกำลังใจมาโดยตลอด ผู้เขียนกราบขอบพระคุณอาจารย์เป็นอย่างยิ่ง เมื่อครั้งอนุโลมให้ผู้เขียนนำเสนองานวิจัยแบบออนไลน์ ณ งานสัมมนาวิชาการแห่งหนึ่ง เนื่องจากผู้เขียนต้องเข้ารักษาตัวที่โรงพยาบาลอย่างกระทันหัน ต้องขอบพระคุณมา ณ ที่นี่ครับ

นอกจากนี้ ผู้เขียนขอกราบขอบพระคุณอาจารย์ ดร.วรพล พงษ์เพชร และ ดร. เอกสิทธิ์ พัชรวงศ์ศักดา อาจารย์ทั้งสองท่านเป็นแรงบันดาลใจทางด้านการใช้ชีวิตและการศึกษา มีความเป็นห่วงลูกศิษย์ตลอดเวลา รวมทั้งกราบขอบพระคุณอาจารย์ ดร.ธนภัทร ฆังคะจิตร ที่ติดตามนักศึกษาเพื่อมาสอบจบวิทยานิพนธ์จนถึงวินาทีสุดท้าย รวมทั้งให้คำแนะนำในการปรับปรุงวิทยานิพนธ์ให้สมบูรณ์ สุดท้ายขอขอบคุณนางสาวกุลธิดา รอดบุญ เป็นอย่างยิ่งที่ให้อำนวยความสะดวกและประสานเรื่องต่างๆ ตั้งแต่การลงทะเบียนเรียนจนถึงสอบจบวิทยานิพนธ์ ขอบคุณเพื่อนร่วมรุ่น Big Data Engineering รุ่นที่ 2 ทุกท่านที่ช่วยเหลือและให้กำลังใจตลอดการศึกษาครับ

วิทวัส แสงสว่าง

สารบัญ

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญภาพ

ภาพที่ **หน้า**

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ปัจจุบันการกำหนดราคาขายเฉลี่ยต่อตารางเมตรของโครงการที่อยู่อาศัย ถือว่าเป็นความท้าทายของผู้ประกอบการอสังหาริมทรัพย์ การกำหนดราคาขายโครงการที่อยู่อาศัย มักจะแบ่งตามขอบเขตความเป็นเมือง (Urban Zoning) โดยหน่วยงานภาครัฐหรือประสบการณ์ส่วนตัวของผู้เชี่ยวชาญวงการพัฒนาโครงการที่อยู่อาศัยเป็นหลัก ทว่าความเป็นสังคมเมืองและโครงสร้างพื้นฐานต่างๆ ซึ่งพัฒนาอย่างต่อเนื่องนั้นมีความซับซ้อนมากกว่าในอดีต การพิจารณาขอบเขตของพื้นที่แบบเดิมๆ จึงสร้างปัญหาทางอคติหรือความล้าสมัยต่อการกำหนดขอบเขตของพื้นที่เพื่อสะท้อนราคาขาย จนทำให้การตั้งราคาขายโครงการที่อยู่อาศัยไม่ได้สะท้อนสภาวะตลาดของแต่ละพื้นที่อย่างแท้จริง

ด้วยเหตุนี้ การจัดกลุ่มข้อมูล (Data Clustering) คือ เทคนิคการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) จึงเข้ามาช่วยพัฒนามุมมองและกลยุทธ์เกี่ยวกับการจัดกลุ่มข้อมูล สำหรับอัลกอริทึมการจัดกลุ่มตามความหนาแน่น (Density-based Clustering Algorithms) เริ่มมีบทบาทสำคัญกับงานด้านวิเคราะห์ข้อมูลเชิงพื้นที่ (Geospatial Analysis) เพราะพื้นฐานแนวคิดเพื่อค้นหาบริเวณที่มีข้อมูลเกาะกันอย่างหนาแน่นนั้น มีความสอดคล้องกับแนวคิดกฎข้อแรกทางภูมิศาสตร์ (The First Law of Geography) กล่าวคือ ตำแหน่งที่ตั้งของวัตถุที่อยู่ใกล้กันจะมีแนวโน้มความสัมพันธ์ทางด้านกายภาพหรือหน้าที่การทำงานคล้าย ๆ กันมากกว่าวัตถุที่ตั้งอยู่ห่างไกลออกไป ทว่าปัญหาการจัดกลุ่มข้อมูลด้วยอัลกอริทึมยังคงเกิดขึ้น เนื่องจากความหนาแน่นของข้อมูลแต่ละพื้นที่มีความคลุมเครือ มีความหนาแน่นที่ไม่แน่นอน (Arbitrary Shape) หรือไม่ได้แบ่งแยกกันอย่างเด็ดขาด ทำให้ส่งผลกระทบต่อความสามารถของอัลกอริทึมการจัดกลุ่มที่เลือกใช้ ปัญหาประการถัดมาคือ การเลือกดัชนีวัดคุณภาพการจัดกลุ่ม (Cluster Validation) เพื่อวัดคุณภาพการเกาะกลุ่มกันของกลุ่มนั้นยังมีการถกเถียงกันอยู่ เนื่องจากดัชนีวัดคุณภาพแต่ละเทคนิคถูกพัฒนาขึ้นบนพื้นฐานแนวคิดที่แตกต่างกัน การเลือกดัชนีวัดคุณภาพการจัดกลุ่มจึงจำเป็นต้องเหมาะสมกับลักษณะของข้อมูลที่กำลังจัดกลุ่มอยู่ รวมทั้งสอดคล้องกับวิธีการทำงานของอัลกอริทึมการจัดกลุ่มที่เลือกใช้ด้วย จึงทำให้การวัดคุณภาพการจัดกลุ่มนั้นมีความสมเหตุสมผล

นอกจากนี้การสร้างแบบจำลองเพื่อกำหนดราคาขายเฉลี่ยของโครงการที่อยู่อาศัย ยังมีความท้าทายเช่นกัน ดังจากการศึกษาที่ผ่านมา ส่วนมากจะนำข้อมูลเชิงพื้นที่มาสร้างแบบจำลองแบบภาพรวมตลาด มีการกำหนดขอบเขตอย่างเฉพาะเจาะจง เช่น เขตเมืองชั้นใน เขตเมืองชั้นกลาง หรือเขตเมืองชั้นนอก เป็นต้น ทำให้ผลการทำนายราคาขายของโครงการที่อยู่อาศัยยึดอยู่กับปัจจัยทางด้านทำเลที่ตั้งเพียงอย่างเดียว จนละเลยปัจจัยด้านลักษณะของโครงการและความเคลื่อนไหวของตลาดที่ส่งผลต่อราคาขายของโครงการที่อยู่อาศัยปัจจุบันและโครงการที่กำลังจะเปิดขายในอนาคต ด้วยเหตุนี้การจัดกลุ่มและการพิจารณาคุณภาพของ

กลุ่มข้อมูลโครงการที่อยู่อาศัยจึงเข้ามาเป็นปัจจัยพิจารณาร่วม เพื่อลดความแปรปรวนและสะท้อนสภาพความเป็นจริงของตลาดโครงการที่อยู่อาศัยมากขึ้น

ดังนั้นผู้วิจัยจึงนำเสนอกรอบแนวคิดการจัดกลุ่มกับข้อมูลเชิงพื้นที่ ด้วยอัลกอริทึมการจัดกลุ่มตามความหนาแน่นหลาย ๆ เทคนิค แล้วสร้างแบบจำลองการทำนายราคาเฉลี่ยต่อตารางเมตรของโครงการที่อยู่อาศัย

1.2 วัตถุประสงค์ของงานวิจัย

1.2.1 เพื่อจัดกลุ่มกับข้อมูลเชิงพื้นที่ของโครงการที่อยู่อาศัย ด้วยอัลกอริทึมการจัดกลุ่มตามความหนาแน่น แล้ววัดผลการจัดกลุ่มด้วยดัชชีเทคนิคต่างๆ เพื่อเปรียบเทียบคุณภาพการจัดกลุ่ม

1.2.2 เพื่อสร้างแบบจำลองการทำนายราคาเฉลี่ยของโครงการที่อยู่อาศัย

1.3 ประโยชน์ที่คาดว่าจะได้รับ

1.3.1 สามารถนำวิธีการที่นำเสนอมาช่วยวิเคราะห์การจัดกลุ่มข้อมูลเชิงพื้นที่อย่างเหมาะสม

1.3.2 สามารถนำวิธีการที่นำเสนอมาประยุกต์กับงานเฉพาะทางที่มีข้อมูลเชิงพื้นที่เข้ามาเกี่ยวข้อง เช่น การทำนายราคาเฉลี่ยโครงการที่อยู่อาศัยตามขอบเขตพื้นที่ (Housing Price Prediction)

1.4 ขอบเขตของงานวิจัย

1.4.1 ขอบเขตของข้อมูล

ข้อมูลสำหรับการวิเคราะห์มาจากแหล่งข้อมูลสารสนเทศของโครงการความร่วมมือเพื่อสร้างแผนที่สาธารณะ หรือ OpenStreetMap ผนวกกับฐานข้อมูลที่ผ่านเว็บเพจของผู้ให้บริการข้อมูลอสังหาริมทรัพย์ www.baania.com ซึ่งประกอบด้วยข้อมูลที่อยู่อาศัย บ้านเดี่ยว คอนโดมิเนียมและทาวเฮ้าท์ ตั้งแต่ปี 2019 – 2020 จำนวนกว่า 2000 โครงการ ทั้งโครงการที่เปิดขายใหม่และที่ยังคงขายอยู่

1.4.2 ขอบเขตของแผนงานการจัดกลุ่ม

งานวิจัยนี้มุ่งเน้นศึกษาอัลกอริทึมการจัดกลุ่มตามความหนาแน่นแบบ DBSCAN และ HDBSCAN ร่วมกับการพิจารณามาตรวัดระยะห่างแบบ Euclidean สำหรับวัดระยะห่างแนวระนาบ และ Haversine สำหรับวัดระยะห่างตามพื้นผิวโค้งของโลก งานวิจัยนี้เป็นการจัดกลุ่มข้อมูลโครงการที่อยู่อาศัยแบบไม่มีลาเบลเป้าหมายกำกับ ดังนั้นผู้วิจัยจึงเลือกดัชนีวัดคุณภาพการจัดกลุ่มบางเทคนิคที่เหมาะสมและไม่เหมาะสมกับข้อมูลเชิงพื้นที่ เพื่อเปรียบเทียบให้เห็นความแตกต่าง

1.4.3 ขอบเขตของแผนงานการทำนายราคาเฉลี่ย

หลังจากพิจารณาผลลัพธ์และเปรียบเทียบประสิทธิภาพการจัดกลุ่มโครงการที่อยู่อาศัย ผู้วิจัยจะนำสร้างแบบจำลอง เพื่อช่วยทำนายราคาขายเฉลี่ยของโครงการที่อยู่อาศัย

1.5 นิยามศัพท์

ข้อมูลเชิงพื้นที่ หมายถึง ข้อมูลเชิงพื้นที่ที่สามารถอ้างอิงพิกัดทางภูมิศาสตร์บนพื้นผิวโลก

การจัดกลุ่มตามความหนาแน่น หมายถึง อัลกอริทึมการจัดกลุ่มประเภทหนึ่ง เพื่อค้นหาบริเวณที่มีการรวมกลุ่มกันของข้อมูลที่มีความหนาแน่นมากกว่า เมื่อเปรียบเทียบกับพื้นที่รอบข้าง

ดัชนีวัดคุณภาพการจัดกลุ่ม หมายถึง เทคนิคการวัดประสิทธิภาพการจัดกลุ่ม เพื่อทดสอบคุณภาพของการจัดกลุ่ม แบ่งออกเป็น 3 ประเภท คือ ดัชนีวัดประสิทธิภาพการจัดกลุ่มแบบภายนอก ดัชนีวัดประสิทธิภาพการจัดกลุ่มแบบภายในและแบบสัมพัทธ์

มาตรวัดระยะห่าง หมายถึง วิธีคำนวณระยะห่างระหว่างข้อมูล เพื่อบ่งบอกความคล้ายกันและความแตกต่างของข้อมูล โดยข้อมูลที่มีความคล้ายกันหรือกลุ่มเดียวกัน จะมีระยะห่างใกล้กันมากกว่าข้อมูลที่มีอยู่คนละกลุ่ม

กลุ่มเสถียร หมายถึง กลุ่มที่มีความอัดแน่นกัน โดยสมาชิกของกลุ่มจะถูกจัดให้อยู่กลุ่มใดกลุ่มหนึ่งเพียงเท่านั้น ถึงแม้ว่าจะมีการเปลี่ยนแปลงมาตรวัดระยะห่างหรืออัลกอริทึมการจัดกลุ่ม สมาชิกของกลุ่มนั้นๆ ยังคงอยู่กลุ่มเดิมเสมอ

บทที่ 2
แนวคิด ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 แนวคิดและทฤษฎี

ผู้วิจัยทำการค้นคว้าความรู้จากฐานข้อมูลงานวิจัยต่าง ๆ และทางอินเตอร์เน็ต แล้วนำมาประมวลเป็นพื้นฐานความรู้สำหรับงานวิจัย ตามหัวข้อดังนี้

1. การจัดกลุ่มข้อมูล (Data Clustering)
2. วิธีวัดความคล้ายกันของข้อมูล (Similarity Measure)
3. กฎข้อแรกทางภูมิศาสตร์ (The First Law of Geography)
4. อัลกอริทึมการจัดกลุ่มตามความหนาแน่น (Density-based Clustering Algorithms)
5. ดัชนีวัดคุณภาพการจัดกลุ่ม (Cluster Validation)
6. การแปลผลการจัดกลุ่มด้วยสายตา (Visual Interpretation)
7. การวิเคราะห์ถดถอย (Regression Analysis)
8. การเปรียบเทียบความความคลาดเคลื่อน (Loss Function)

2.1.1 การจัดกลุ่มข้อมูล (Data Clustering)

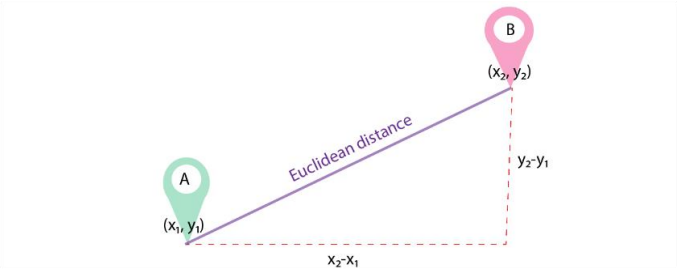
การจัดกลุ่มข้อมูลจัดเป็นการเรียนรู้ของเครื่อง (Machine Learning) อย่างหนึ่ง แบบไม่มีผู้สอน (Unsupervised Learning) กล่าวคือ เครื่องคอมพิวเตอร์สามารถเรียนรู้ข้อมูลโดยที่ไม่มีลาเบลเป้าหมายกำกับมาก่อน (Un-labelling) การทำงานของการเรียนรู้ประเภทนี้จะต้องนำชุดข้อมูลมาทำการจัดกลุ่มข้อมูล (Clustering) ด้วยการคำนวณหาความคล้ายกัน (Similarity Measure) ก่อน แล้วทำงานร่วมกับอัลกอริทึมการจัดกลุ่ม (Clustering Algorithm) เพื่อค้นหาความเป็นกลุ่มก้อนของข้อมูล (Cluster) จากความรู้ที่ซ่อนอยู่ของชุดข้อมูล (Intrinsic Pattern)

2.1.2 วิธีวัดความคล้ายกันของข้อมูล (Similarity Measure)

การวัดความคล้ายกันของข้อมูล ถือว่าปัจจัยสำคัญต่อการจัดกลุ่มข้อมูล เนื่องจากการประเมินความคล้ายกันจำเป็นต้องวัดความคล้ายกันผ่านการวัดระยะห่าง (Distance Function) ระหว่างข้อมูล วิธีวัดระยะห่างนั้นมีหลายเทคนิคด้วยกัน ถึงแม้แต่ละวิธีจะมีสูตรการคำนวณที่แตกต่างกัน ทว่าแนวคิดพื้นฐานของการวัดความคล้ายกันกลับมีความเหมือนกัน คือ ข้อมูลที่อยู่ใกล้กันหรือกลุ่มเดียวกันนั้น จะมีระยะห่างน้อยกว่าข้อมูลที่อยู่คนละกลุ่ม (Proximity)

มาตรวัดระยะห่างแบบ Euclidean (Euclidean Distance) เทคนิคการวัดระยะห่างประเภทนี้จัดเป็นมาตรวัดระยะห่างพื้นฐานระหว่างข้อมูล เพื่อคำนวณหาระยะทางที่สั้นที่สุด (Shortest Path) ระหว่างข้อมูล 2 จุด ณ แนวเส้นตรง บนพิกัดพื้นผิวราบ มีสูตรการคำนวณหาระยะห่างมาจากทฤษฎีบทปีทาโกรัส (Pythagorean Theorem) ดังนี้

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$



ภาพที่ 2.1 สมการคำนวณระยะห่างแบบ Euclidean

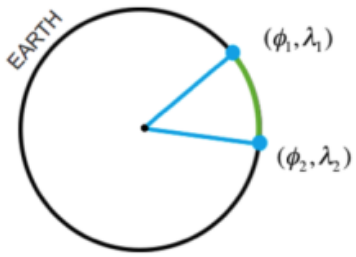
ที่มา: Role of Distance Metrics in Machine Learning, Analytic Vidhya

มาตรวัดระยะห่างแบบ Haversine (Haversine Distance)

มาตรวัดระยะห่างแบบ Haversine เป็นการวัดระยะห่างที่สั้นที่สุดระหว่างข้อมูลเช่นเดียวกับวิธีวัดระยะห่างแบบ Euclidean แต่จะคำนวณระยะห่างตามพิกัดรูปทรงรี (Spherical Coordinate System) หรือระยะห่างเชิงภูมิศาสตร์ (Geodesic Distance) เนื่องจากงานวิจัยนี้มีการกระจายตัวของข้อมูลเชิงพื้นที่ครอบคลุมบริเวณกว้าง วิธีวัดระยะห่างที่เลือกใช้จึงต้องคิดความโค้งของพื้นผิวโลกด้วย

นอกจากนี้เมื่อเลือกมาตรวัดระยะห่างประเภทนี้มีข้อควรระวัง คือ ก่อนคำนวณระยะห่างระหว่างข้อมูลเชิงพื้นที่ เราจะต้องเปลี่ยนหน่วยวัดของจุดพิกัดภูมิศาสตร์เป็นองศาก่อน โดยละติจูดจะเปลี่ยนเป็นมุมที่วัดระหว่างจุดใด ๆ กับเส้นศูนย์สูตร (มีค่าสูงสุด 90 องศา) ส่วนลองจิจูดจะเปลี่ยนเป็นมุมที่วัดระหว่างจุดใด ๆ กับเส้นเมอริเดียนที่ศูนย์ (มีค่าสูงสุด 180 องศา) หากจุดพิกัดทางภูมิศาสตร์ไม่ได้อยู่ในหน่วยวัดที่เหมาะสมอาจเป็นสาเหตุที่ทำให้ผลลัพธ์การจัดกลุ่มมีโอกาสผิดพลาดได้

$$d = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right)$$



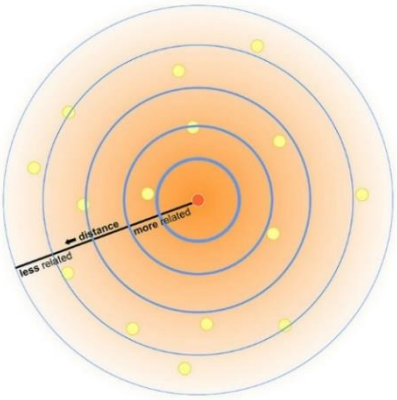
ภาพที่ 2.2 สมการคำนวณระยะห่างแบบ Haversine

ที่มา: https://github.com/DaniilSydorenko/haversine-geolocation/issues/1

2.1.3 กฎข้อแรกทางภูมิศาสตร์ (The First Law of Geography)

จากหัวข้อ 2.1.2 ซึ่งกล่าวถึงวิธีวัดความคล้ายกันของข้อมูลจากระยะห่างนั้นพบว่า มีแนวคิดสอดคล้องกับกฎข้อแรกทางภูมิศาสตร์ของ Tobler (1970) ว่าด้วยกฎข้อแรกของภูมิศาสตร์ (Tobler's first law of geography) กล่าวคือ "ทุกสิ่งทุกอย่างมีความสัมพันธ์กับสิ่งอื่น โดยสิ่งที่อยู่ใกล้กันจะสัมพันธ์กันมากกว่าสิ่งที่อยู่ไกลออกไป" ทำให้การอธิบายเชิงภูมิศาสตร์มีความเป็นวิทยาศาสตร์มากขึ้น มีการใช้สถิติ มีพิกัดตำแหน่งของสถานที่ นอกจากนี้ Phattharaphon (2015) ยังอธิบายการกระจายของสิ่งต่าง ๆ บนโลกออกเป็น 3 ลักษณะ คือ การกระจายแบบเกาะกลุ่ม (Clustered), การกระจายแบบไม่เป็นระเบียบ (Random) และการกระจายแบบกระจัดกระจาย (Dispersed)

ด้วยเหตุนี้ อัลกอริทึมการจัดกลุ่มตามความหนาแน่นแต่ละเทคนิค จึงเข้ามามีบทบาทสำคัญกับงานด้านวิเคราะห์ข้อมูลเชิงพื้นที่ (Geospatial Analysis) เพราะมีพื้นฐานแนวคิดเพื่อค้นหาบริเวณที่มีข้อมูลเกาะกันอย่างหนาแน่น เช่น การตรวจสอบการเจริญเติบโตของชุมชนเมือง ซึ่งมีความหนาแน่นของประชากรและชุมชนสูง ก่อนความเจริญของเมืองจะค่อย ๆ กระจายตัวออกสู่พื้นที่รอบนอกที่มีความหนาแน่นน้อยกว่า เป็นต้น

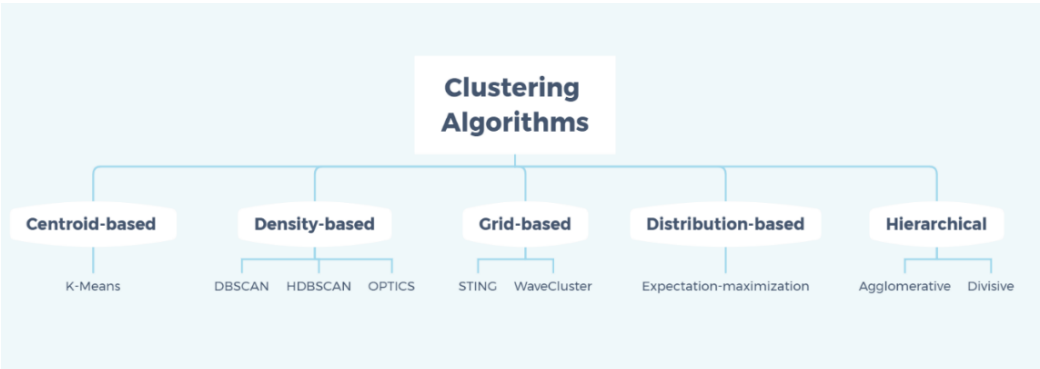


ภาพที่ 2.3 ภาพแสดงตำแหน่งที่ตั้งของวัตถุที่อยู่ใกล้กัน จะมีความสัมพันธ์ทางด้านกายภาพคล้ายกัน

2.1.4 อัลกอริทึมการจัดกลุ่มตามความหนาแน่น (Density-based Clustering Algorithms)

จากการศึกษาของ Xu & Tian (2015) แบ่งอัลกอริทึมการจัดกลุ่มออกเป็น 4 ประเภท ดังนี้ 1) อัลกอริทึมการจัดกลุ่มตามศูนย์กลาง (Centroid-based Clustering), 2) อัลกอริทึมการจัดกลุ่มตามความหนาแน่น (Density-based Clustering), 3) อัลกอริทึมการจัดกลุ่มตามการกระจาย (Distribution-based Clustering) และ 4) อัลกอริทึมการจัดกลุ่มตามลำดับขั้น (Hierarchical Clustering) ดังภาพที่ 2.4 งานวิจัยของ Halkidi et al. (2001) ยังจัดประเภทของอัลกอริทึมการจัดกลุ่มประเภทอื่นๆ คือ อัลกอริทึมการจัดกลุ่มตามตาราง (Grid-based Clustering) ซึ่งพัฒนาขึ้นเพื่อการวิเคราะห์ข้อมูลเชิงพื้นที่ (Spatial Analysis) สำหรับงานวิจัยทางด้านภูมิศาสตร์โดยเฉพาะ

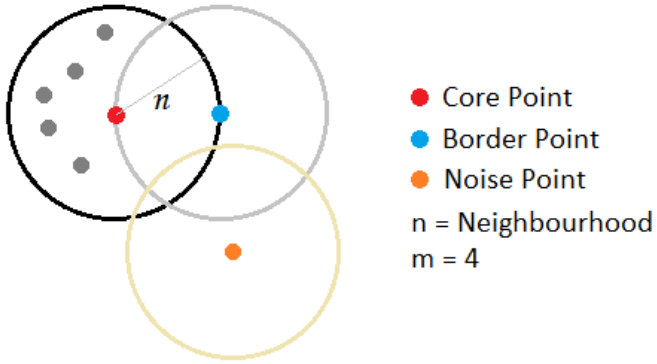
สำหรับงานวิจัยนี้จะมุ่งศึกษาด้วยอัลกอริทึมการจัดกลุ่มตามความหนาแน่น เนื่องจากมีความเหมาะสมกับข้อมูลเชิงพื้นที่และสอดคล้องกับกฎข้อแรกทางภูมิศาสตร์ โดยอัลกอริทึมการจัดกลุ่มตามความหนาแน่น เช่น เทคนิค DBSCAN และ HDBSCAN จะกล่าวถึงรายละเอียดถัดไป



ภาพที่ 2.4 แผนภาพอัลกอริทึมการจัดกลุ่มและตัวอย่างอัลกอริทึมแต่ละประเภท

DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

DBSCAN นำเสนอครั้งแรกโดย Ester et al. (1996) อัลกอริทึมนี้พัฒนามาเพื่อแก้ปัญหาข้อมูลรบกวน (Noise) ที่ถูกจัดกลุ่มรวมกับข้อมูลข้อมูลปกติ แนวคิดของ DBSCAN อ้างอิงจากการจัดกลุ่มตามความหนาแน่นของข้อมูล โดนการทำงานขั้นตอนแรกของอัลกอริทึม จะคำนวณระยะห่างระหว่างข้อมูล แล้วจึงทำการพิจารณาข้อมูลล้อมรอบทุกจุด ด้วยการกำหนดระยะห่าง (Epsilon) เป็นระยะห่างระหว่างข้อมูล (Search Distance) และนับจำนวนสมาชิกอย่างน้อยของกลุ่ม (Minimum Points หรือ minPts) ของแต่ละกลุ่ม จากนั้นจึงเริ่มจัดกลุ่ม แล้วระบุลาเบลเป้าหมายกำกับ



ภาพที่ 2.5 แผนภาพวิธีการทำงานของ DBSCAN

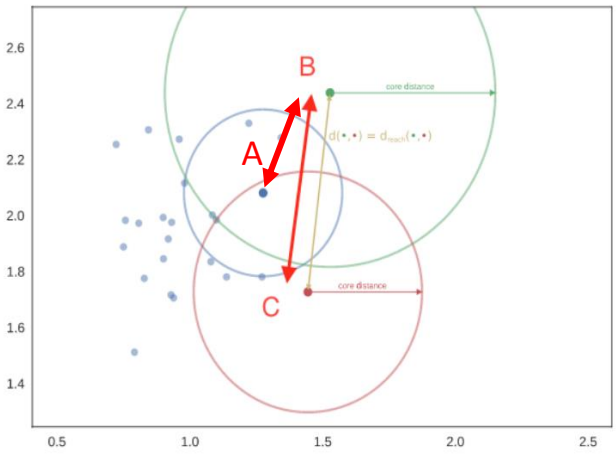
ที่มา: https://mineracaodedados.wordpress.com/2018/02/09/a-gentle-introduction-to-dbscan/

ข้อดีของ DBSCAN คือ อัลกอริทึมสามารถจัดกลุ่มข้อมูลตามความหนาแน่นได้ดี รวมถึงทนทานต่อข้อมูลรบกวน รวมถึงผู้ใช้งานหากมีความคุ้นเคยกับข้อมูลมากพอก็จะสามารถกำหนดค่าระยะห่างด้วยตัวเองและไม่จำเป็นต้องกำหนดจำนวนกลุ่มที่ต้องการแบ่งอีกด้วย อย่างไรก็ตาม การกำหนดพารามิเตอร์ค่าระยะห่างนั้นคงที่ หากข้อมูลชุดนั้นมีความหนาแน่นของข้อมูลในแต่ละบริเวณที่แตกต่างกัน (Varied Density) อาจจะทำให้การจัดกลุ่มข้อมูลด้วยวิธีนี้ไม่มีประสิทธิภาพเท่าที่ควร

HDBSCAN (Hierarchical of DBSCAN)

Campello et al. (2013) พยายามพัฒนาอัลกอริทึม DBSCAN แบบดั้งเดิมร่วมกับหลักการของการจัดกลุ่มแบบลำดับขั้น (Hierarchical Clustering) จนเกิดเป็น HDBSCAN ขึ้นมา ความพิเศษของ HDBSCAN นั้นสามารถเปลี่ยนค่าระยะห่างเองโดยอัติโนมัติ แล้ววาดแผนภาพต้นไม้ออกมา ซึ่งเป็นองค์ประกอบสำคัญของการจัดกลุ่มแบบแบบลำดับขั้น เพื่อบอกผลลัพธ์ของการจัดกลุ่มทีละขั้นตอน ผู้ใช้งาน

เพียงแค่กำหนดจำนวนสมาชิกอย่างน้อยของกลุ่มเท่านั้น ถึงแม้ว่า HDBSCAN จะไม่ได้พิจารณาพารามิเตอร์ของค่าระยะห่าง แต่เราสามารถกำหนดจำนวนกลุ่มที่ต้องการ (Self-adjusting) ด้วยการกำหนดค่าระยะห่างภายหลัง ขณะวาดแผนภาพต้นไม้ (Dendrogram) แล้วทำการประมวลผลอีกครั้ง เพื่อให้ได้ผลลัพธ์ของจำนวนกลุ่มตามต้องการ



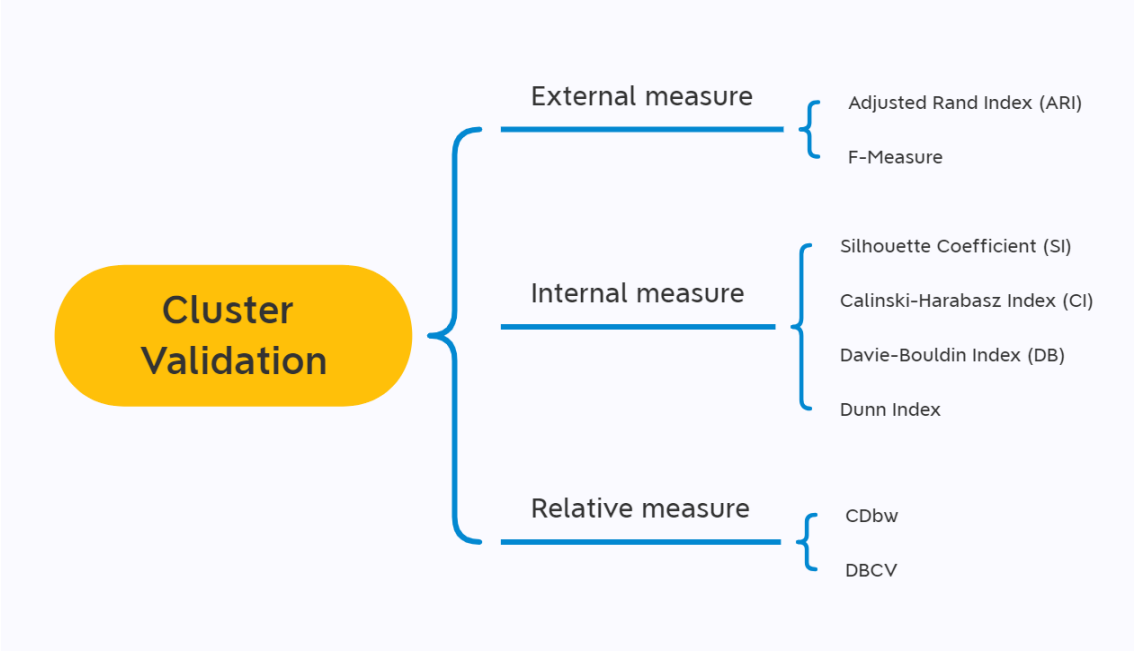
ภาพที่ 2.6 แผนภาพวิธีการทำงานของ HDBSCAN

ที่มา: https://hdbscan.readthedocs.io

2.1.5 ดัชนีวัดคุณภาพการจัดกลุ่ม (Cluster Validation)

การจัดกลุ่มข้อมูลซึ่งเป็นการเรียนรู้ของเครื่องแบบไม่มีผู้สอน การสร้างแบบจำลองการจัดกลุ่มจึงไม่มีการแบ่งแยกชุดข้อมูลออกเป็นชุดสำหรับฝึกสอน (Training set) และชุดสำหรับทดสอบ (Testing set) เหมือนกับวิธีการเรียนรู้ของเครื่องแบบมีผู้สอน (Supervised Learning) จึงกล่าวได้ว่า การวัดประสิทธิภาพของการจัดกลุ่มมักประเมินจากกฎเกณฑ์หรือประสบการณ์ส่วนตัวของผู้เชี่ยวชาญว่า กลุ่มนั้น ๆ สามารถตีความหรือนำมาประยุกต์อย่างไรให้เกิดประโยชน์สูงสุดกับงานด้านธุรกิจอย่างไร

ด้วยเหตุนี้ ดัชนีวัดคุณภาพการจัดกลุ่ม จึงเข้ามาทำหน้าที่ประเมินประสิทธิภาพของอัลกอริทึมการจัดกลุ่มและคุณภาพของกลุ่มเพื่อเป็นมาตราฐานเดียวกันและลดอคติการเลือกจัดกลุ่มข้อมูลอีกด้วย ดัชนีวัดประสิทธิภาพการจัดกลุ่ม แบ่งออกเป็น 3 ประเภท ดังนี้



ภาพที่ 2.7 แผนภาพสรุปดัชนีวัดประสิทธิภาพการจัดกลุ่มประเภทต่างๆ

ดัชนีวัดคุณภาพการจัดกลุ่มแบบภายนอก (External Cluster Validation)

เทคนิคการวัดคุณภาพนี้ เหมาะสำหรับข้อมูลที่มีลาเบลเป้าหมายที่กำกับมาก่อนแล้ว โดยนำลาเบลเป้าหมายมาเปรียบเทียบกับผลลัพธ์ของการจัดกลุ่ม กล่าวอีกนัยหนึ่ง คือ เราสามารถเลือกใช้ดัชนีวัดคุณภาพประเภทนี้กับชุดข้อมูลที่มาลาเบลเป้าหมาย คล้ายกับการทำนายประเภทข้อมูล (Classification) ตัวอย่างดัชนีที่นิยมใช้ เช่น Adjusted Rand Index และ F-Measure เป็นต้น อย่างไรก็ตาม งานวิจัยนี้ไม่ได้พิจารณาดัชนีวัดคุณภาพการจัดกลุ่มแบบภายนอก เนื่องจากข้อมูลที่นำมาวิเคราะห์ ไม่มีลาเบลเป้าหมายกำกับหรือคำตอบที่แท้จริงตั้งแต่แรก

ดัชนีวัดคุณภาพภาพการจัดกลุ่มแบบภายใน (Internal Cluster Validation)

หลักการพื้นฐานของดัชนีวัดคุณภาพการจัดกลุ่มแบบภายใน จะพิจารณาจากสัดส่วนของการอัดแน่น (Compaction) หรือระยะห่างของข้อมูลภายในกลุ่มเดียวกัน (Intra-cluster Distance) เทียบกับการแยกกัน (Separation) หรือระยะห่างระหว่างกลุ่ม (Inter-cluster Distance) ตัวอย่างดัชนี เช่น Silhouette Coefficient และ Calinski-Harabasz Index (CH) เป็นต้น

ด้วยเหตุนี้ ผู้วิจัยเลือกดัชนีวัดคุณภาพการจัดกลุ่มแบบภายในมาวัดคุณภาพ เนื่องจากมีความเหมาะสมกับข้อมูลที่ไม่มีลาเบลเป้าหมายกำกับมาตั้งแต่แรก โดยเลือก Silhouette Coefficient และ Calinski-Harabasz Index (CH) มาเป็นตัวแทนดัชนีวัดคุณภาพภาพการจัดกลุ่มแบบภายใน

ภาพที่ 2.8 ระยะห่างระหว่างข้อมูลภายในกลุ่มและระยะห่างระหว่างกลุ่ม

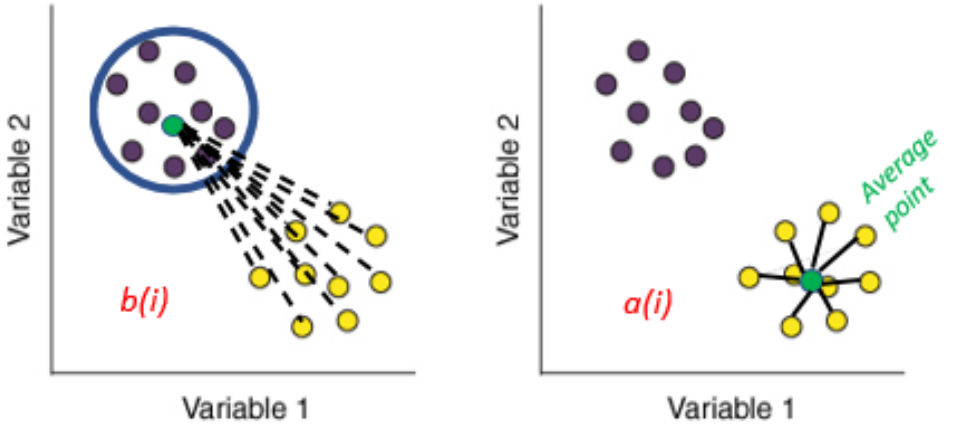
ที่มา: The Most Comprehensive Guide to K-Means Clustering You'll Ever Need, Analytic Vidhya

1) Silhouette Coefficient คือ ดัชนีพื้นฐานที่นิยมนำมาวัดคุณภาพการจัดกลุ่ม เพื่อวัดว่าข้อมูลแต่ละจุดนั้นมีความเหมือนกับกลุ่มที่ตัวเองอยู่มากหรือน้อย โดยประเมินจากระยะห่างระหว่างข้อมูลเป็นหลักและระยะห่างภายในกลุ่มเดียวกัน ตามสมการ โดยทั่วไปแล้ว Silhouette Coefficient จะมีค่าระหว่าง -1 และ 1 โดยกลุ่มที่มีคุณภาพจะมีค่าดัชนีเข้าใกล้ 1 กล่าวคือ การจัดกลุ่มจะมีระยะห่างระหว่างกลุ่มมากและระยะห่างภายในกลุ่มน้อย แต่หากดัชนีมีค่าเข้าใกล้ 0 แสดงว่า การจัดกลุ่มมีการทับซ้อนกัน อย่างไรก็ตามข้อจำกัดของ Silhouette Coefficient คือ มีความอ่อนไหวต่อข้อมูลรบกวน อาจทำให้ดัชนีวัดประสิทธิภาพการจัดกลุ่มผิดจากความเป็นจริง

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

โดยที่ b (i) หมายถึง ระยะห่างระหว่างกลุ่ม

a (i) หมายถึง ระยะห่างภายในกลุ่มเดียวกัน



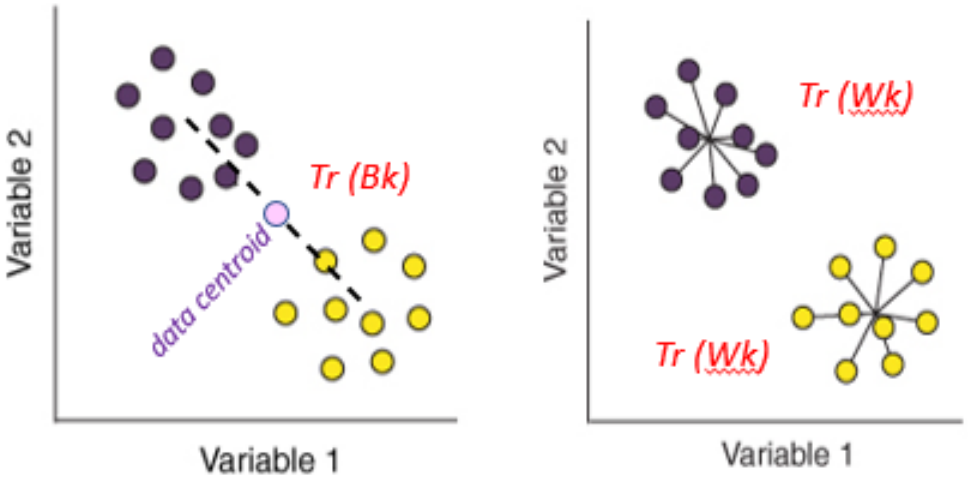
ภาพที่ 2.9 แผนภาพแสดงการทำงานของ Silhouette Coefficient

2) Calinski-Harabasz Index คือ อัตราส่วนของความแปรปรวนระหว่างกลุ่มกับความแปรปรวนภายในกลุ่ม เรียกอีกชื่อหนึ่งว่า Variance Ratio Criterion โดยระยะห่างระหว่างกลุ่ม คำนวณมาจากจุดศูนย์กลางของกลุ่ม (Cluster Centroid) ถึงจุดศูนย์กลางของข้อมูล (Data Centroid) ส่วนระยะห่างระหว่างกลุ่ม คำนวณจากระยะห่างของข้อมูลกับจุดศูนย์กลางของกลุ่ม ดังสมการ โดยทั่วไปแล้ว Calinski-Harabasz Index จะมีค่าอยู่ระหว่าง 0 จนถึงอนันต์ หากมีค่าดัชนีมาก แสดงว่าการจัดกลุ่มข้อมูลนั้นมีคุณภาพหรือไม่มีการซ้อนทับกันระหว่างกลุ่ม ดังนั้นกลุ่มที่แบ่งแยกกันชัดเจน ควรจะมี tr (Bk) หรือค่าความแปรปรวนระหว่างกลุ่มมากนั้นเอง

$$s = \frac{\mathrm{tr}(B_k)}{\mathrm{tr}(W_k)} \times \frac{n_E - k}{k - 1}$$

โดยที่ tr (Wk) ความแปรปรวนภายในกลุ่ม

tr (Bk) หมายถึง ความแปรปรวนระหว่างกลุ่ม



ภาพที่ 2.10 แผนภาพแสดงการทำงานของ Calinski-Harabasz Index (CH)

ดัชนีวัดคุณภาพการจัดกลุ่มแบบสัมพัทธ์ (Relative Cluster Validation)

ดัชนีวัดคุณภาพการจัดกลุ่มแบบสัมพัทธ์ มีแนวคิดการวัดคุณภาพสอดคล้องกับอัลกอริทึมการจัดกลุ่มตามความหนาแน่น เนื่องจากดัชนีประเภทนี้ วัดการอัดแน่นของข้อมูล บนพื้นฐานของความหนาแน่น (Density-Based Validation) เช่นเดียวกับอัลกอริทึมการจัดกลุ่มตามความหนาแน่น

1) CDbw (Composed Density between and within cluster) คือ ดัชนีแบบสัมพัทธ์เทคนิคแรก ๆ ที่ประยุกต์พื้นฐานของความหนาแน่น เพื่อนำมาวัดประสิทธิภาพของการจัดกลุ่ม หลักการพอสังเขปของเทคนิคนี้ คือ คำนวณหาการเกาะเกี่ยว (Cohesion) เพื่อเป็นตัวแทนความหนาแน่นภายในกลุ่มเดียวกัน รวมทั้งยังคำนวณหาการแยกกัน (Separation) เพื่อเป็นตัวแทนของระยะห่าง/ความหนาแน่นระหว่างกลุ่มด้วย ถ้าหาก CDbw มีค่าสูง นั้นแสดงว่า มีการจัดกลุ่มที่ดีและมีคุณภาพ อีกทั้งจุดเด่น CDbw นั้นยังสามารถจัดการข้อมูลที่มีการกระจายตัวไร้ระเบียบ อย่างไรก็ตาม ข้อมูลรบกวนยังคงเป็นอุปสรรคต่อการวัดประสิทธิภาพ

2) DBCV (Density-Based Cluster Validation) คือ ดัชนีที่พัฒนาขึ้นมาจาก CDbw มีหลักการพอสังเขป คือ ดัชนีจะคำนวณหาการกระจายตัวภายในกลุ่ม (Density Spareness of Cluster: DSC) และความหนาแน่นระหว่างกลุ่ม (Density Separation of a Pair of Clusters: DSPC) โดยหาก DSC มีค่ามาก แสดงว่า การกระจายตัวของข้อมูลนั้นมาก (ความหนาแน่นน้อย) จะทำให้ดัชนี DBCV มีค่าน้อย ซึ่งบ่งบอกว่าการจัดกลุ่มนั้นมีคุณภาพที่ไม่ดี นอกจากนี้ จุดเด่นของ DBCV นั้นสามารถจัดการกับชุดข้อมูลที่มีข้อมูลรบกวนและข้อมูลที่มีการกระจายตัวแบบไร้รูปแบบได้ ปกติแล้วค่า DBCV จะมีค่าอยู่ระหว่าง -1 และ 1 หากการจัด

กลุ่มมีการแบ่งแยกกันอย่างชัดเจน จะมีค่ามาทางบวก หากค่าดัชนีเข้าใกล้ 0 แสดงว่า มีการทับซ้อนกันของกลุ่มข้อมูล

$$V_C(C_i) = \frac{\min_{1 \le j \le l, j \ne i}(DSPC(C_i, C_j)) - DSC(C_i)}{\max(\min_{1 \le j \le l, j \ne i}(DSPC(C_i, C_j), DSC(C_i))}.$$

$$DBCV(C) = \sum_{i=1}^{i=l} \frac{|C_i|}{|O|} V_C(C_i)$$

โดยที่ | O | คือ จำนวนสมาชิกของข้อมูลทั้งหมด รวมถึงข้อมูลรบกวนด้วย

| C | คือ จำนวนกลุ่มที่แบ่งออกมา

ตารางที่ 2.1 การเปรียบเทียบความสามารถระหว่างดัชนีวัดคุณภาพการจัดกลุ่ม

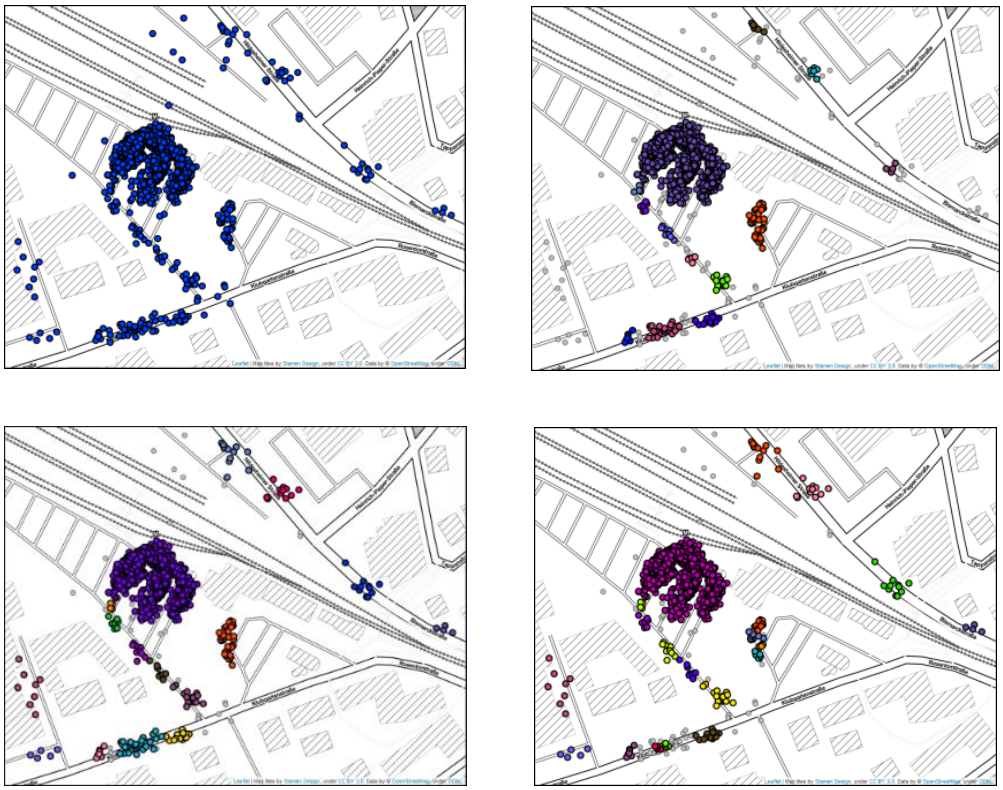
Cluster Validation	Support to Varied Density	Support to Arbitrary shape	Handling with Noises
Silhouette Coefficient	-	-	-
Calinski-Harabasz Index	-	-	-
CDbw	✓	✓	-
DBCV	✓	✓	✓

2.1.6 การแปลผลการจัดกลุ่มด้วยสายตา (Visual Interpretation)

การเลือกใช้ดัชนีวัดคุณภาพการจัดกลุ่มมักถูกนำมาใช้ประโยชน์ร่วมกับการแปลผลการจัดกลุ่มด้วยสายตาโดยทั่วไปแล้วพฤติกรรมของข้อมูลที่ถูกจัดกลุ่มแบ่งออกเป็น 2 ประเภท ดังนี้

ประเภทที่มีการแบ่งกลุ่มอย่างชัดเจน (Hard Cluster) เป็นการแบ่งข้อมูลออกจากกันเป็นกลุ่มๆ อย่างสิ้นเชิง โดยแต่ละข้อมูลนั้นจะถูกจัดให้อยู่ในกลุ่มใดกลุ่มหนึ่งเท่านั้น ประเภทของกลุ่มข้อมูลประเภทนี้จะมีความเสถียร (Stable Cluster) กล่าวคือ ไม่ว่าเราจะเลือกใช้อัลกอริทึมการจัดกลุ่มแบบใดก็ตามทำการจัดกลุ่ม จำนวนสมาชิกของกลุ่มมักจะเหมือนเดิมเสมอ ดังภาพที่ 2.11

ประเภทที่มีการแบ่งกลุ่มแบบไม่ชัดเจน (Soft Cluster) เป็นเทคนิคการแบ่งที่ข้อมูลสามารถอยู่ในหลายๆ กลุ่มได้ ณ การจัดกลุ่มแต่ละครั้ง โดยขึ้นอยู่กับความน่าจะเป็นของตัวข้อมูล การเลือกใช้อัลกอริทึมการจัดกลุ่ม การเลือกค่าพารามิเตอร์ที่แตกต่างกันออกไป เป็นต้น



ภาพที่ 2.11 กลุ่มเสถียรที่เกิดจากการแบ่งกลุ่มอย่างชัดเจน (สังเกตกลุ่มข้อมูลขนาดใหญ่มุมซ้ายบน)

ที่มา: https://hdbscan.readthedocs.io

2.1.7 การวิเคราะห์ถดถอย (Regression Analysis)

การวิเคราะห์ถดถอยเป็นวิธีการทางสถิติเพื่อศึกษาความสัมพันธ์ระหว่างตัวแปรตั้งแต่ 2 ตัวแปรขึ้นไป ประกอบด้วย ตัวแปรอิสระ (Independent Variable) ซึ่งสามารถนำมาประมาณค่าของตัวแปรอีกตัวหนึ่งที่เรียกว่า ตัวแปรทำนาย (Predictor หรือ Dependent Variable) สำหรับงานวิจัยนี้ เทคนิคการวิเคราะห์ถดถอยที่นำมาใช้ประเมินราคาขายเฉลี่ยมีเทคนิคต่างๆ ดังนี้

การวิเคราะห์ถดถอยเชิงเส้น (Linear Regression)

การวิเคราะห์ถดถอยเชิงเส้น ถือว่าเป็นเครื่องมือทางสถิติที่มีความซับซ้อนน้อยที่สุด แต่นิยมนำมาประยุกต์ใช้กันอย่างกว้างขวาง (Dependent Variable) เพื่อศึกษาขนาดหรือสัมประสิทธิ์ความสัมพันธ์

(Coefficient) ระหว่างตัวแปรอิสระ มีผลต่อตัวแปรทำนาย (Predictor หรือ Dependent Variable) รวมทั้งศึกษาทิศทางของความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรทำนายด้วยว่า ข้อมูลนั้นมีความสัมพันธ์ในรูปแบบเส้นตรง (Linearity) มากน้อยเพียงใด

เทคนิคนี้สามารถแบ่งออกเป็น 2 ประเภท ตามจำนวนจำนวนของตัวแปรอิสระ คือ 1) การวิเคราะห์ถดถอยเชิงเส้นแบบตัวแปรเดียว (Simple Linear Regression) และ 2) การวิเคราะห์ถดถอยเชิงเส้นแบบหลายตัวแปร (Multiple Linear Regression) ซึ่งการศึกษานี้เลือกใช้การวิเคราะห์ถดถอยเชิงเส้นแบบหลายตัวแปร เนื่องจากปัจจัยต่างๆ นั้นมีผลต่อการตั้งราคาขายเฉลี่ยของโครงการที่อยู่อาศัย เช่น ด้านทะเลที่ตั้งและราคาขายตลาด ณ ปัจจุบัน เป็นต้น

$$Y \approx \beta_0 + \beta_1 X$$

โดยที่ Y คือ ตัวแปรตาม

B0 คือ ค่าจุดตัดแกน ณ แกน Y

B1 คือ ค่าสัมประสิทธิ์ของตัวแปรอิสระ

การวิเคราะห์ถดถอยแลซโซ (Lasso Regression)

การวิเคราะห์ถดถอยแลซโซมีจุดเด่นคือ การประมาณค่าและการคัดเลือกตัวแปรเข้าสู่แบบจำลองการวิเคราะห์ในเวลาเดียวกัน กล่าวคือ ค่าสัมประสิทธิ์การถดถอยพหุคูณของวิธีแลซโซจะอยู่ในรูปผลบวกระหว่างผลรวมความคลาดเคลื่อนกำลังสองและผลรวมสัมบูรณ์ของค่าสัมประสิทธิ์ถ่วงน้ำหนักให้มีค่าต่ำสุด หากค่าถ่วงน้ำหนักมากจะทำให้สัมประสิทธิ์นั้นเหลือเท่ากับค่าเท่ากับศูนย์หรือหายไป โดยส่วนใหญ่แล้ว ค่าสัมประสิทธิ์มีค่าเป็นศูนย์และค่าสัมประสิทธิ์บางส่วนไม่เท่ากับศูนย์

ข้อดีของการวิเคราะห์ถดถอยแลซโซจะช่วยลดปัญหาอิทธิพลของกลุ่มตัวแปรอิสระที่มีความสัมพันธ์เชิงเส้นต่อกัน (Multicollinearity) โดยจะคัดเลือกเฉพาะตัวแปรอิสระเพียงแค่ตัวแปรเดียว แล้วนำมาสร้างแบบจำลอง ทำให้ความผิดพลาดของการทำนายผลลดลง อย่างไรก็ตาม ข้อจำกัดของวิธีนี้คือ การคัดเลือกเฉพาะตัวแปรอิสระเพียงแค่ตัวแปรเดียวดังกล่าวนั้น มักจะเลือกตัวแปรใดตัวแปรหนึ่งก็ได้จากกลุ่มตัวแปรอิสระที่มีความสัมพันธ์เชิงเส้นต่อกันสูง ดังนั้นการวิเคราะห์ถดถอยแลซโซจะมีประสิทธิภาพสูง เมื่อตัวแปรอิสระมีความสัมพันธ์เชิงเส้นต่อกันไม่มากนัก

การวิเคราะห์ถดถอยอิลาสติคเน็ต (Elastic Net Regression)

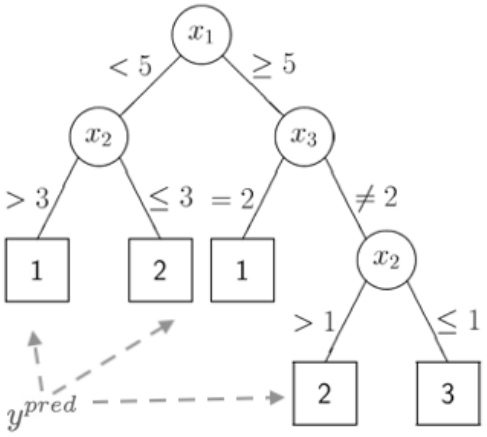
การวิเคราะห์ถดถอยอิลาสติคเน็ตเป็นวิธีที่สามารถเลือกตัวแปรและประมาณค่าได้ในคราว คล้ายกับการวิเคราะห์ถดถอยแลซโซ ถูกพัฒนาขึ้นเพื่อแก้ไขข้อจำกัดของการวิเคราะห์ถดถอยแลซโซ ด้วยการผนวกระหว่างเทคนิคการวิเคราะห์ถดถอยริดจ์กับแลซโซ โดยค่าสัมประสิทธิ์การถดถอยพหุคูณของวิธีอิลาสติคเน็ตจะอยู่ในรูปผลบวกของผลรวมความคลาดเคลื่อนกำลังสองผลรวมสัมบูรณ์ของค่าสัมประสิทธิ์ถ่วงน้ำหนักและผลรวมกำลังสองของค่าสัมประสิทธิ์ถ่วงน้ำหนักให้มีค่าต่ำสุด ทำให้เทคนิคนี้มีความเหมาะสมต่อการวิเคราะห์ข้อมูลที่มีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่างเป็นจำนวนมากและกรณีที่ตัวแปรอิสระมีความสัมพันธ์ต่อกันสูง

ต้นไม้ตัดสินใจแบบถดถอย (Regression Tree)

ต้นไม้ตัดสินใจถือว่าเป็นแบบจำลองทางคณิตศาสตร์ที่มีการสร้างแบบจำลองแบบเงื่อนไข (Rule-based Model) คล้ายกับการตัดสินใจของมนุษย์ เพื่อแสดงลำดับขั้นตอนของการตัดสินใจ ประกอบด้วยโหนด (node) เป็นจุดแสดงลักษณะข้อมูล ส่วนกิ่ง (Branch) ทำหน้าที่เป็นทางเลือกการตัดสินใจจากการพิจารณาโหนด จนนำมาสู่ใบ (Leaf) ที่แสดงคำตอบของการตัดสินใจ

รายละเอียดของการสร้างต้นไม้ตัดสินใจจะเริ่มต้นนำตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรทำนายมากที่สุด ขึ้นมาเป็นโหนดเริ่มต้น (Root Node) ของต้นไม้ตัดสินใจ หากกล่าวทางเทคนิคคือ ต้นไม้ตัดสินใจจะเลือกตัวแปรอิสระที่มีค่าความคลาดเคลื่อน (Residual) น้อยที่สุดในการหาจุดที่ดีที่สุดในการแบ่งข้อมูล (Split point) ก่อน หลังจากนั้นแบบจำลองจะทำแบบนี้ซ้ำเรื่อยๆ (Recursive Iteration) จนได้ต้นไม้ตัดสินใจที่อธิบายตัวแปรทำนายได้ชัดเจนที่สุดหรือมีค่าผลรวมความคลาดเคลื่อนกำลังสอง (Residual Sum of Square: RSS) ของทั้งแบบจำลองน้อยที่สุดนั่นเอง

ต้นไม้ตัดสินใจแบ่งออกเป็น 2 ประเภท ตามชนิดของตัวแปรทำนาย ได้แก่ 1)ต้นไม้ตัดสินใจแบบถดถอย (Regression Tree) สำหรับตัวแปรทำนายเป็นตัวเลข และ 2)ต้นไม้ตัดสินใจแบบจำแนก (Classification Tree) สำหรับปัญหาที่มีตัวแปรทำนายแบบลาเบล



ภาพที่ 2.12 แผนภาพแสดงการตัดสินใจของต้นไม้แบบที่มีตัวแปรตามเป็นตัวเลข

ที่มา: bookdown.org/tpinto_home/Beyond-Additivity/regression-and-classification-trees.html

การวิเคราะห์ถดถอยการสุ่มของป่า (Random Forest Regression)

การสุ่มของป่าเป็นเทคนิคที่พัฒนาต่อยอดมาจากต้นไม้ตัดสินใจ โดยการสุ่มของป่าประกอบด้วยต้นไม้ตัดสินใจหลาย ๆ ต้น แต่ละต้นเป็นอิสระต่อกัน จึงทำให้ประสิทธิภาพของการพยากรณ์สูงขึ้น หลักการทำงานของการสุ่มของป่าจะเริ่มด้วยต้นไม้ตัดสินใจหลาย ๆ ต้นจะถูกสร้างขึ้นมา โดยแต่ละต้นจะได้รับตัวแปรอิสระและข้อมูลเพียงบางส่วน (Subset) แบบสุ่ม (Random) จากจำนวนตัวแปรอิสระและจำนวนข้อมูลทั้งหมดของข้อมูลทดลอง เพื่อทำให้ได้ต้นไม้ตัดสินใจที่มีความหลากหลายและเป็นเอกเทศต่อกัน ลดปัญหาตัวแปรอิสระมีความสัมพันธ์เชิงเส้น หลังจากนั้นต้นไม้ตัดสินใจจะเติบโตเพิ่มจำนวนโหนดและกิ่งมากขึ้นเรื่อย ๆ จนได้ค่าพยากรณ์จากต้นไม้แต่ละต้น แล้วทำการสรุปออกมาเป็นภาพรวมพยากรณ์ของการสุ่มของป่า

หากเป็นผลลัพธ์ของปัญหาการวิเคราะห์แบบถดถอย เช่น การหาราคาเฉลี่ยของโครงการที่อยู่อาศัย ค่าพยากรณ์สุดท้ายจะมาจากการหาค่าเฉลี่ย (Mean) ของตัวเลขผลการพยากรณ์ กล่าวคือ การสุ่มของป่าจะนำค่าพยากรณ์ของทุกต้นไม้ตัดสินใจ มาคำนวณหาค่าเฉลี่ย ส่วนหากผลลัพธ์ของปัญหาการวิเคราะห์แบบจำแนก การสุ่มของป่าจะใช้วิธีผลโหวต (Majority vote) กล่าวคือ ค่าพยากรณ์ของต้นไม้ตัดสินใจหลาย ๆ ต้นที่ได้รับค่าผลโหวตมากที่สุด จะถูกเลือกให้เป็นคำตอบ

การวิเคราะห์ถดถอยเอ็กซ์จีบูสท์ (Extreme Gradient Boosting: XGB)

การวิเคราะห์ถดถอยเอ็กซ์จีบูสท์เป็นอีกเทคนิคที่พัฒนาต่อยอดมาจากต้นไม้ตัดสินใจ ทั้งการสุ่มของป่าและการวิเคราะห์ถดถอยเอ็กซ์จีบูสท์ถือว่าเป็นการเรียนรู้แบบกลุ่ม (Ensemble Model) กล่าวคือ มีการสร้างแบบจำลองหลาย ๆ แบบเพื่อช่วยกันตัดสินใจ อย่างไรก็ตาม การวิเคราะห์ถดถอยเอ็กซ์จีบูสท์นั้นจะ

สร้างต้นไม้ตัดสินใจที่ไม่เป็นเอกเทศต่อกัน แต่จะเรียนรู้ข้อผิดพลาดจากต้นไม้ก่อนหน้า เพื่อนำมาปรับปรุงการสร้างต้นไม้ต้นใหม่อย่างต่อเนื่อง ทำให้แบบจำลองมีความแม่นยำมากขึ้น โดยแบบจำลองจะหยุดสร้างตัดไม้ต้นใหม่ เมื่อเรียนรู้ข้อผิดพลาดจากต้นไม้ก่อนหน้าจนครบหมดแล้ว

2.1.8 การเปรียบเทียบค่าความคลาดเคลื่อน (Loss Function)

หลังจากที่นำแบบจำลองต่าง ๆ มาอธิบายความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรทำนายแล้ว ปกติการทำนายผลนั้นไม่ได้มีความถูกต้องแม่นยำสมบูรณ์ กล่าวคือ แบบจำลองมีความคลาดเคลื่อนเกิดขึ้นอยู่ ดังนั้นการหาวิธีประเมินความแม่นยำหรือค่าความคลาดเคลื่อนจึงมีความสำคัญ เพื่อพิจารณาว่าแบบจำลองแบบใดมีความเหมาะสมที่สุดในการอธิบายความสัมพันธ์ การศึกษานี้เลือกค่า RMSE และ MAPE มาประเมินประสิทธิภาพของแบบจำลอง โดยค่าความคลาดเคลื่อนน้อยจะให้ค่าพยากรณ์ใกล้เคียงกับความเป็นจริงมากที่สุด มีรายละเอียดดังนี้

รากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Square Error: RMSE) วิธีนี้เป็นวิธีการวัดค่าความคลาดเคลื่อนแบบมาตรฐานที่นิยมใช้กันอย่างแพร่หลาย โดยค่าความคลาดเคลื่อนจะมีค่าบวกเสมอด้วยการยกกำลังสอง แล้วจึงนำค่าความคลาดเคลื่อนมารวมกันเพื่อหาค่าเฉลี่ยความคลาดเคลื่อนของแบบจำลอง โดย RMSE มีค่าน้อยจะบ่งบอกว่า แบบจำลองมีความแม่นยำ หน่วยของความคลาดเคลื่อนยังแปลความง่าย เนื่องจากมีหน่วยวัดเดียวกับตัวแปรทำนายอีกด้วย

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

โดยที่ y หมายถึง ค่าพยากรณ์ตัวแปรตาม

y (i) หมายถึง ค่าจริงของตัวแปรตาม

n หมายถึง จำนวนข้อมูลทั้งหมด

ค่าเฉลี่ยของร้อยละความผิดพลาดสัมบูรณ์ (Mean Absolute Percent Error: MAPE) วิธีนี้เป็นวิธีการวัดค่าความคลาดเคลื่อนที่นิยมใช้วัดความคลาดเคลื่อนเช่นเดียวกับ RMSE แต่มีวิธีการคำนวณค่าความคลาดเคลื่อนแตกต่างกันคือ MAPE คำนวณจากค่าเฉลี่ยของร้อยละความคลาดเคลื่อนระหว่างค่าจริงและค่าพยากรณ์ ทำให้ค่า MAPE นั้นมีความอ่อนไหวต่อข้อมูลรบกวนน้อยและแสดงค่าความคลาดเคลื่อนแบบร้อยละโดยไม่ต้องคำนึงถึงขนาดของตัวแปรที่ทำนายอยู่

2.2 งานวิจัยที่เกี่ยวข้อง

ปัจจุบันการพิจารณาขอบเขตของพื้นที่มักจะแบ่งตามขอบเขตความเป็นเมือง (Urban Zoning) จากหน่วยงานภาครัฐหรือประสบการณ์ส่วนตัวของผู้เชี่ยวชาญ อาจทำให้เกิดปัญหาทางอคติต่อการกำหนดขอบเขตของพื้นที่ จึงมีงานวิจัยหลาย ๆ ฉบับที่พยายามนำอัลกอริทึมการจัดกลุ่มมาผนวกกับงานทางด้านภูมิศาสตร์ เพื่อแบ่งขอบเขตของพื้นที่ ตามสถานที่สำคัญต่าง ๆ เช่น POI หรือตามเส้นทางการจราจรและขนส่งสาธารณะ

Boeing (2018) ศึกษาและเปรียบเทียบการจัดกลุ่มระหว่างอัลกอริทึม K-Means และ DBSCAN กับข้อมูลระบุตำแหน่ง GPS เนื่องจากข้อมูลมีการกระจายตัวครอบคลุมอย่างกว้างขวางและได้รับอิทธิพลจากความโค้งของพื้นผิวโลก ดังนั้นการวัดความคล้ายกันของข้อมูล จึงเลือกวิธีวัดระยะห่างแบบ Haversine มาทดสอบงานวิจัย จากผลการทดลองพบว่า อัลกอริทึม DBSCAN จัดกลุ่มได้ดีกว่าและมีความเหมาะสมกับข้อมูลเชิงพื้นที่มากกว่า K-Means

ถัดมา Wang et al. (2019) นำเสนอวิธีการประยุกต์อัลกอริทึมการจัดกลุ่มมาผนวกกับงานทางด้านภูมิศาสตร์อย่างชัดเจน โดยทำการเชื่อมต่อ (Projection) ข้อมูล POI ของเขต Hanyang ประเทศจีนลงบนข้อมูลเส้นทางถนน (Street Network) สาธารณะของ OpenStreetMap แล้วทำการพัฒนาอัลกอริทึม NS-DBSCAN ขึ้นมาจัดการข้อมูลจุดพิกัดกับข้อมูลเส้นทางถนน โดยมีการประเมินคุณภาพของกลุ่มด้วยดัชนีวัดประสิทธิภาพแบบภายใน เช่น Silhouette Coefficient และ Davie-Bouldin Index แต่ตามทฤษฎีแล้วอาจจะไม่เป็นธรรมกับอัลกอริทึมจัดกลุ่มตามความหนาแน่นมากนัก

Aksac et al. (2019) ประยุกต์อัลกอริทึมการจัดกลุ่มเข้ากับทฤษฎีกราฟ (Graph Theory) ด้วยการสร้างเป็นโครงข่ายของชุดข้อมูล กล่าวคือ งานทดลองเลือกใช้วิธีการทางด้านเรขาคณิต (Geometry) เช่น Delaunay Triangulation เชื่อมต่อจุดพิกัดเข้าด้วยกันเป็นโครงร่างตาข่าย แล้วทำการจัดกลุ่มด้วย งานวิจัยนี้มีข้อควรระวัง คือ เราจะสูญเสียการวัดระยะห่างของข้อมูลโดยปริยาย เพราะการสร้างเป็นกราฟไม่ได้คำนึงถึงความใกล้-ความไกลของระยะทางเชิงพื้นที่จริงๆ แต่คำนึงถึงเพียงหลักการความใกล้-ไกลทางเรขาคณิตเท่านั้น ซึ่งขัดกับกฎข้อแรกทางภูมิศาสตร์ที่ว่า ตำแหน่งที่ตั้งของวัตถุที่อยู่ใกล้กัน ควรจะมีแนวโน้มความสัมพันธ์ทางด้านกายภาพหรือหน้าที่การทำงานคล้าย ๆ กัน

โดยทั่วไปแล้ว การจัดกลุ่มมักจะเริ่มต้นด้วยคำถามหลัก คือ *อัลกอริทึมการจัดกลุ่มแบบใด ให้ผลลัพธ์การจัดกลุ่มที่ดีที่สุด* การวัดประสิทธิภาพการจัดกลุ่มจึงเป็นความท้าทายสำคัญของงานการเรียนรู้ของเครื่องแบบไม่มีผู้สอน เพราะวิธีการทำงานของอัลกอริทึมแต่ละเทคนิคมีความแตกต่างกัน การวัดประสิทธิภาพการจัดกลุ่มจึงเป็นหลักฐานแสดงประสิทธิภาพ รวมทั้งสามารถเปรียบเทียบความสามารถระหว่างอัลกอริทึมการจัดกลุ่มได้อย่างเป็นธรรม

เมื่อพิจารณานิยามของอัลกอริทึมการจัดกลุ่มตามความหนาแน่นแล้ว ดัชนีวัดคุณภาพการจัดกลุ่มแบบภายในนั้นกลับมีพื้นฐานแนวคิดขัดแย้งกับแนวคิดของอัลกอริทึมการจัดกลุ่ม กล่าวคือ ดัชนีวัดคุณภาพ

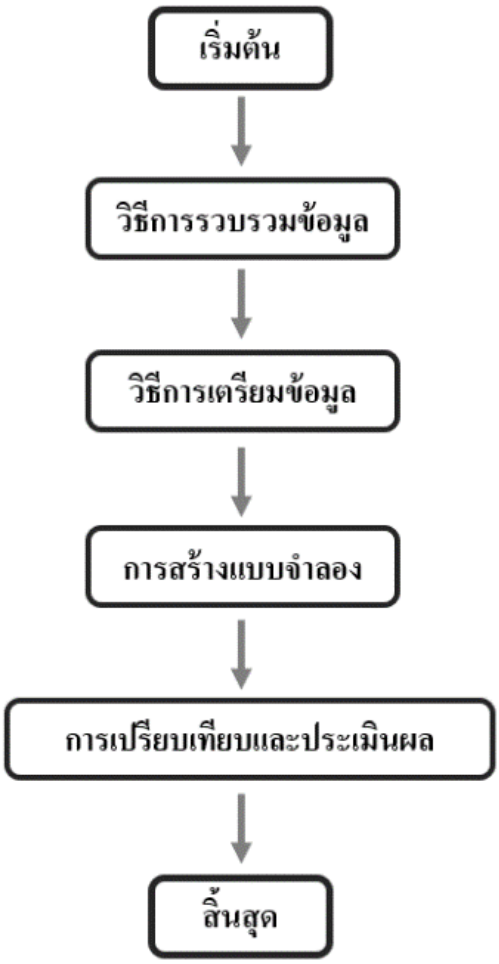
ดังกล่าวจะคำนวณประสิทธิภาพด้วยการยึดตำแหน่งศูนย์กลางของกลุ่มหรือศูนย์กลางของข้อมูลเป็นหลัก (Globular Clustering) จึงทำให้เหมาะกับชุดข้อมูลที่มีการกระจายตัวเท่า ๆ กันรอบจุดศูนย์กลางมากกว่าข้อมูลที่มีการกระจายตัวแบบไร้รูปแบบ (Arbitrary Shape) นอกจากนี้การวัดคุณภาพการจัดกลุ่มแบบภายในยังพิจารณาข้อมูลรบกวนเสมือนเป็นกลุ่ม ๆ หนึ่ง แต่ตามนิยามของการจัดกลุ่มแล้ว ข้อมูลรบกวนไม่ควรถูกจัดเป็นกลุ่ม ๆ หนึ่งเลยด้วยซ้ำ

จากข้อจำกัดข้างต้น Halkidi et al. (2002 และ 2008) จึงพัฒนาดัชนีวัดคุณภาพการจัดกลุ่ม CDbw และ Moulavi et al. (2014) พัฒนาดัชนีวัดคุณภาพการจัดกลุ่ม DBCV ขึ้นมา โดยทั้งสองเทคนิคเป็นดัชนีวัดคุณภาพแบบเทียบสัมพันธ์ โดยนำหลักการความหนาแน่น (Density-based Validation) มาประยุกต์เพื่อให้การวัดคุณภาพสะท้อนธรรมชาติของข้อมูล ส่วนงานวิจัยของ Craenendonck และ Blockeel (2015) วางแผนทดสอบการจัดกลุ่มคล้าย ๆ กัน แล้วนำอัลกอริทึมการจัดกลุ่มตามศูนย์กลาง เช่น K-Means อัลกอริทึมการจัดกลุ่มตามความหนาแน่น เช่น Meanshift หรือ DBSCAN และอัลกอริทึมการจัดกลุ่มตามการกระจายตัว เช่น EM มาทดสอบกับชุดข้อมูลทั้งหมด 27 ชุด (Public Data) แล้ววัดคุณภาพการจัดกลุ่ม พบว่า Calinski-Harabasz Index วัดประสิทธิภาพดีเมื่อทำงานร่วมกับอัลกอริทึม K-Means และ Davie-Bouldin Index กับ DBCV วัดประสิทธิภาพดีเมื่อทำงานร่วมกับอัลกอริทึมการจัดกลุ่มตามความหนาแน่น อย่างเช่น DBSCAN เป็นต้น

บทที่ 3
ระเบียบวิธีวิจัย

งานวิจัยนี้เป็นงานวิจัยเชิงประจักษ์ (Empirical Research) โดยนำอัลกอริทึมการจัดกลุ่มตามความหนาแน่นมาประยุกต์กับการจัดกลุ่มกับข้อมูลจริง (Real-world Data) โดยมีจุดประสงค์เพื่อช่วยวางกรอบแนวคิดการจัดกลุ่มข้อมูลเชิงพื้นที่ให้มีความเหมาะสม ระเบียบวิธีวิจัยมีขั้นตอน ดังนี้

3.1 แนวทางการวิจัย



ภาพที่ 3.1 แผนผังภาพรวมของแนวทางการวิจัย

3.1 ขั้นตอนการทำงานโดยละเอียด

3.2.1 วิธีการรวบรวมข้อมูล (Data Collection)

ข้อมูลสำหรับงานวิจัยนี้มาจากฐานข้อมูลจากเว็บเพจ www.baania.com ซึ่งเป็นเว็บไซต์ให้บริการข้อมูลที่อยู่อาศัย บ้านเดี่ยว คอนโดมิเนียม ทาวน์เฮ้าส์ โดยข้อมูลที่ดึง (Web Scraping) ออกมาประกอบด้วยโครงการที่อยู่อาศัยที่เปิดขายอยู่และกำลังจะเปิดขายช่วงครึ่งหลังปี 2019 และครึ่งปีแรก 2020 จำนวน กว่า 2,000 โครงการ ดังตารางที่ 3.1 และ ภาพที่ 3.1

ตารางที่ 3.1 ตัวแปรต่าง ๆ ของข้อมูลโครงการที่อยู่ซึ่งดึงผ่านเว็บเพจบริการข้อมูลที่อยู่อาศัย

textperiod	project_code	prj_name_en	market_name_en	latitude	longitude	total_sold_percent	onsale_absorption
2019H2	SKV00222	168 Residence Sukhumvit 36	Condo	13.717052	100.576298	99.05	0.33
2019H2	CBD100036	28 Chidlom	Condo	13.746497	100.544186	54.80	0.00
2019H2	CNT00114	88 The Terminal	Condo	13.791689	100.474149	98.20	1.00
2019H2	CBD100037	98 Wireless	Condo	13.741750	100.546956	96.10	0.50
2019H2	CNT00176	A Plus Inspire Rattanathibet 11	Condo	13.859024	100.506653	100.00	0.17

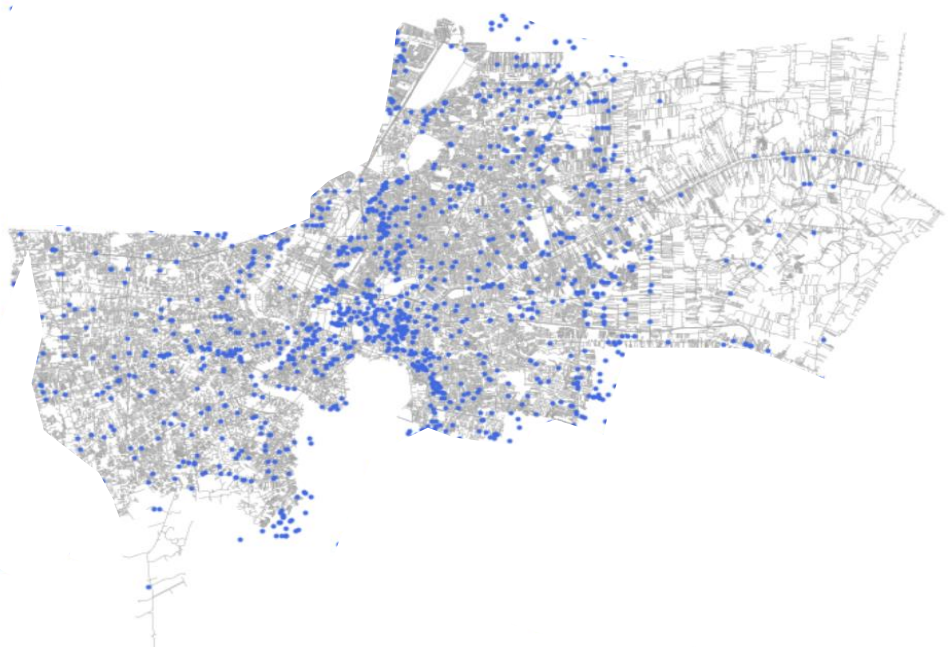
3.2.2 วิธีการเตรียมข้อมูล (Data Preparation)

สำหรับการสร้างแบบจำลองการจัดกลุ่ม เนื่องจากอัลกอริทึมการจัดกลุ่มตามความหนาแน่นบางเทคนิค เช่น HDBSCAN ไม่สามารถจัดการข้อมูลเชิงคุณภาพ (Attributed Data) ที่มาพร้อมกับ เช่น ราคาขาย ขนาดพื้นที่ จำนวนชั้นของตึก ผู้วิจัยจึงเลือกเฉพาะข้อมูลตำแหน่งที่ตั้งของโครงการที่อยู่อาศัยเท่านั้น โดยมีขั้นตอนพอสังเขป ดังนี้

1) เลือกเฉพาะตัวแปรทางพิกัดภูมิศาสตร์ เช่น ละติจูดและลองจิจูด

2) ทำการแปลงหน่วยของรูปร่างเรขาคณิต (Geometry) แบบต่าง ๆ ของโครงการที่อยู่อาศัยให้เป็นพิกัดทางภูมิศาสตร์แบบจุด (Point Geometry)

3) คำนวณความคล้ายกันของข้อมูล ณ ลำดับถัดไป



ภาพที่ 3.1 ตัวอย่างการแสดงผลข้อมูลตำแหน่งที่อยู่อาศัย กว่า 2,000 โครงการ

3.2.3 การสร้างแบบจำลอง (Modeling)

การวัดความคล้ายกันของข้อมูล

เมื่อเตรียมข้อมูลเสร็จเรียบร้อยแล้ว ก่อนการเริ่มจัดกลุ่ม อัลกอริทึมเดียวกันจะถูกทดสอบด้วยวิธีวัดระยะห่าง 2 วิธี คือ วิธีวัดระยะห่างแบบ Euclidean และ Haversine ข้อควรระวังเมื่อเลือกมาตรวัดระยะห่าง คือ การคำนวณระยะห่างระหว่างข้อมูลเชิงพื้นที่ เราจะต้องเปลี่ยนหน่วยวัดของจุดพิกัดภูมิศาสตร์ให้อยู่ในหน่วยวัดที่เหมาะสม มาตรวัดระยะห่างแบบ Haversine จะต้องเปลี่ยนจากละติจูดเป็นมุมที่วัดระหว่างจุดใด ๆ กับเส้นศูนย์สูตร ส่วนลองจิจูดจะต้องเปลี่ยนเป็นมุมที่วัดระหว่างจุดใด ๆ กับเส้นเมอริเดียนที่ศูนย์ดังกล่าวที่หัวข้อ 2.1.2 และตารางที่ 3.2

ตารางที่ 3.2 ตัวอย่างการเปลี่ยนหน่วยวัดจากจุดพิกัดเป็นองศา

โครงการ	ละติจูด	ลองจิจูด	ละติจูด (องศา)	ลองจิจูด (องศา)
1	13.717052	100.576298	0.239408	1.755388
2	13.46497	100.544186	0.239922	1.754827
3	13.791689	100.474149	0.240710	1.753605
4	13.741750	100.546956	0.239839	1.754875

อัลกอริทึมการจัดกลุ่มตามความหนาแน่น

เมื่อเตรียมข้อมูลเสร็จเรียบร้อยแล้ว ก่อนการจัดกลุ่ม อัลกอริทึมเดียวกันจะถูกทดสอบด้วยวิธีวัดระยะห่าง 2 วิธี คือ วิธีวัดระยะห่างแบบ Euclidean และ Haversine อย่างไรก็ตาม ข้อควรระวังเมื่อเลือกมาตรวัดระยะห่าง คือ ก่อนคำนวณระยะห่างระหว่างข้อมูลเชิงพื้นที่ เราจะต้องเปลี่ยนหน่วยวัดของจุดพิกัดภูมิศาสตร์ให้อยู่ในหน่วยวัดที่เหมาะสม หลังจากนั้น เราจึงมาจัดกลุ่มด้วยอัลกอริทึมการจัดกลุ่มตามความหนาแน่น ดังหัวข้อ 2.1.5 พร้อมทั้งกำหนดพารามิเตอร์ที่เหมาะสม แล้วทำการแปลผลด้วยดัชนีวัดคุณภาพการจัดกลุ่มและการแปลผลด้วยสายตา

ตารางที่ 3.3 ตัวอย่างผลลัพธ์การจัดกลุ่ม

โครงการ	ละติจูด	ลองจิจูด	ละติจูด (องศา)	ลองจิจูด (องศา)	กลุ่ม
1	Lat 1	Lon 1	Lat 1 Rad	Lon 1 Rad	กลุ่มที่ 1
2	Lat 2	Lon 2	Lat 2 Rad	Lon 2 Rad	กลุ่มที่ 2
3	Lat 3	Lon 3	Lat 3 Rad	Lon 3 Rad	กลุ่มที่ 1
4	Lat 4	Lon 4	Lat 4 Rad	Lon 4 Rad	กลุ่มที่ 1
5	Lat 5	Lon 5	Lat 5 Rad	Lon 5 Rad	กลุ่มที่ 2
6	Lat 6	Lon 6	Lat 6 Rad	Lon 6 Rad	ข้อมูลรบกวน

ดัชนีวัดคุณภาพการจัดกลุ่มและการแปลผลการจัดกลุ่มด้วยสายตา

ข้อมูลแบบที่ไม่มีลาเบลเป้าหมายกำกับก่อน งานวิจัยนี้จึงประเมินคุณภาพการจัดกลุ่ม ด้วยดัชนีวัดประสิทธิภาพการจัดกลุ่มแบบภายในและแบบสัมพันธ์ จำนวนทั้งหมด 4 เทคนิค ประกอบด้วย Silhouette Coefficient, Calinski-Harabasz, CDbw และ DBCV ดังกล่าวรายละเอียดที่หัวข้อ 2.1.6 เพื่อเปรียบเทียบผลลัพธ์การจัดกลุ่มที่มีความเหมาะสมกับชุดข้อมูลเชิงพื้นที่มากที่สุด นอกจากการใช้ดัชนีวัดคุณภาพการจัดกลุ่มแล้ว การแปลผลด้วยสายตาจะถูกนำมาใช้เพื่อคัดเลือกกลุ่มที่น่าสนใจสำหรับการสร้างแบบจำลองทำนายราคาเฉลี่ยของโครงการที่อยู่อาศัยด้วย

การสร้างแบบจำลองเพื่อทำนายราคาเฉลี่ยของโครงการที่อยู่อาศัย

สำหรับการศึกษานี้ใช้ข้อมูล ประกอบด้วยโครงการที่อยู่อาศัยที่เปิดขายอยู่และกำลังจะเปิดขายช่วงครึ่งหลังปี 2019 และครึ่งปีแรก 2020 จำนวน กว่า 2,000 โครงการ ข้อมูลต่างๆ สามารถแบ่งออกเป็น 3 ปัจจัยที่ส่งผลต่อแบบจำลองราคาเฉลี่ยของโครงการที่อยู่อาศัย ดังตารางที่ 3.2

1) ปัจจัยด้านทำเลที่ตั้ง มีรายละเอียดดังต่อไปนี้

- สถานะของสถานีรถไฟฟ้าที่ใกล้ที่สุด (Station Status) แบ่งออกเป็น สถานีรถไฟฟ้าที่เปิดใช้งานอยู่แล้ว กับ สถานีรถไฟฟ้าที่อยู่ระหว่างการก่อสร้าง
- สถานีรถไฟฟ้าจุดเชื่อมต่อ (Interchange Station) คือ ตัวแปรระบุว่าโครงการที่อยู่อาศัยอยู่ใกล้กับจุดเชื่อมต่อของโครงข่ายสถานีรถไฟฟ้าหรือไม่
- ระยะทางจากโครงการที่อยู่อาศัยถึงสถานีรถไฟฟ้าที่ใกล้ที่สุด (Station Tier) แบ่งออกเป็น 4 กลุ่ม โดยใช้ระยะห่างจากโครงการที่อยู่อาศัยจนถึงสถานีรถไฟฟ้าเป็นเกณฑ์การแบ่ง โดยกลุ่มที่ 1 มีระยะห่างน้อยกว่า 3 กิโลเมตร, กลุ่มที่ 2 มีระยะห่างมากกว่า 3 จนถึง 5 กิโลเมตร, กลุ่มที่ 3 มีระยะห่างมากกว่า 5 จนถึง 10 กิโลเมตร และกลุ่มสุดท้ายมีระยะห่างมากกว่า 10 กิโลเมตร
- ตำแหน่งที่ตั้งของโครงการที่อยู่อาศัย (Location) ประกอบด้วย ตำแหน่งละติจูดและลองจิจูด อย่างไรก็ตาม ตัวแปรทั้งสองนี้จะไม่นำมาพิจารณาการสร้างแบบจำลองเพื่อทำนายราคาเฉลี่ยของโครงการที่อยู่อาศัย เพราะถูกนำมาพิจารณาตั้งแต่การจัดกลุ่มของโครงการที่อยู่อาศัยแล้ว

2) ปัจจัยด้านลักษณะของโครงการ มีรายละเอียดดังต่อไปนี้

- โครงการของบริษัทมหาชน (Developers) พิจารณาจากผู้ประกอบการที่พัฒนาโครงการโดยแบ่งออกเป็น บริษัทมหาชน ได้แก่ แสนสิริ, อนันดา และ SC Asset และผู้ประกอบการรายย่อย
- อายุโครงการหรือปีที่สร้าง (Project Duration) ข้อมูลจะระบุปีที่สร้าง แล้วคำนวณอายุของโครงการตั้งแต่เริ่มเปิดขายมาจนถึงสิ้นปี 2562

- ความหรูหราของโครงการ (Segment) คือ ระดับของโครงการที่อยู่อาศัย ปกติจะสะท้อนช่วงราคาขายของโครงการหรือถูกนำมาใช้โฆษณาการขาย เพื่อดึงดูดกลุ่มลูกค้าเป้าหมาย เช่น ระดับโครงการคอนโดมิเนียม Segment A จะสะท้อนช่วงราคาขายมากกว่า 300,000 บาทต่อตารางเมตร
- จำนวนยูนิตทั้งหมด (Total Unit) คือ จำนวนห้องที่เปิดขายของโครงการ

3) ปัจจัยความเคลื่อนไหวของการขายและตลาด มีรายละเอียดดังต่อไปนี้

- จำนวนยูนิตที่ขายได้ (Total Sold) คือ จำนวนยูนิตที่ขายได้แล้วของทั้งโครงการ
- ร้อยละจำนวนยูนิตที่ขายได้ (Percent of Total Sold) คือ จำนวนยูนิตที่ขายได้แล้วของทั้งโครงการ เทียบกับจำนวนยูนิตทั้งหมด
- จำนวนยูนิตที่ยังเปิดขาย (Total Unsold) คือ จำนวนยูนิตคงเหลือของทั้งโครงการ
- ร้อยละจำนวนยูนิตที่ยังเปิดขาย (Percent of Total Unsold) คือ จำนวนยูนิตที่ยังเปิดขายอยู่ เทียบกับจำนวนยูนิตทั้งหมด
- อัตราการดูดซับ (Total Absorption Rate) คือ ดัชนีสะท้อนถึงอุปสงค์หรือปริมาณความต้องการซื้อ คำนวณจากจำนวนยูนิตทั้งหมดที่ขายได้ตลอดอายุโครงการ หารด้วยจำนวนยูนิตทั้งหมดของโครงการ แล้วคูณด้วย 100 ออกมาเป็นร้อยละ กล่าวคือ เมื่ออัตราดูดซับมีค่าสูง โครงการนั้นๆ มีโอกาสในการลงทุน ปล่อยขายต่อหรือปล่อยเช่าได้ง่าย
- อัตราการดูดซับระหว่างรอบการขาย (On-sale Absorption Rate) คือ ดัชนีสะท้อนถึงอุปสงค์หรือปริมาณความต้องการซื้อเ คล้ายกับอัตราการดูดซับ แต่จะคำนวณจากจำนวนยูนิตที่ขายได้ ณ รอบเวลานั้นๆ หารด้วยจำนวนยูนิต
- ระยะเวลาที่คาดว่าโครงการจะยังมียูนิตเหลือขายอยู่ (Remaining Time to Sell) คือ สัดส่วนของจำนวนยูนิตที่ยังเปิดขายอยู่ เทียบกับอัตราการดูดซับ
- ราคาเฉลี่ยต่อตารางเมตร (Price per SQM) ถือว่าเป็นตัวแปรเป้าหมาย เนื่องจากโครงการหนึ่งๆ มีห้องหลายรูปแบบ จึงใช้ราคาเฉลี่ยของทั้งโครงการเป็นตัวแทน

ตารางที่ 3.4 รายละเอียดของข้อมูลต่างๆ ที่ใช้ในการวิเคราะห์

ลำดับ	ลักษณะของตัวแปร	ตัวแปร	หน่วยวัด
1	**ทำเลที่ตั้ง**	สถานะสถานีรถไฟฟ้าที่ใกล้ที่สุด	ตัวแปรหุ่น
2		สถานีรถไฟฟ้าจุดเชื่อมต่อ	ตัวแปรหุ่น
3		ระยะทางจากโครงการที่อยู่อาศัย ถึงสถานีรถไฟฟ้าที่ใกล้ที่สุด	กิโลเมตร
4		ตำแหน่งที่ตั้ง (ละติจูด ลองจิจูด)	องศา
5	**ลักษณะของโครงการ**	โครงการที่พัฒนาของบริษัทมหาชน	ตัวแปรหุ่น
6		อายุโครงการ	เดือน
7		ความหรูหราของโครงการ	ตัวแปรหุ่น
8		จำนวนยูนิตทั้งหมด	ยูนิต
9	**ความเคลื่อนไหวของการขายและตลาด**	จำนวนยูนิตที่ขายได้	ยูนิต
10		ร้อยละจำนวนยูนิตที่ขายได้	ร้อยละ
11		จำนวนยูนิตที่ยังเปิดขาย	ยูนิต
12		ร้อยละจำนวนยูนิตที่ยังเปิดขาย	ร้อยละ
13		อัตราการดูดซับ	ร้อยละ
14		อัตราดูดซับระหว่างรอบการขาย	ร้อยละ
15		ระยะเวลาที่คาดว่าโครงการจะยังมียูนิตเหลือขายอยู่	เดือน
16		ราคาเฉลี่ยต่อตารางเมตร	พันบาท/ตารางเมตร

หลังจากนั้นนำข้อมูลมาสร้างแบบจำลองถดถอยทั้งหมด 6 แบบ ด้วย Python Library ชื่อว่า Lazypredict มีการกำหนดค่าพารามิเตอร์เป็นค่าเริ่มต้น ประกอบด้วยเทคนิคต่าง ๆ ดังนี้ 1) การวิเคราะห์ถดถอยเชิงเส้น 2) การวิเคราะห์ถดถอยแลซโซ 3) การวิเคราะห์ถดถอยอิลาสติคเน็ต 4) ต้นไม้ตัดสินใจแบบถดถอย 5) การวิเคราะห์ถดถอยการสุ่มของป่า และ 6) การวิเคราะห์ถดถอยเอ็กซ์จีบูสท์

แบบจำลองแต่ละแบบจะแบ่งข้อมูลออกเป็น 2 ส่วน ส่วนแรกคือ ข้อมูลชุดข้อมูลที่นำมาสอนแบบจำลอง คิดเป็นร้อยละ 70 และส่วนที่สองคือ ข้อมูลสำหรับทดสอบประสิทธิภาพของแบบจำลองอีกร้อยละ 30 แล้วจึงทำการวัดผลความแม่นยำด้วยค่าความคลาดเคลื่อน ณ ลำดับถัดไป

3.2.4 การเปรียบเทียบและประเมินผล (Comparison and Evaluation)

ค่าความคลาดเคลื่อนที่นำมาใช้วัดความแม่นยำของแบบจำลอง ประกอบด้วย 1) รากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ย หรือ RMSE และ 2) ค่าเฉลี่ยของร้อยละความผิดพลาดสัมบูรณ์ หรือ MAPE เพื่อรับมือกับกรณีชุดข้อมูลมีการกระจายตัวแบบปกติและกรณีที่ชุดข้อมูลที่มีค่าผิดปกติ ตามลำดับ ถึงแม้ทั้ง 2 วิธีจะมีการคำนวณค่าความคลาดเคลื่อนและหน่วยวัดของความคลาดเคลื่อนแตกต่างกัน แต่มีจุดร่วมกันคือ หากแบบจำลองมีค่าความคลาดเคลื่อนน้อยจะให้ค่าพยากรณ์ใกล้เคียงกับความเป็นจริงมากที่สุด

3.3 เครื่องมือที่ใช้ในการวิจัย

3.3.1 คอมพิวเตอร์ลูกข่ายสำหรับการสร้างโมเดล มีคุณสมบัติ ดังนี้

1) หน่วยประมวลผลกลาง CPU Intel R i7-6500U 2.50 GHz 80 Core

2) หน่วยความจำ 8 GB

3) พื้นที่จัดเก็บ 237 GB

3.3.2 ซอร์ฟแวร์ที่ใช้ในงานวิจัยมีรายละเอียดดังนี้

1) Jupyter Notebook

2) Python (version 3.4.4)

3) Python Library ต่าง ๆ เช่น NetworkX, OSMNX, Sklearn และ Lazypredict

บทที่ 4
ผลการทดลอง

4.1 ผลการศึกษา

4.1.1 การจัดกลุ่มข้อมูลของโครงการที่อยู่อาศัย

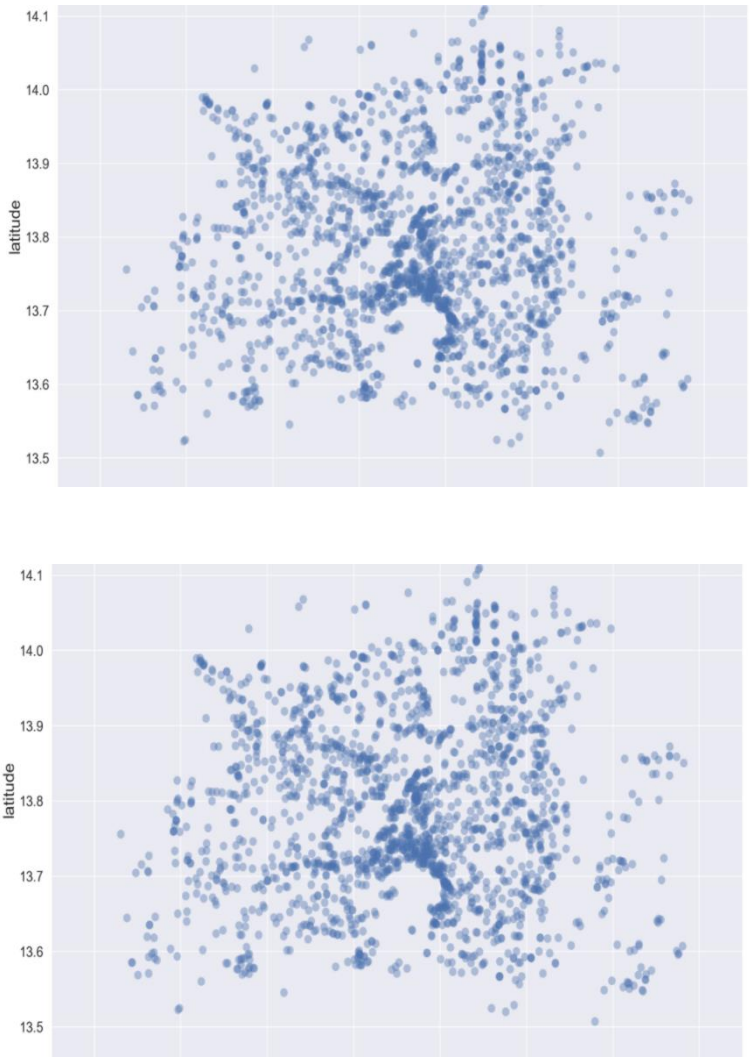
งานวิจัยนี้ศึกษาการจัดกลุ่มข้อมูลเชิงพื้นที่แบบไม่มีลาเป้าหมายกำกับ เพื่อนำเป็นตัวแทนสำหรับกลุ่มของโครงการที่อยู่อาศัย โดยใช้อัลกอริทึมการจัดกลุ่ม ประกอบด้วย DBSCAN และHDBSCAN ทั้งหมด 4 แบบจำลองซึ่งกำหนดค่าพารามิเตอร์ที่แตกต่างกัน หลังจากนั้นจึงวัดคุณภาพของการจัดกลุ่มด้วย Silhouette coefficient, CH index, CDbw และ DBCV และแปลผลการจัดกลุ่มของข้อมูลด้วยสายตาโดยเบื้องต้น ผลการศึกษาของแบบจำลองการจัดกลุ่มทั้ง 4 แบบ มีรายละเอียดดังนี้

ตารางที่ 4.1 ผลการศึกษาแบบจำลองการจัดกลุ่มและดัชนีวัดคุณภาพทั้ง 4 แบบจำลอง

Algorithms	Distance	Parameters	# Cluster	Silhouette	CH	CDbw	DBCV
DBSCAN	Euclidean	eps = 2, minpt = 1	1	-	-	-	-
	Haversine	eps = 2, minpt = 2	42	-0.46	19.0	**0.84**	2.41
	Haversine	eps = 1, minpt = 1	445	0.12	**216.7**	0.71	1.98
HDBSCAN	Haversine	eps = 2, minpt = 1	33	**0.19**	85.5	0.75	**2.45**

แบบจำลองการจัดกลุ่มที่ 1: DBSCAN, Euclidean’s distance, eps = 2, minpt = 1

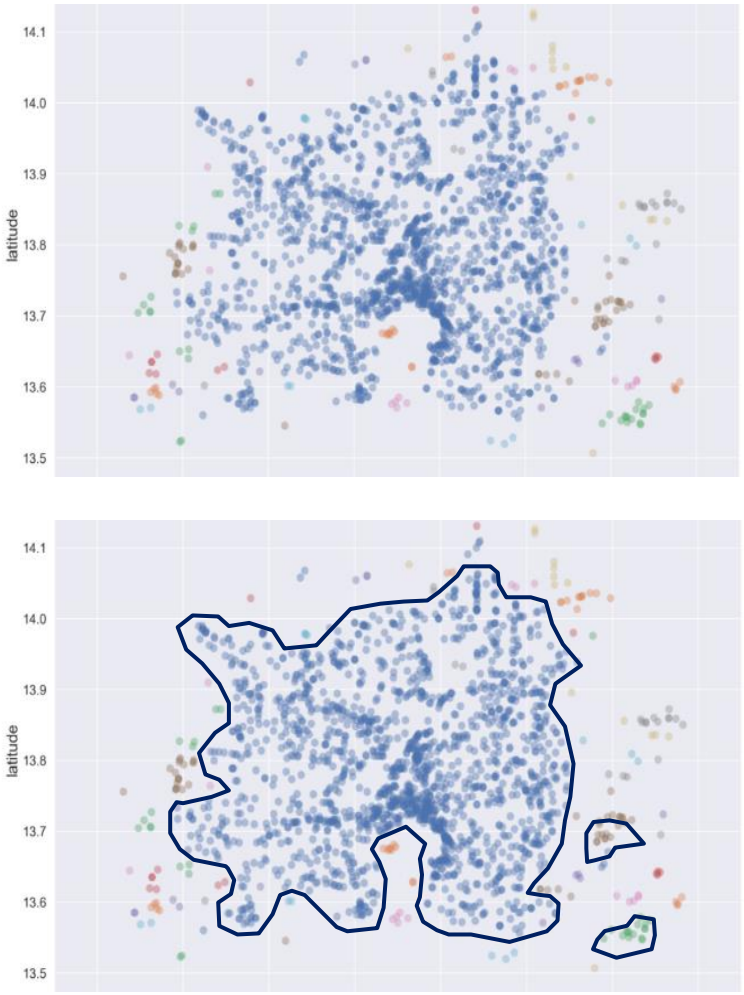
แบบจำลองนี้ถือว่าเป็นความล้มเหลวของการจัดกลุ่มข้อมูล เพราะไม่สามารถจัดกลุ่มโครงการที่อยู่อาศัยได้ โดยผลลัพธ์การจัดกลุ่มมีเพียง 1 กลุ่มเท่านั้น เนื่องจากวิธีการวัดความคล้ายกันของข้อมูลด้วยวิธีวัดระยะห่างแบบ Euclidean นั้น ไม่เหมาะสมกับข้อมูลที่กระจายตัวอยู่บนพื้นที่ขนาดใหญ่ที่มีอิทธิผลความโค้งของพื้นผิวโลกเข้ามาเป็นตัวแปรด้วย



ภาพที่ 4.1 ผลลัพธ์การจัดกลุ่มและการแปลผลด้วยสายตาของแบบจำลองการจัดกลุ่มที่ 1

แบบจำลองการจัดกลุ่มที่ 2: DBSCAN, Haversine distance, eps = 2, minpt = 2

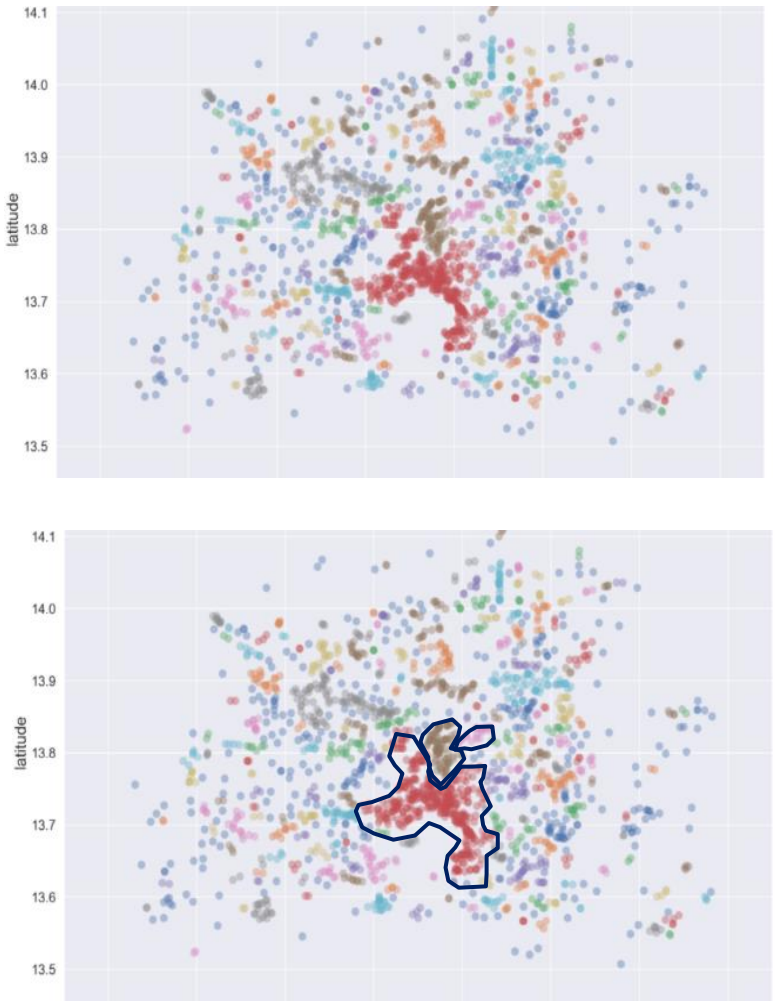
สำหรับกรณีนี้ แบบจำลองสามารถแบ่งกลุ่มออกเป็น 42 กลุ่ม เนื่องจากการเลือกใช้วิธีวัดระยะห่างระห่างข้อมูลแบบ Haversine เหมาะสมมากกว่า Euclidean โดยมีดัชนีวัดคุณภาพการจัดกลุ่ม CDbw สูงมากที่สุดเท่ากับ 0.84 และ DBCV มีค่าสูงใกล้เคียงกับแบบจำลองอื่น ๆ อย่างไรก็ตาม เมื่อพิจารณารายละเอียดการจัดกลุ่มด้วยสายตา พบว่า ณ จุดศูนย์กลาง แบบจำลองมีการจัดกลุ่มที่มีขนาดใหญ่มากเกินไป ส่วนกลุ่มขนาดเล็กๆ มีการกระจัดกระจายอยู่บริเวณขอบนอกของพื้นที่



ภาพที่ 4.2 ผลลัพธ์การจัดกลุ่มและการแปลผลด้วยสายตาของแบบจำลองการจัดกลุ่มที่ 2

แบบจำลองการจัดกลุ่มที่ 3: DBSCAN, Haversine distance, eps = 1, minpt = 1

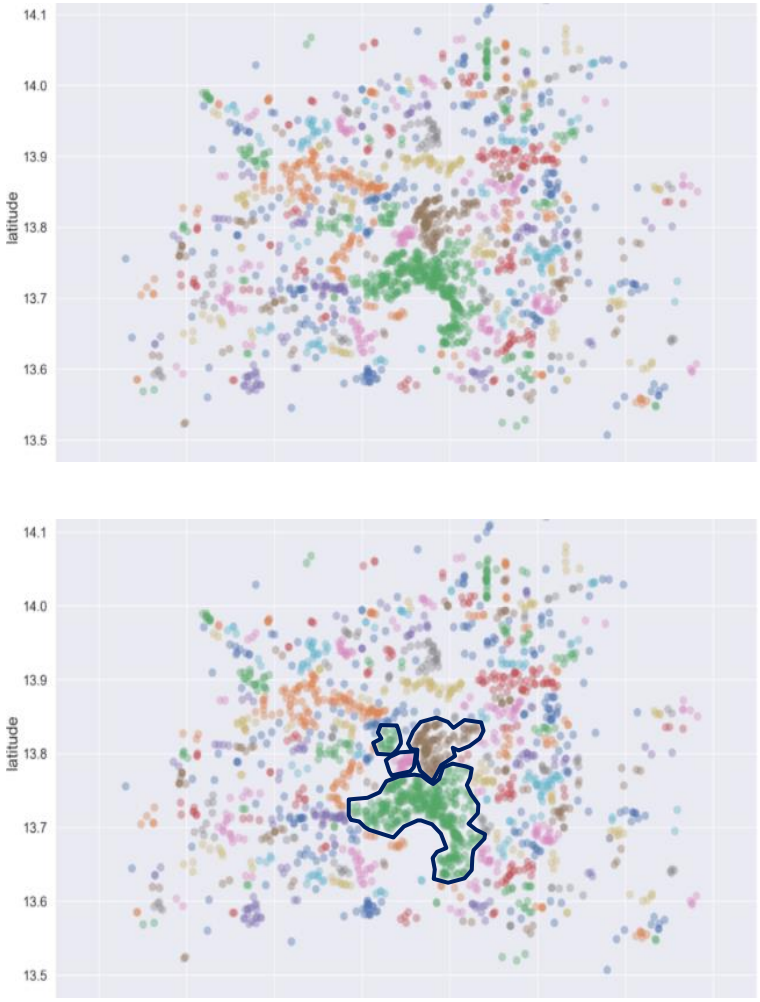
แบบจำลองนี้มีการกำหนดค่าพารามิเตอร์ต่าง ๆ คล้ายกับแบบจำลองที่ 2 แต่กำหนดรัศมีจากจุดศูนย์กลางให้สั้นลงและจำนวนสมาชิกของกลุ่มลดลงด้วย พบว่า เมื่อเปรียบเทียบกับแบบจำลองอื่นๆ ถึงแม้ว่าแบบจำลองนี้จะมีดัชนีวัดคุณภาพ CH เท่ากับ 216.7 ซึ่งมากที่สุด แต่กลุ่มข้อมูลกลับมีจำนวนมากเกินไป จำนวนทั้งหมด 445 กลุ่ม ทำให้ไม่สามารถแปลผลการจัดกลุ่มอย่างมีนัยยะได้



ภาพที่ 4.3 ผลลัพธ์การจัดกลุ่มและการแปลผลด้วยสายตาของแบบจำลองการจัดกลุ่มที่ 3

แบบจำลองการจัดกลุ่มที่ 4: HDBSCAN, Haversine distance, eps = 2, minpt = 1

จากผลลัพธ์พบว่าแบบจำลองนี้ มีจำนวนกลุ่มทั้งหมด 33 กลุ่ม รวมทั้งสามารถทำงานได้ดีบนพื้นที่ขนาดใหญ่ เพราะมีดัชนีวัดคุณภาพ Silhouette coefficient เท่ากับ 0.19 และ DBCV เท่ากับ 2.45 ซึ่งสูงกว่าแบบจำลองอื่น ๆ บ่งบอกถึงการจัดกลุ่มที่มีคุณภาพ สามารถแยกออกจากกลุ่มที่อยู่ห่างไกลอย่างชัดเจน นอกจากนี้แบบจำลองที่ 4 ยังพบกลุ่มเสถียร (Stable cluster) ณ จุดกึ่งกลางของพื้นที่ (กลุ่มสีเขียว) เช่นเดียวกับแบบจำลองที่ 3 (กลุ่มสีแดง) ซึ่งเป็นกลุ่มที่น่าสนใจและควรศึกษาถัดไป



ภาพที่ 4.4 ผลลัพธ์การจัดกลุ่มและการแปลผลด้วยสายตาของแบบจำลองการจัดกลุ่มที่ 4

ด้วยผลการทดลองพบกลุ่มเสถียรที่มีความหนาแน่น ณ กลางพื้นที่เสมอ นอกจากดัชนีวัดคุณภาพกลุ่มของแต่ละแบบจำลองการจัดกลุ่มที่เป็นตัวบ่งบอกคุณภาพของกลุ่มข้างต้นแล้ว เรายังมีเหตุผลด้านอื่นๆ ที่สนับสนุนความน่าสนใจของกลุ่มดังกล่าว เพื่อนำมาสร้างแบบจำลองทำนาย ดังนี้

1. จำนวนโครงการที่อยู่อาศัยที่มีอยู่แล้ว เป็นตัวสะท้อนอุปสงค์ที่มีอยู่ของผู้ซื้อ/ผู้เช่า บริเวณที่มีจำนวนโครงการที่อยู่อาศัยหนาแน่น จึงเป็นพื้นที่ศักยภาพที่น่าลงทุน

2. พื้นที่ดินรอการพัฒนา (Land Bank) บริเวณพื้นที่หนาแน่นอาจจะยังมีเหลืออยู่ แล้วแต่ความสามารถของบริษัทพัฒนาอสังหาริมทรัพย์ที่สามารถครอบครองได้ ขณะเดียวกันบริเวณพื้นที่ห่างไกลชุมชนเมือง ถึงแม้จะมีพื้นที่ว่างมากกว่า แต่อาจจะยังไม่มีพื้นที่ดินที่มีศักยภาพเพียงพอ

3. การลากเส้นแบ่งพื้นที่ถือว่าเป็นเรื่องประสบการณ์ส่วนบุคคล (Subjective) การนำอัลกอริทึมมาสร้างแบบจำลองการจัดกลุ่ม จึงเป็นอีกหนึ่งมาตราฐานที่ทำให้อภิปรายบนหลักการเดียวกัน

4.1.2 การทำนายราคาเฉลี่ยและประเมินผลค่าความคลาดเคลื่อน

งานวิจัยเลือกกลุ่มเสถียรที่ตั้งอยู่ตรงกลางของพื้นที่ตามเหตุผลข้างต้น แล้วนำมาสร้างแบบจำลองทำนายราคาเฉลี่ยต่อตารางเมตร โดยแบ่งกลุ่มเสถียรดังกล่าวออกเป็น 2 กลุ่มย่อยตามตัวแปรความหรูหราของโครงการที่อยู่อาศัย คือ 1) กลุ่มที่มีราคาสูงกว่า (Higher Segment) และ 2) กลุ่มที่มีราคาต่ำกว่า (Lower Segment) เพื่อลดความแปรปรวนและเพิ่มความแม่นยำของแบบจำลอง

กลุ่มที่มีราคาสูงกว่า หรือ Higher Segment

จากผลลัพธ์ของกลุ่มที่มีราคาสูงกว่าพบว่า การวิเคราะห์ถดถอยเชิงเส้นมีประสิทธิภาพสูงกว่าเมื่อเทียบกับแบบจำลองอื่น ๆ มีค่า RMSE เท่ากับ 88.96 นั้นหมายความว่า ค่าพยากรณ์ราคาเฉลี่ยต่อตารางเมตรเบี่ยงเบนมากกว่าราคาจริงเฉลี่ย 88,960 บาทต่อตารางเมตร หรือ น้อยกว่าราคาจริงเฉลี่ย 88,960 บาทต่อตารางเมตร ส่วนค่าความคลาดเคลื่อน MAPE มีค่าเท่ากับ 0.25 หมายความว่า ผลลัพธ์จากการทำนายมีความคลาดเคลื่อนจากราคาเฉลี่ยจริงต่อตารางเมตรอยู่ร้อยละ 25

การศึกษานี้ยังพบอีกว่า แบบจำลองถดถอยประเภทต้นไม้ เช่น ต้นไม้ตัดสินใจถดถอย และ การวิเคราะห์ถดถอยการสุ่มของป่ามีค่าความคลาดเคลื่อนทั้ง RMSE และ MAPE มากกว่าแบบจำลองถดถอยแบบเส้นตรง ส่วนการวิเคราะห์ถดถอยเอ็กซ์จีบูสท์ ถึงแม้จะเป็นแบบจำลองถดถอยประเภทต้นไม้ตัดสินใจเหมือนกัน กลับมีค่าความคลาดเคลื่อนที่น้อยกว่า เพราะพื้นฐานของแบบจำลองที่เรียนรู้ความผิดพลาดจากต้นไม้ก่อนหน้า แล้วนำปรับปรุงประสิทธิภาพของแบบจำลองให้ดีขึ้น

ตารางที่ 4.2 ผลลัพธ์ค่าความคลาดเคลื่อนของกลุ่มที่มีราคาสูงกว่า

แบบจำลอง	RMSE	MAPE
การวิเคราะห์ถดถอยเชิงเส้น (Linear Regression)	**88.96**	**0.25**
การวิเคราะห์ถดถอยแลซโซ (Lasso Regression)	92.18	0.26
การวิเคราะห์ถดถอยอิลาสติคเน็ต (Elastic Net Regression)	89.80	**0.25**
ต้นไม้ตัดสินใจแบบถดถอย (Regression Tree)	111.74	0.32
การวิเคราะห์ถดถอยการสุ่มของป่า (Random Forest Regression)	94.25	0.26
การวิเคราะห์ถดถอยเอ็กซ์จีบูสท์ (XGB)	90.76	0.26

กลุ่มที่มีราคาต่ำกว่า หรือ Lower Segment

ผลลัพธ์ของกลุ่มที่มีราคาต่ำกว่าพบว่า ค่าความคลาดเคลื่อน RMSE มีขนาดน้อยกว่าเมื่อเทียบกับกลุ่มที่อยู่อาศัยที่มีราคาสูง จากแบบจำลองทั้ง 6 แบบมีค่า RMSE ตั้งแต่ 24.86 จนถึง 39.14 โดยการวิเคราะห์ถดถอยแลซโซมีค่า RMSE ที่น้อยที่สุด ส่วนค่าความคลาดเคลื่อน MAPE มีขนาดใกล้เคียงกับแบบจำลองกลุ่มที่มีราคาสูงกว่า เนื่องจากคำนวณเป็นค่าสัมบูรณ์ โดยการวิเคราะห์ถดถอยแลซโซมีค่า MAPE น้อยที่สุดเท่ากับ 0.20 หรือคลาดเคลื่อนจากราคาเฉลี่ยจริงอยู่ร้อยละ 20

อย่างไรก็ตามแบบจำลองถดถอยประเภทต้นไม้ยังคงมีค่าความคลาดเคลื่อนทั้ง RMSE และ MAPE สูงกว่าแบบจำลองถดถอยแบบเส้นตรง หรือกล่าวอีกนัยหนึ่งคือ สำหรับกรณีนี้แบบจำลองถดถอยแบบเส้นตรงมีความแม่นยำของการทำนายสูงกว่าแบบจำลองถดถอยประเภทต้นไม้

ตารางที่ 4.3 ผลลัพธ์ค่าความคลาดเคลื่อนของกลุ่มที่มีราคาต่ำกว่า

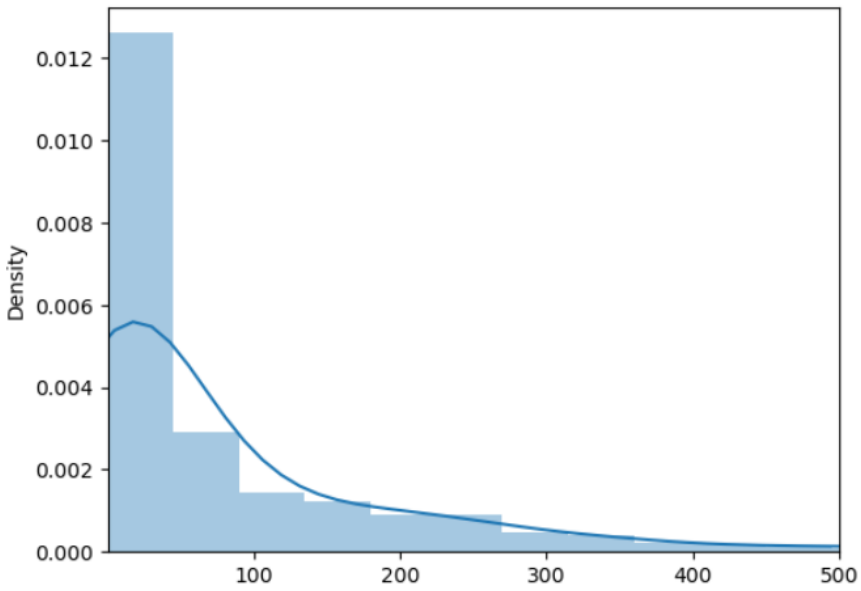
แบบจำลอง	RMSE	MAPE
การวิเคราะห์ถดถอยเชิงเส้น (Linear Regression)	36.27	0.30
การวิเคราะห์ถดถอยแลซโซ (Lasso Regression)	**24.86**	**0.20**
การวิเคราะห์ถดถอยอิลาสติคเน็ต (Elastic Net Regression)	26.56	0.21
ต้นไม้ตัดสินใจแบบถดถอย (Regression Tree)	36.23	0.27
การวิเคราะห์ถดถอยการสุ่มของป่า (Random Forest Regression)	34.70	0.30
การวิเคราะห์ถดถอยเอ็กซ์จีบูสท์ (XGB)	39.14	0.33

การกระจายตัวของข้อมูลโครงการที่อยู่อาศัย

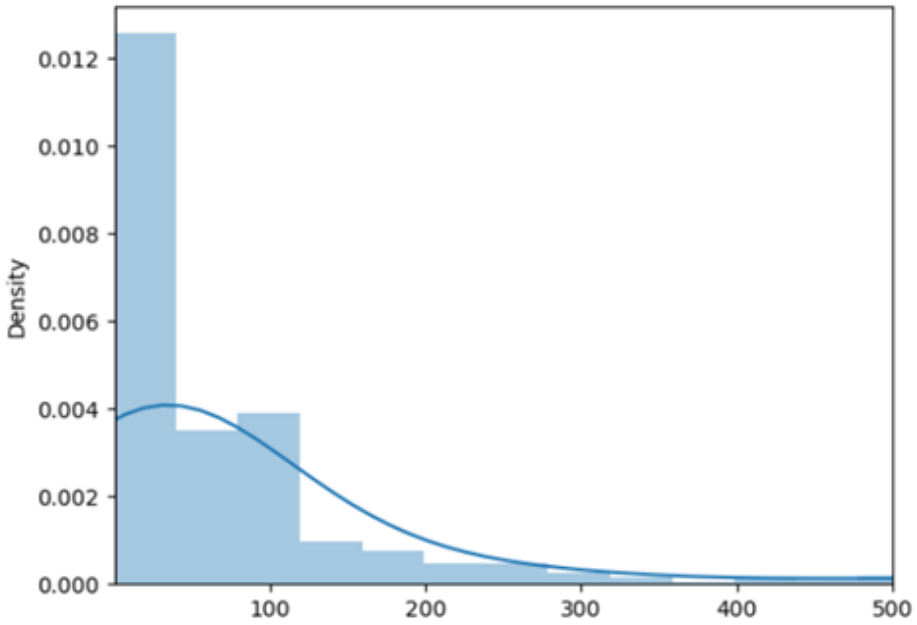
ดังกล่าว ณ บทที่ 3 ตารางที่ 3.4 รายละเอียดของข้อมูลต่างๆ ที่ใช้ในการวิเคราะห์ว่า ราคาเฉลี่ยต่อตารางเมตรของโครงการที่อยู่อาศัยคำนวณมาจากรูปแบบห้องทุกแบบที่มีอยู่ แล้วนำมาเฉลี่ยเป็นค่าเดียวเพื่อเป็นราคาตัวแทนของทั้งโครงการ เมื่อพิจารณาการกระจายตัวของราคาเฉลี่ยต่อตารางเมตรพบว่า ราคาเฉลี่ยของกลุ่มที่มีราคาสูงกว่าและกลุ่มที่มีราคาต่ำกว่ามีความซ้อนทับกัน

การกระจายตัวของข้อมูลราคาเฉลี่ยยังส่งผลกระทบต่อความแม่นยำของแบบจำลองด้วย กล่าวคือ กลุ่มที่มีราคาสูงกว่านั้น ประกอบด้วยโครงการที่มีราคาเฉลี่ยต่อตารางเมตรน้อยกว่า 100,000 บาท จนถึง 500,000 บาทต่อตารางเมตร ดังภาพที่ 4.5 ด้วยเหตุนี้แบบจำลองถดถอยประเภทต้นไม้จึงมีค่าความคลาดเคลื่อนสูง เพราะมีความอ่อนไหวต่อค่าผิดปกติ ทำให้การเปลี่ยนแปลงข้อมูลเพียงเล็กน้อยอาจจะส่งผลไปถึงโครงสร้างและลำดับการสร้างตัดไม้ตัดสินใจ

ส่วนกลุ่มที่มีราคาต่ำกว่ามีความผันผวนน้อยกว่า โดยจำนวนโครงการส่วนใหญ่แล้วมีราคาเฉลี่ยไม่เกิน 100,000 บาทต่อตารางเมตร ดังภาพที่ 4.6 ด้วยเหตุนี้แบบจำลองของกลุ่มที่มีราคาต่ำกว่า จึงมีค่าความคลาดเคลื่อนแบบ RMSE น้อยกว่าเมื่อเทียบกับแบบจำลองของกลุ่มที่มีราคาสูงกว่า เพราะราคาเฉลี่ยของแต่ละโครงการมีความใกล้เคียงกันนั้นเอง



ภาพที่ 4.5 การกระจายตัวราคาเฉลี่ย (พันต่อตารางเมตร) ของกลุ่มที่มีราคาสูงกว่า (ซ้าย)



ภาพที่ 4.6 การกระจายตัวราคาเฉลี่ย (พันต่อตารางเมตร) ของกลุ่มที่มีราคาต่ำกว่า

บทที่ 5
บทสรุปและข้อเสนอแนะ

5.1 สรุปผลงานวิจัย

5.1.1 การจัดกลุ่มข้อมูลของโครงการที่อยู่อาศัย

งานวิจัยนี้ศึกษาการจัดกลุ่มข้อมูลเชิงพื้นที่แบบไม่มีลาเป้าหมายกำกับ เพื่อนำเป็นตัวแทนสำหรับการทำความเป็นกลุ่มก้อนของโครงการที่อยู่อาศัย โดยใช้เทคนิคอัลกอริทึมการจัดกลุ่ม ประกอบด้วย DBSCAN และ HDBSCAN จำนวนทั้งหมด 4 แบบจำลอง แล้ววัดผลด้วยดัชนีวัดคุณภาพ Silhouette coefficient, CH index, CDbw และ DBCV พบว่า เทคนิค HDBSCAN (Haversine, eps = 2, minpt = 1) มีค่า Silhouette coefficient เท่ากับ 0.19 และ DBCV เท่ากับ 2.45 ซึ่งถือว่ามีคุณภาพของกลุ่มมากกว่าแบบจำลองอื่นๆ บ่งบอกถึงการจัดกลุ่มที่มีความเป็นกลุ่มก้อนและสามารถแบ่งแยกออกจากกลุ่มรอบข้างอย่างชัดเจน อย่างไรก็ตาม ถึงแม้เทคนิคข้างต้นจะมีค่าดัชนี Silhouette coefficient สูง แต่ยังคงต่ำกว่า 1 ซึ่งเป็นค่าที่มากที่สุด เนื่องจากชุดข้อมูลมีข้อมูลรอบนอกค่อนข้างกระจายตัว จึงส่งผลกระทบต่อการคำนวณดัชนีวัดคุณภาพ Silhouette coefficient

ส่วนผลลัพธ์อื่นๆ ที่น่าสนใจคือ เทคนิค DBSCAN (Euclidean, eps = 2, minpt = 1) ถือว่าเกิดความล้มเหลวของการจัดกลุ่ม เนื่องจากผลลัพธ์การจัดกลุ่มมีเพียงกลุ่มข้อมูลเดียว ทำให้ไม่สามารถประเมินความเป็นกลุ่มด้วยดัชนีวัดคุณภาพได้ ดังนั้นสิ่งที่ควรระมัดระวังเมื่อเลือกใช้วิธีวัดระยะห่างระหว่างข้อมูลคือ วิธีคำนวณระยะห่างแบบ Euclidean ไม่เหมาะสมกับข้อมูลเชิงพื้นที่ซึ่งกระจายตัวอยู่บนพื้นที่ขนาดใหญ่

5.1.2 การทำนายราคาเฉลี่ยของโครงการที่อยู่อาศัย

หลังจากการจัดกลุ่มข้อมูลโครงการที่อยู่อาศัยแล้วพบว่า มีการผสมกันของโครงการที่มีราคาขายเฉลี่ยต่อตารางเมตรที่หลากหลาย ซึ่งอาจทำให้การสร้างแบบจำลองทำนายราคาเฉลี่ยได้ผลออกมาไม่ดีเท่าที่ควร จึงแบ่งกลุ่มข้อมูลเดียวกันออกเป็น 2 กลุ่มย่อยตามราคาเฉลี่ยของโครงการที่อยู่อาศัย คือ 1) กลุ่มที่มีราคาสูงกว่า (Higher Segment) และ 2) กลุ่มที่มีราคาต่ำกว่า (Lower Segment) ก่อนจะนำมาสร้างแบบจำลองทำนายราคาเฉลี่ย แล้ววัดประสิทธิภาพด้วยวิธี RMSE และ MAPE โดยภาพรวมแล้ว กลุ่มที่มีราคาต่ำกว่าให้ผลการทำนายผิดพลาดน้อยกว่ากลุ่มที่มีราคาสูงกว่า เนื่องจากการการกระจายตัวของราคาขายเฉลี่ยเกาะกลุ่มกัน กล่าวอีกนัยหนึ่งคือ ช่วงราคาขายเฉลี่ยของกลุ่มที่มีราคาต่ำกว่ามีราคาขายเฉลี่ยต่อตารางเมตรใกล้เคียงกันนั้นเอง

5.2 ข้อเสนอแนะ

5.2.1 การจัดกลุ่มข้อมูลของโครงการที่อยู่อาศัย

สำหรับการจัดกลุ่มข้อมูล การศึกษานี้พิจารณาเฉพาะปัจจัยตำแหน่งที่ตั้งของโครงการที่อยู่อาศัยเพียงอย่างเดียว หากอนาคตมีการพิจารณาตัวแปรดัชนีเชิงพื้นที่อื่นๆ เช่น เส้นถนนที่ตัดผ่าน (Street Network) หรือระยะทางถึงสถานที่อำนวยความสะดวก (Amenities) รวมทั้งอาจจะพัฒนาการจัดกลุ่มแบบอื่นๆ ร่วมกัน เช่น การจัดกลุ่มที่คำนึงถึงความสัมพันธ์กับเพื่อนบ้าน (Spatial Constraint Multivariate Cluster: SCMC) เทคนิคต่างๆ ประกอบด้วย การพิจารณาความสัมพันธ์แบบด้านข้าง (Contiguity Edge Only), การพิจารณาความสัมพันธ์แบบเชิงมุม (Contiguity Edge Corner) หรือ การพิจารณาสร้างโครงข่ายเพื่อนบ้านสามเหลี่ยม (Delaunay Triangulation) อาจจะทำให้ได้ผลลัพธ์การจัดกลุ่มที่หลากหลายและสอดคล้องกับสภาพแวดล้อมความของพื้นที่มากยิ่งขึ้น

5.2.2 การทำนายราคาเฉลี่ยของโครงการที่อยู่อาศัย

การศึกษานี้เลือกประเมินประสิทธิภาพของการทำนายราคาเฉลี่ยของโครงการที่อยู่อาศัย โดยเปรียบเทียบค่าจริงกับผลลัพธ์ที่เกิดขึ้นจากการทำนายผล ด้วยวิธีการประเมินค่า RMSE และ MAPE ซึ่งทั้งสองวิธีจะให้ผลลัพธ์ขึ้นกับลักษณะของชุดข้อมูลและจุดประสงค์ของการนำไปใช้ กล่าวคือ แม้ว่า RMSE จะมีความอ่อนไหวต่อข้อมูลรบกวนมากกว่า MAPE แต่หน่วยวัดของประสิทธิภาพนั้นมีหน่วยเดียวกับตัวแปรตาม ทำให้ตัวผู้ทดลองคุ้นเคยและง่ายต่อการสร้างแบบจำลองทางคณิตศาสตร์เพื่อการหาทางเลือกที่ดีที่สุด (Model Optimization) ด้วยการปรับค่าพารามิเตอร์ต่างๆ ของแบบจำลอง ส่วน MAPE นั้นมีความอ่อนไหวต่อข้อมูลรบกวนน้อยกว่าและแสดงประสิทธิภาพในรูปแบบร้อยละ ทำให้ง่ายต่อการสื่อสารกับผู้มีส่วนเกี่ยวข้อง (Stakeholder) ว่าแบบจำลองการทำนายราคาเฉลี่ยแบบใดมีประสิทธิภาพที่ดีมากกว่ากัน

ข้อจำกัดของ Python Library ที่เลือกใช้นั้นมีให้เลือกวิธีประเมินค่าความผิดพลาดจากการทำนายอย่างจำกัด อาจจะมีวิธีประเมินค่าความผิดพลาดอื่นๆ ที่สามารถประเมินประสิทธิภาพหรืออธิบายผลของแบบจำลองได้เหมาะสมมากกว่า เช่น วิธี Root Mean Square Logarithmic Error: RMSLE ที่มีหน่วยวัดความผิดพลาดเช่นเดียวกับตัวแปรที่ต้องการทำนาย อีกทั้งยังลดอิทธิพลของข้อมูลรบกวนด้วยการมาตราส่วนลอการิทึม

ประการสุดท้าย หากพิจารณาตามมุมมองของธุรกิจ เราควรจะพิจารณารูปแบบห้องและจำนวนของห้องแต่ละรูปแบบ เพื่อนำมาเป็นตัวแทนของราคาเฉลี่ยต่อตารางเมตรที่เหมาะสมกับห้องแต่ละรูปแบบ เพราะการเลือกใช้ราคาเฉลี่ยต่อตารางเมตรเพียงแค่ 1 ราคาเพื่อเป็นตัวแทนของทั้งโครงการ อาจไม่ได้สะท้อนราคาเฉลี่ยที่แท้จริงของทุกรูปแบบห้องและอาจมองเป็นการเหมารวมมากเกินไป

บรรณานุกรม

บรรณานุกรม

[1] N. Ahmed and A. Razak, A comparative study of different density based spatial clustering algorithms, *International Journal of Computer Applications*, vol. 99, no. 5, pp.18-25, 2014

[2] G. Boeing, Clustering to reduce Spatial data set size, SocArXiv, 2018

[3] T. Wang, C. Ren, Y. Luo and J. Tian, NS-DBSCAN: A Density-Based Clustering Algorithm in Network Space, *SPRS International Journal of Geo-Information*, vol. 8, 2018

[4] A. Aksac Wang, T. Ozyer, and R. Alhajj, CutESC: Cutting edge spatial clustering technique based on proximity graphs, *Pattern Recognit*, vol. 96, 2019

[5] M. Halkidi and M. Vazirgiannis, A density-based cluster validity approach using multi representatives, Pattern Recognition Letters, pp. 773–786, 2008

[6] D. Moulavi and P. A. Jaskowiak, Density-based clustering validation, *Proceedings of the 2014 SIAM International Conference on Data Mining*, 2014

[7] M. Ester and H. P. Kriegel, A density-based algorithm for discovering cluster in large spatial database with noise, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD96)*, vol.33, 1996

[8] M. Leland and H. John, Accelerated Hierarchical Density Based Clustering, IEEE International Conference on Data Mining Workshops 2017 (ICDMW), pp. 33-42, 2017

[9] T. V. Craenendonck, and H. B. K. Leuven, Using internal validity measures to compare clustering algorithms, ICML, 2015

[10] P. J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics, vol. 20, pp. 53–65, 1987

[11] T. Cali´nski and J. Harabasz, A dendrite method for cluster analysis, Communications in Statistics, pp. 1-27, 1974

ภาคผนวก

ภาคผนวก ก

ผลงานตีพิมพ์

IBDAP and BDAP 2021, August 26-27, Bangkok, Thailand, 2021

การจัดกลุ่มข้อมูลเชิงพื้นที่ตามความหนาแน่นของโครงการที่อยู่อาศัยในกรุงเทพมหานคร
Density-based Clustering for Residential Locations: A Bangkok Use Case

วิทวัส สรสวัง (Witawat Sangsawang)[1] และ ดวงใจ อัศวชื่น (Duangjai Askongchuen)[2]
[1]สาขาวิชาวิศวกรรมข้อมูลขนาดใหญ่ วิทยาลัยนวัตกรรมดิจิทัลและวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์
[2]สาขาวิชาวิศวกรรมข้อมูลขนาดใหญ่ วิทยาลัยนวัตกรรมดิจิทัลและวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์
[1]5951620200012@dpu.ac.th, [2]duangjai.as@dpu.ac.th

บทคัดย่อ

ปัญหาของพิจารณาขอบเขตพื้นที่ มักจะแบ่งตามขอบเขตการปกครองหรือประสบการณ์ของผู้เชี่ยวชาญเป็นหลัก การจัดกลุ่มข้อมูลเชิงพื้นที่จึงเข้ามาช่วยกำหนดมุมมองใหม่ เช่น การจัดกลุ่มที่อยู่อาศัยตามตำแหน่งที่ตั้งหรือตามราคาที่ขาย เป็นต้น งานวิจัยก่อนหน้านี้มุ่งเน้นเพื่อพัฒนาอัลกอริทึมการจัดกลุ่มที่มีประสิทธิภาพ แต่เฉพาะเจาะจงกับชุดข้อมูลทดลองมากเกินไป รวมถึงการเลือกดัชนีวัดคุณภาพไม่สอดคล้องกับแนวคิดอัลกอริทึมการจัดกลุ่ม ด้วยเหตุนี้ การศึกษานี้จึงนำอัลกอริทึมการจัดกลุ่มเชิงพื้นที่ตามความหนาแน่น ประกอบด้วย DBSCAN และ HDBSCAN มาทดลองกับข้อมูลโครงการที่อยู่อาศัยจำนวน 2010 โครงการทั่วกรุงเทพฯ แล้วทำการเปรียบประเมินคุณภาพด้วยดัชนีวัดคุณภาพกลุ่ม จำนวน 4 เทคนิค พบว่า HDBSCAN (Haversine, eps = 2, minpt = 1) มีคุณภาพการจัดกลุ่มที่ดีที่สุด มีค่า Silhouette Coefficient และ DBCV เท่ากับ 0.19 และ 2.45 ตามลำดับ

คำสำคัญ: การจัดกลุ่ม ดัชนีวัดคุณภาพกลุ่ม โครงการที่อยู่อาศัย

Abstract

Spatial clustering analysis plays such an important task to divide residential locations into clusters. It can cluster the locations better than traditionally or experiencedly allowed. Most recent works have developed many sufficient clustering algorithms while they are just suitable for only the experimental synthetic dataset themselves. Therefore, rather than just defined those locations according to predetermined areas for a purpose of administration, our study is to apply well-known density-based clustering algorithms i.e., DBSCAN and HDBSCAN on real-world dataset of residential locations entire Bangkok metropolitan. Thus, we comparatively evaluate the clustering results with 4 cluster validation techniques. Our results show that HDBSCAN (Haversine, eps = 2, minpt = 1) outperforms than the others with silhouette coefficient and DBCV, 0.19 and 2.45, respectively.

Keyword: Clustering, Cluster Validation, Residential

1. บทนำ

การวิเคราะห์ข้อมูลของผู้ประกอบการอสังหาริมทรัพย์มักจะแบ่งพื้นที่ (Zoning) ตามขอบเขตการปกครองที่กำหนดโดยหน่วยงานรัฐหรือประสบการณ์ผู้เชี่ยวชาญเป็นหลัก การกำหนดขอบเขตพื้นที่ดังกล่าว อาจเกิดอคติต่อมุมมองทางด้านพื้นที่ เพราะไม่ได้คำนึงถึงสภาพแวดล้อมทางกายภาพรอบข้าง เช่น โครงการที่อยู่อาศัยข้างเคียง ดังนั้น การจัดกลุ่มข้อมูล (Data Clustering) ซึ่งเป็นเทคนิคการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) จึงเข้ามาช่วยกำหนดมุมมองและช่วยทำความเข้าใจสภาวะตลาดของการพัฒนาโครงการที่อยู่อาศัยในอนาคต

อย่างไรก็ตาม การจัดกลุ่มข้อมูลเชิงพื้นที่ มีข้อควรพิจารณา 3 ประเด็นด้วยกัน ประกอบด้วย การวัดความคล้ายกัน (Similarity Measure) ด้วยการคำนวณระยะห่าง

ระหว่างข้อมูล จัดเป็นพื้นฐานสำคัญและมีอิทธิพลอย่างมากต่อผลการจัดกลุ่ม ข้อมูลชุดเดียวกันแต่หากมีวิธีคำนวณระยะห่างแตกต่างกัน อาจจะทำให้ผลลัพธ์การจัดกลุ่มแตกต่างกันด้วย ประการที่สอง การเลือกอัลกอริทึมการจัดกลุ่ม (Clustering Algorithms) ต้องสอดคล้องกับธรรมชาติของข้อมูลที่นำมาวิเคราะห์ รวมทั้งการกำหนดค่าพารามิเตอร์ซึ่งส่งผลต่อจำนวนกลุ่มที่แบ่งได้ และประการสุดท้าย การเลือกดัชนีวัดคุณภาพกลุ่ม (Cluster Validation) สำหรับข้อมูลเชิงพื้นที่ ยังมีความท้าทายอย่างมาก เนื่องจากดัชนีวัดคุณภาพบางเทคนิคพัฒนาบนแนวคิดที่อาจจะขัดแย้งกับอัลกอริทึมการจัดกลุ่มเชิงพื้นที่ตามความหนาแน่น ดังนั้นการเลือกดัชนีวัดคุณภาพควรจะเหมาะสมกับประเภทและการกระจายตัวของข้อมูล รวมถึงสอดคล้องกับวิธีการทำงานของอัลกอริทึมการจัดกลุ่มด้วย

จากประเด็นที่กล่าวข้างต้น ผู้วิจัยจึงจัดกลุ่มข้อมูลด้วยอัลกอริทึมการจัดกลุ่มเชิงพื้นที่ตามความหนาแน่น 2 อัลกอริทึม จำนวน 4 เทคนิคกับข้อมูลเชิงพื้นที่จริงของโครงการที่อยู่อาศัย หลังจากนั้นประเมินคุณภาพด้วยดัชนีวัดคุณภาพกลุ่ม 4 ดัชนี เพื่อเปรียบเทียบผลลัพธ์การจัดกลุ่ม แล้วจึงนำมาอธิบายความหมายของกลุ่ม

2. วรรณกรรมที่เกี่ยวข้อง

การจัดกลุ่มข้อมูลเชิงพื้นที่ด้วยวิธีการเรียนรู้ของเครื่องแบบไม่มีผู้สอนนั้นมีความท้าทายอย่างยิ่ง เพราะไม่มีข้อสรุปร่วมกันว่าวิธีการใดคือวิธีปฏิบัตินิยมที่ให้ผลดีที่สุด [1] จากการทบทวนวรรณกรรม หลักการการจัดกลุ่มข้อมูลเชิงพื้นที่ ประกอบด้วย 3 ส่วน คือ การวัดความคล้ายกันของข้อมูล อัลกอริทึมการจัดกลุ่มและดัชนีวัดคุณภาพการจัดกลุ่ม ส่วน Boeing [2] จัดกลุ่มข้อมูลระบุตำแหน่ง GPS ด้วย DBSCAN งานวิจัยนี้เลือกวิธีวัดระยะห่างแบบ Haversine ซึ่งคำนึงถึงอิทธิพลจากความโค้งของพื้นผิวโลก เหมาะกับข้อมูลที่มีการกระจายตัวครอบคลุมพื้นที่กว้างขวาง จากผลการทดลองพบว่า DBSCAN จัดกลุ่มได้ผลลัพธ์น่าพอใจ ลำดับถัดมามีการพัฒนา DBSCAN ประยุกต์เข้ากับทฤษฎีกราฟและเราขาคณิต เพื่อนำมาเป็นเงื่อนไขก่อนการจัดกลุ่ม Wang et al. [3] และ Aksac et al. [4] แล้วค่อยทำการจัดกลุ่ม แต่วิธีวัดระยะห่างนี้ ทำให้สูญเสียระยะห่างจริงของข้อมูล เพราะคำนึงถึงหลักการหาวงจรปิดเท่านั้น

การจัดกลุ่มมักจะตามมาด้วยคำถาม คือ อัลกอริทึมการจัดกลุ่มแบบใด ให้ผลลัพธ์การจัดกลุ่มที่ดีที่สุด โดยทั่วไปการวัดคุณภาพการจัดกลุ่มมักอธิบายผลทางด้านคุณภาพเป็นหลัก ด้วยเหตุนี้การวัดคุณภาพการจัดกลุ่มเชิงปริมาณจึงมีบทบาทเปรียบเสมือนการวัดค่าความถูกต้องของวิธีการเรียนรู้ของเครื่องแบบมีผู้สอน โดยทั่วไปแล้วงานวิจัยนิยมนำ Silhouette Coefficient และ Calinski-Harabasz Index ซึ่งเป็นดัชนีวัดคุณภาพการจัดกลุ่มแบบภายในมาวัดคุณภาพการจัดกลุ่ม ข้อเสียคือ ดัชนีทั้งสองเทคนิคนี้ไม่เหมาะกับชุดข้อมูลที่มีข้อมูลรบกวน (Noise) หรือข้อมูลที่มีการกระจายตัวแบบไม่แน่นอน (Arbitrary shape) เพราะข้อมูลรบกวนจะถูกมองเสมือนเป็นกลุ่ม ๆ หนึ่ง ทำให้การวัดคุณภาพการจัดกลุ่มคลาดเคลื่อนจากความเป็นจริง ต่อมา Halkidi et al. [5] และ Moulavi et al. [6] จึงพัฒนาดัชนีวัดคุณภาพแบบเฉพาะขึ้นมา คือ CDbw และ DBCV ตามลำดับ โดยประยุกต์หลักการความหนาแน่น (Density-based Validation) เพื่อความสอดคล้องกับอัลกอริทึมการจัดกลุ่มเชิงพื้นที่ตามความหนาแน่น

3. ขั้นตอนวิธีการดำเนินการวิจัย

งานวิจัยนี้เป็นงานวิจัยเชิงประยุกต์ ด้วยการนำอัลกอริทึมการจัดกลุ่มหลากหลายวิธีมาประยุกต์กับข้อมูลจริงของโครงการที่อยู่อาศัยทั่วกรุงเทพมหานคร

3.1 แหล่งที่มาและการรวบรวมข้อมูล

ฐานข้อมูลจากเว็บเพจ www.baania.com ซึ่งเป็นเว็บไซต์ให้บริการข้อมูลที่อยู่อาศัย ข้อมูลที่ดึงออกมา (Web Scraping) ประกอบด้วยโครงการที่เปิดขายอยู่ช่วงปี 2019-2020 จำนวน 2,010 โครงการ (ภาพที่ 1) ข้อมูลตัวแปรที่นำมาวิเคราะห์ คือ ข้อมูลด้านทำเลที่ตั้ง

3.2 วิธีการวิเคราะห์

3.2.1 วิธีวัดความคล้ายกัน (Similarity Measures)

การวัดความคล้ายกันเป็นปัจจัยสำคัญอย่างยิ่งต่อการจัดกลุ่มข้อมูล ซึ่งสามารถวัดความคล้ายกัน ด้วยการวัด

ภาพที่ 1: แสดงตำแหน่งโรงงานที่อยู่อาศัยที่มีในปี พ.ศ. 2019-2020

ระยะห่าง (Distance Function) โดยแนวคิดพื้นฐานของความคล้ายกัน คือ ข้อมูลที่อยู่กลุ่มเดียวกัน จะมีระยะห่างของข้อมูลของสมาชิกน้อยกว่าข้อมูลที่อยู่คนละกลุ่ม

1) มาตรวัดระยะห่างแบบ Euclidean

เทคนิคการวัดระยะห่างประเภทนี้ จัดเป็นมาตรวัดระยะห่างที่นิยมเทคนิคหนึ่งเพื่อคำนวณหาระยะทางที่สั้นที่สุดระหว่างข้อมูล 2 จุด มีสูตรการคำนวณหาระยะห่างมาจากทฤษฎีพีทาโกรัส ตามสมการที่ 1

$$d = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2} \quad (1)$$

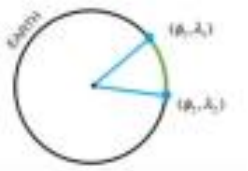
2) มาตรวัดระยะห่างแบบ Haversine

มาตรวัดระยะห่างแบบ Haversine มีหลักการวัดระยะห่างที่สั้นที่สุดระหว่างข้อมูล เช่นเดียวกับมาตรวัดระยะห่างแบบ Euclidean แต่จะพิจารณาความโค้งของพื้นผิวโลกตามพิกัดรูปทรงรี (Spherical Coordinate System) ร่วมด้วย ดังสมการที่ 2

มาตราวัดระยะห่างวิธีนี้ ผู้ใช้จะต้องเปลี่ยนหน่วยข้อมูลเป็นค่าระยะเชิงมุมก่อน โดยละติจูดเปลี่ยนเป็นมุมที่วัดระหว่างจุดใด ๆ กับเส้นศูนย์สูตร ส่วนลองจิจูดเปลี่ยนเป็นมุมที่วัดระหว่างจุดใด ๆ กับเส้นแนวเดียวที่ศูนย์ หากจุดพิกัดไม่ได้ถูกเปลี่ยนให้อยู่ในหน่วยที่เหมาะสม อาจทำให้ผลลัพธ์การจัดกลุ่มเกิดความคลาดเคลื่อนได้ (ภาพที่ 2)

$$d = 2r\arcsin\sqrt{\sin^2\left(\frac{\varphi_2-\varphi_1}{2}\right) + \cos(\varphi_1)\cos(\varphi_2)\sin^2\left(\frac{\lambda_2-\lambda_1}{2}\right)} \quad (2)$$

โดยที่ φ แสดงมุมของละติจูดที่วัดระหว่างจุดใด ๆ กับเส้นศูนย์สูตร ส่วน λ แสดงมุมของลองจิจูดที่วัดระหว่างจุดใด ๆ กับเส้นแนวเดียวที่ศูนย์

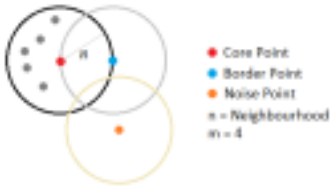


ภาพที่ 2: แสดงการคำนวณระยะห่างแบบ Haversine

3.2.2 อัลกอริทึมการจัดกลุ่ม (Cluster Algorithms)

1) DBSCAN (Density-Based Spatial Clustering of Applications with Noise) เริ่มดำเนินค่ารัศมีระยะห่าง (Epsilon) และจำนวนสมาชิกขั้นต่ำ (Minimum Points) เมื่อเริ่มการจัดกลุ่ม หากข้อมูลใด ๆ ที่เป็นจุดศูนย์กลางรวมกับข้อมูลที่อยู่ภายในรัศมี มีจำนวนเท่ากับจำนวนสมาชิกขั้นต่ำ เราจะเรียกข้อมูลนั้นว่า จุดศูนย์กลาง (Core Point) และจุดข้อมูลที่อยู่ใกล้ที่สุดที่อีกจุดอยู่ภายในรัศมี เรียกว่า จุดขอบ (Border Point) แล้วจึงรวมตัวกลายเป็นกลุ่มเดียวกัน จากนั้นดำเนินจุดขอบจะเปลี่ยนเป็นจุดศูนย์กลาง แล้วทำซ้ำ จนกระทั่งไม่มีจุดข้อมูลใดเข้าข่ายคำพารามิเตอร์ที่กำหนดแล้ว ส่วนข้อมูลที่อยู่เหลืออยู่ เราจะเรียกว่า ข้อมูลรบกวน (Noise) [7]

จุดเด่นของ DBSCAN นั้นสามารถจัดกลุ่มข้อมูลที่มีการกระจายตัวไม่เป็นรูปแบบและยังแยกและข้อมูลรบกวนได้ หากผู้ใช้งานคุ้นเคยกับข้อมูลแล้ว สามารถกำหนดค่ารัศมีระยะห่างและไม่จำเป็นต้องกำหนดจำนวนกลุ่มที่ต้องการแบ่ง อย่างไรก็ตาม ภายในชุดข้อมูลเดียวกัน หากมีการกระจายตัวหรือความหนาแน่นของข้อมูลที่แตกต่างกัน (Varied Density) แม้ยังคงกำหนดพารามิเตอร์รับเข้าได้ จะทำให้การจัดกลุ่มด้วยวิธีนี้ไม่มีประสิทธิภาพ (ภาพที่ 3)



ภาพที่ 3: แผนภาพแสดงวิธีการทำงานของ DBSCAN

2) HDBSCAN (Hierarchical of DBSCAN) มีหลักการจัดกลุ่มแบบพื้นฐานความหนาแน่น (Density-based) เหมือนกับ DBSCAN สำหรับค่าพารามิเตอร์ HDBSCAN กำหนดเพื่องจำนวนสมาชิกขั้นต่ำ ทำให้มีโอกาสจัดกลุ่มที่มีขนาดใหญ่ขึ้น เพราะไม่มีข้อจำกัดทางด้านรัศมีระยะห่างเข้ามาเกี่ยวข้อง อัลกอริทึมนี้พัฒนาขึ้นเพื่อแก้ปัญหาความหนาแน่นและขนาดของข้อมูลที่แตกต่างซึ่งเกิดขึ้นภายในชุดข้อมูลเดียวกัน รวมทั้งเรื่องกระจายตัวแบบไม่แน่นอนด้วย [8]

3.3.3. ดัชนีวัดคุณภาพกลุ่ม (Cluster Validation)

โดยทั่วไปแล้ว ดัชนีคุณภาพการจัดกลุ่ม แบ่งออกเป็น 3 ประเภท [9] คือ ดัชนีวัดประสิทธิภาพการจัดกลุ่มแบบภายนอก (External Validation) แบบภายใน (Internal Validation) และแบบสัมพัทธ์ (Relative Validation) งานวิจัยนี้เลือกเฉพาะดัชนีวัดคุณภาพแบบภายในและแบบสัมพัทธ์ (ตารางที่ 1) เพราะมีความเหมาะสมกับปัญหาแบบไม่มีสามารถกำกับ มีรายละเอียด ดังนี้

1) Silhouette Coefficient คำนวณหาความเหมือนกันหรือใกล้เคียงกันของข้อมูล จากอัตราส่วนของผลต่างระยะห่างระหว่างกลุ่มกับระยะห่างภายในกลุ่มเดียวกันเทียบกับระยะห่างที่มากที่สุด ปกติแล้วค่าดัชนีนี้จะอยู่ระหว่าง -1 และ 1 โดยกลุ่มที่มีคุณภาพหรือความเหมือนกันสูง จะมีค่าเข้าใกล้ 1 [10]

2) Calinski-Harabasz Index คือ อัตราส่วนของความแปรปรวนระหว่างกลุ่มกับความแปรปรวนภายในกลุ่ม โดยความแปรปรวนระหว่างกลุ่ม คำนวณมาจากจุดศูนย์กลางของกลุ่ม (Cluster Centroid) เทียบกับจุดศูนย์กลางของข้อมูลทั้งหมด (Data Centroid) ส่วนความแปรปรวนภายในกลุ่ม คำนวณจากระยะห่างของสมาชิกภายในกลุ่มเทียบกับจุดศูนย์กลางของกลุ่ม หากดัชนีมีค่ามาก แสดงว่า การจัดกลุ่มข้อมูลมีคุณภาพดี [11]

3) **CDbw (Composed Density between and within Cluster)** เทคนิคนี้คำนวณหาการเกาะเกี่ยว (Cohesion) เพื่อเป็นตัวแทนความหนาแน่นภายในกลุ่ม และคำนวณหาการแยกกัน (Separation) เพื่อเป็นตัวแทนของระยะห่างระหว่างกลุ่ม ถ้าหาก CDbw มีค่าสูง แสดงว่ามีการจัดกลุ่มที่มีคุณภาพ จุดเด่น CDbw นั้นสามารถจัดการข้อมูลที่มีกระจายตัวไม่แน่นอนได้ดี

4) **DBCV (Density-Based Cluster Validation** พัฒนาหลังจากจาก CDbw เพื่อแก้ไขข้อบกพร่องจากข้อมูลรบกวน ดัชนีจะคำนวณหาความหนาแน่นระหว่างกลุ่ม หรือ DSPC และการกระจายตัวภายในกลุ่ม หรือ DSC แล้วหาอัตราส่วนความหนาแน่นระหว่างกลุ่มต่อการกระจายตัวภายในกลุ่ม ถ้าหาก DSC มีค่าน้อย แสดงว่าข้อมูลมีการกระจายตัวต่ำ ทำให้ดัชนี DBCV มีค่ามาก แสดงว่าการจัดกลุ่มข้อมูลมีความคล้ายกัน

4. ผลการดำเนินงาน

4.1 ผลการจัดกลุ่ม

ผลลัพธ์ของการจัดกลุ่ม (ภาพที่ 4) เปรียบเทียบ 2 อัลกอริทึม จำนวน 4 เทคนิค ด้วยการเปลี่ยนค่าพารามิเตอร์ต่าง ๆ พบว่าเทคนิค DBSCAN (Haversine, eps = 1, minpt = 1) มีจำนวนกลุ่ม 445 กลุ่ม (ภาพที่ 4 - ซ้ายล่าง) พบกลุ่มที่มีรูปร่างชัดเจน บริเวณแม่น้ำเจ้าพระยาและส่วนใกล้ทะเล นอกเหนือจากนั้นมีความกระจัดกระจายค่อนข้างมาก เนื่องจากธรรมชาติของข้อมูลจริง ส่วนเทคนิค HDBSCAN (Haversine, eps = 2, minpt = 1) มี

ตารางที่ 1: การเปรียบเทียบของดัชนีวัดคุณภาพการจัดกลุ่ม

Cluster validation	Varied density	Arbitrary shape	Noise handling
Silhouette			
CH Index			
CDbw	✓	✓	
DBCV	✓	✓	✓

จำนวนทั้งหมด 33 กลุ่ม (ภาพที่ 4 - ขวาล่าง) มีผลลัพธ์การจัดกลุ่มใกล้เคียงกับกับเทคนิคก่อนหน้า แต่ควรพบกลุ่มขนาดเล็ก (กลุ่มสีชมพู) แทรกอยู่ระหว่างกลุ่มใหญ่ (กลุ่มสีเขียวและน้ำตาล)

อย่างไรก็ตาม เทคนิค DBSCAN (Haversine, eps = 2, minpt = 2) พบกลุ่มขนาดใหญ่ (กลุ่มสีฟ้า) ครอบคลุมพื้นที่เกือบทั้งกรุงเทพฯ ชนะที่โครงการที่อยู่อาศัยรอบนอกเมือง พบเป็นกลุ่มก้อนขนาดเล็ก (ภาพที่ 4 - ขวาบน) ส่วนเทคนิค DBSCAN (Euclidean, eps = 2, minpt = 1) พบเพียง 1 กลุ่มเดียวเท่านั้น (ภาพที่ 4 - ซ้ายบน)

4.2 ผลการวัดคุณภาพการจัดกลุ่ม

จากผลการศึกษาเมื่อวัดคุณภาพการจัดกลุ่มพบว่าเทคนิคการจัดกลุ่มแต่ละแบบ มีคุณภาพการจัดกลุ่มที่แตกต่างกัน ดังตารางที่ 2

5. สรุป

งานวิจัยนี้ศึกษาการจัดกลุ่มข้อมูลเชิงพื้นที่แบบไม่มีตัวแปรเป้าหมายกำกับ เพื่อนำเป็นตัวแทนสำหรับการทำความเป็นกลุ่มก้อนของโครงการที่อยู่อาศัย โดยใช้อัลกอริทึมการจัดกลุ่ม ประกอบด้วย DBSCAN และ HDBSCAN จำนวนทั้งหมด 4 เทคนิค แล้ววัดผลด้วย Silhouette Coefficient, CH Index, CDbw และ DBCV พบว่าเทคนิค HDBSCAN (Haversine, eps = 2, minpt = 1) มีค่า Silhouette Coefficient 0.19 และ DBCV 2.45 ซึ่งมากกว่าเทคนิคอื่น ๆ บ่งบอกถึงการแบ่งกลุ่มที่มีคุณภาพดี ถึงแม้เทคนิคช้างต้นจะมีค่าดัชนี Silhouette coefficient สูง แต่ยังคงต่ำกว่า 1 ซึ่งเป็นค่าที่มากที่สุด เพราะข้อมูลรบกวนส่งผลกระทบต่อการคำนวณดัชนีวัดคุณภาพ

นอกจากนี้จากผลลัพธ์ เทคนิค DBSCAN (Euclidean, eps = 2, minpt = 1) มีข้อวิเคราะห์ความล้มเหลวของการจัดกลุ่ม เนื่องจากผลลัพธ์การจัดกลุ่มมีเพียงกลุ่มข้อมูลเดียว ทำให้ไม่สามารถประเมินด้วยดัชนีวัดคุณภาพได้

สำหรับงานวิจัยในอนาคต ด้านการเลือกพื้นที่ อาจจะมีการขอบพื้นที่ศึกษาให้มีขนาดเล็กลง เพื่อจะได้เจอกลุ่มขนาดเล็กที่มีความหมาย นอกจากนี้ด้านปัจจัยตัวแปร หากนำข้อมูลยอดขายของโครงการ เช่น จำนวนยูนิตทั้งหมด (Total Unit) จำนวนยูนิตที่ขายได้ (Total Sold) อัตราการดูดซับ (Total Absorption Rate) รวมทั้งปัจจัยด้านความเป็นเมือง อาจจะทำให้การจัดกลุ่มข้อมูลเชิงพื้นที่สะท้อนความเป็นจริงมากขึ้น แล้วนำมาต่อยอดกับงานทางด้านอสังหาริมทรัพย์ เช่น การสร้างโมเดลพยากรณ์ราคาขายโครงการที่อยู่อาศัยเป็นต้น

เอกสารอ้างอิง

[1] N. Ahmed and A. Razak, "A comparative study of different density based spatial clustering algorithms," *International Journal of Computer Applications*, vol. 99, no. 5, pp. 18-25, 2014.

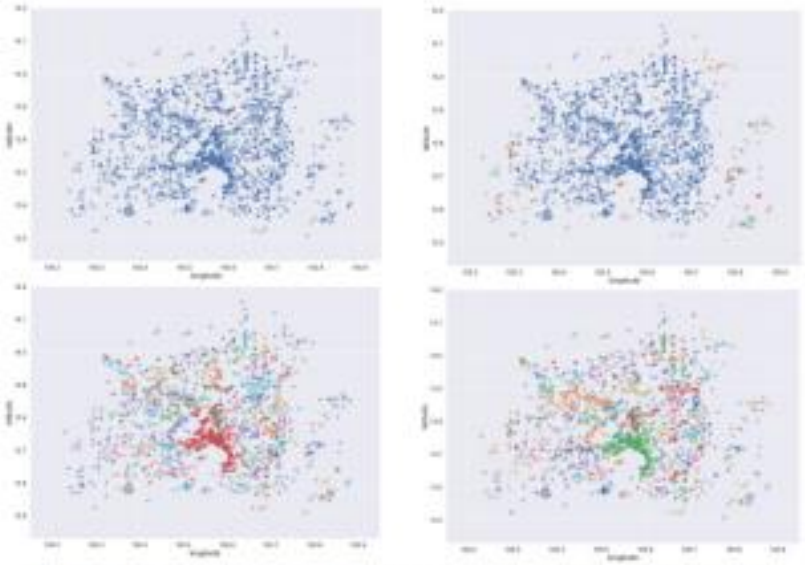
[2] G. Boeing, "Clustering to reduce Spatial data set size," *SocArXiv*, 2018.

[3] T. Wang, C. Ren, Y. Luo and J. Tian "NS-DBSCAN: A Density-Based Clustering Algorithm in Network Space," *SPRS International Journal of Geo-Information*, vol. 8, 2018.

[4] A. Aksac Wang, T. Oryer, and R. Alhajj, "CutESC: Cutting edge spatial clustering technique based on proximity graphs," *Pattern Recognit*, vol. 96, 2019.

ตารางที่ 2: ผลลัพธ์ของดัชนีวัดคุณภาพกลุ่ม

Algorithms	Distance	Parameters	# Cluster	Silhouette	CH	CDbw	DBCV
DBSCAN	Euclidean	eps = 2, minpt = 1	1	-	-	-	-
	Haversine	eps = 2, minpt = 2	42	-0.46	19.0	**0.84**	2.41
	Haversine	eps = 1, minpt = 1	445	0.12	**216.7**	0.71	1.98
HDBSCAN	Haversine	eps = 2, minpt = 1	33	**0.19**	85.5	0.75	**2.45**



ภาพที่ 4: ผลลัพธ์การจัดกลุ่มที่เหมาะสมที่สุดจากการจัดกลุ่มจุดพื้นที่ตามความหนาแน่นด้วย (ก) อัลกอริทึม DBSCAN (Euclidean, eps = 2, minpt = 1), (ข) อัลกอริทึม DBSCAN (Haversine, eps = 2, minpt = 2), (ค) อัลกอริทึม DBSCAN (Haversine, eps = 1, minpt = 1) และ (ง) อัลกอริทึม HDBSCAN (Haversine, eps = 2, minpt = 1)

[5] M. Halkidi and M. Vazirgiannis, "A density-based cluster validity approach using multi-representatives" *Pattern Recognition Letters*, pp. 773–786, 2008.

[6] D. Moulavi and P. A. Jaskowiak, "Density-based clustering validation," *Proceedings of the 2014 SIAM International Conference on Data Mining*, 2014.

[7] M. Ester and H. P. Kriegel, "A density-based algorithm for discovering cluster in large spatial database with noise," *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD96)*, vol.33, 1996.

[8] M. Leland and H. John, "Accelerated Hierarchical Density Based Clustering," *IEEE International Conference on Data Mining Workshops 2017 (ICDMW)*, pp. 33-42, 2017.

[9] T. V. Craenendonck, and H. B. K. Leuven, "Using internal validity measures to compare clustering algorithms," *ICML*, 2015.

[10] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[11] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, pp. 1-27, 1974.

ประวัติผู้เขียน

ชื่อ – นามสกุล วิทวัส แสงสว่าง

ประวัติการศึกษา

พ.ศ. 2554 ภาควิชาธรณีวิทยา คณะวิทยาศาสตร์
จุฬาลงกรณ์มหาวิทยาลัย

ประสบการณ์ทำงาน

พ.ศ. 2566 Business Process Optimization
บริษัท เอไอเอ (ประเทศไทย) จำกัด